Iconic prosody is deeply connected to iconic gesture, and it may occur just as frequently

Perlman, Marcus

Iconic prosody is deeply connected to iconic gesture, and it may occur just as frequently

Marcus Perlman

To appear in: Handbook on Iconicity in Language. Edited by Olga Fischer, Kimi Akita, and Pamela Perniss. Oxford University Press.

Abstract: This chapter examines iconic prosody – how speakers modify their voice in ways that are motivated by their meaning. It argues that iconic prosody is deeply connected to iconic gesture, and may occur just as frequently. Section X.1 presents four detailed examples illustrating the formal and semantic breadth of iconic prosody. Section X.2 defines iconic prosody against the backdrop of traditional studies of prosody. Section X.3 discusses pioneering research in the study of iconic prosody, especially Bolinger's work on emotional expression and Ohala's work on the size frequency code. Section X.4 reviews recent experiments showing that iconic prosody can be investigated in the psycholinguistics laboratory, including studies demonstrating people's ability to create iconic vocalizations to communicate various meanings. Section X.5 examines iconic prosody in the wild, where it features in the multimodal context of quotation and ideophones. Finally, section X.6 concludes with some key questions for future research on iconic prosody.

Keywords: expressive voice; frequency code; intonation; pitch; sound symbolism; vocalization; vocal iconicity

X.1 Introduction and examples of iconic prosody in action

Discussing the use of sign languages, Adam Kendon observed that, in actual performance, signs are "rarely produced as static or 'citation' forms. They are always modified in various ways that refine their expressive role in any given context of the utterance discourse at hand" (in press: 501). This point is equally true for speech. In any given context, the words of a spoken utterance can be articulated louder, slowed down, at a higher pitch, in a creaky voice, with stronger aspiration, extra trills, or various other modifications. This variation in speech is influenced by a wide range of factors ranging from a speaker's emotional state, to the information status of referents, to the type of speech act intended, to name just a few. The focus of this chapter is on how speakers modify their voice in ways that are motivated by the *meaning* they are expressing – what I call 'iconic prosody'.

As we will see, iconic prosody is not a particularly well-defined phenomenon in the research literature. But before jumping into the terminological weeds, let us begin by looking at a few examples of iconic prosody in action.[1] These four examples illustrate some of the formal and semantic breadth of iconic prosody. First, consider the following two utterances spoken by a man describing to a news reporter an automobile accident in which he had been involved just a few minutes before.[2][3] In a highly animated narrative, the speaker, George Lindell, recreates a vivid, multimodal account of what transpired.

[1] The examples are all from American English, spoken by the author. Notably, English is actually an interesting test case for the study of iconicity in spoken languages because it has often been described as extremely impoverished in iconicity (e.g., Perniss et al. 2010), and thus may represent a sort of lower bound case.
[2] https://www.youtube.com/watch?v=6L94Qy_D998&t=114s (beginning at about 1 minute and 44 seconds)
[3] Video excerpts available upon request if the URL is broken.

(1) All of a sudden I was just minding my own business… [Bam!].[4]

(2) And the wires come down. [Foom]. And then arc [arr] [bam] that fire was coming everywhere. It was arcing, sparking, blowing up.

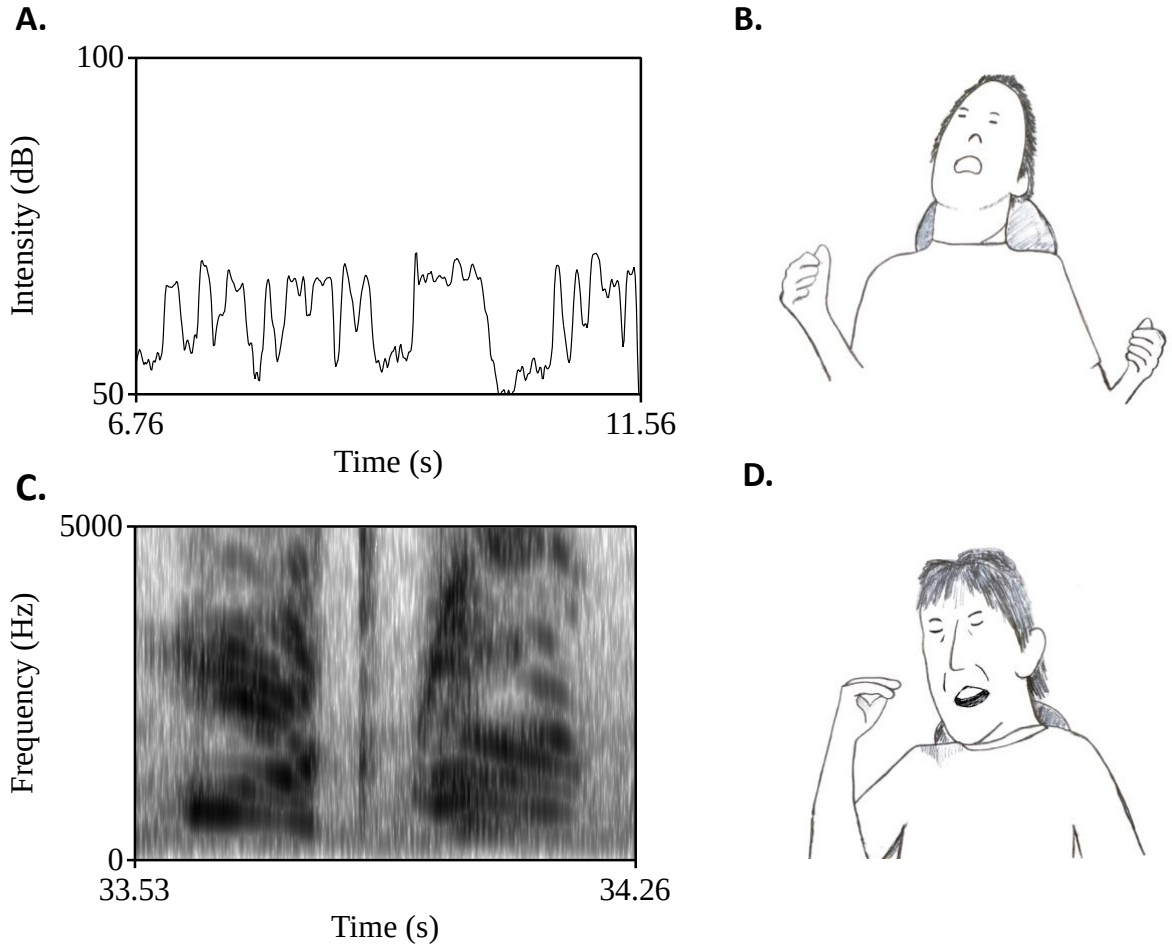

Figure X.1: A. Plot of the intensity of a speaker describing the moment of impact in a car accident. Note the extended peak of intensity that corresponds with the word "bam". B. The speaker re-enacts the moment he was slammed back in the car, timed with "bam". C. Spectrogram of "arc [arr]", produced by the speaker to describe the arc of an electrical spark. Note the release of the /k/ visible at the end of the first syllable "arc", which is not visible in the more wild second syllable. D. Speaker produces an iconic gesture timed to represent the "arc" of the spark.

Lindell produces the first utterance (1) as he sets up the scene and recalls the initial moment of the crash (see Figure X.1A and X.1B). In the culminating "[Bam!]", he transforms the onomatopoeic word into an explosive depiction of the jarring impact experienced inside of the car. Watching the video, we see this articulation is part of a dramatic, full-bodied pantomime of the event. The intensified build-up and explosive release of the word is timed to the speaker's visual re-enactment – jerking his body backwards as if being slammed back

[4] Brackets indicate that a word is pronounced distinctly as a sound effect and may be only vaguely reminiscent of an actual word. Underlining indicates the word was produced at the same time as a gesture.

into the seat. In (2), Lindell again uses his voice in coordination with iconic gesture to present a vivid rendition of events, but this time from an observer viewpoint (see Figure X.1C and X.1D). Recounting the sparks and explosions that resulted from damage to electrical wires, Lindell produces a series of imitative words and sound effects – what sound like variations of "foom", "arc" (and perhaps "spark"?) and "bam" (see Figure X.1C). These words, which seem to reflect aspects of both the sound and motion of the event, are produced in tandem with hand gestures that trace the sparking wires and explosions through the air (see Figure X.1D). Lindell's highly modified words contrast with the more grammatically integrated rendition of these same words produced in the subsequent sentence.

Iconic prosody can also be used to depict dimensions that are more static and silent. Take the following utterance spoken by a mycologist discussing a particular variety of matsutake mushroom found in Japan (originally described in Perlman et al. 2015a).[5] Talking with a small group of people while seated at a table in a diner, the speaker notes the mushroom's distinctive stem.

(3) Like the ones from Nagano have a real silky top … and a [looong] nice stem.

The video shows that as the speaker describes the stem's "long" length, he also depicts it manually with an iconic gesture, using his curled right fingers to trace its length downward from his left palm (see Figure X.2). But this does not constitute all the iconicity in the speaker's utterance: in pronouncing the word "long", the speaker simultaneously uses his voice to produce another sort of iconic effect, extending the temporal duration of the vowel to depict the spatial extension of the stem.

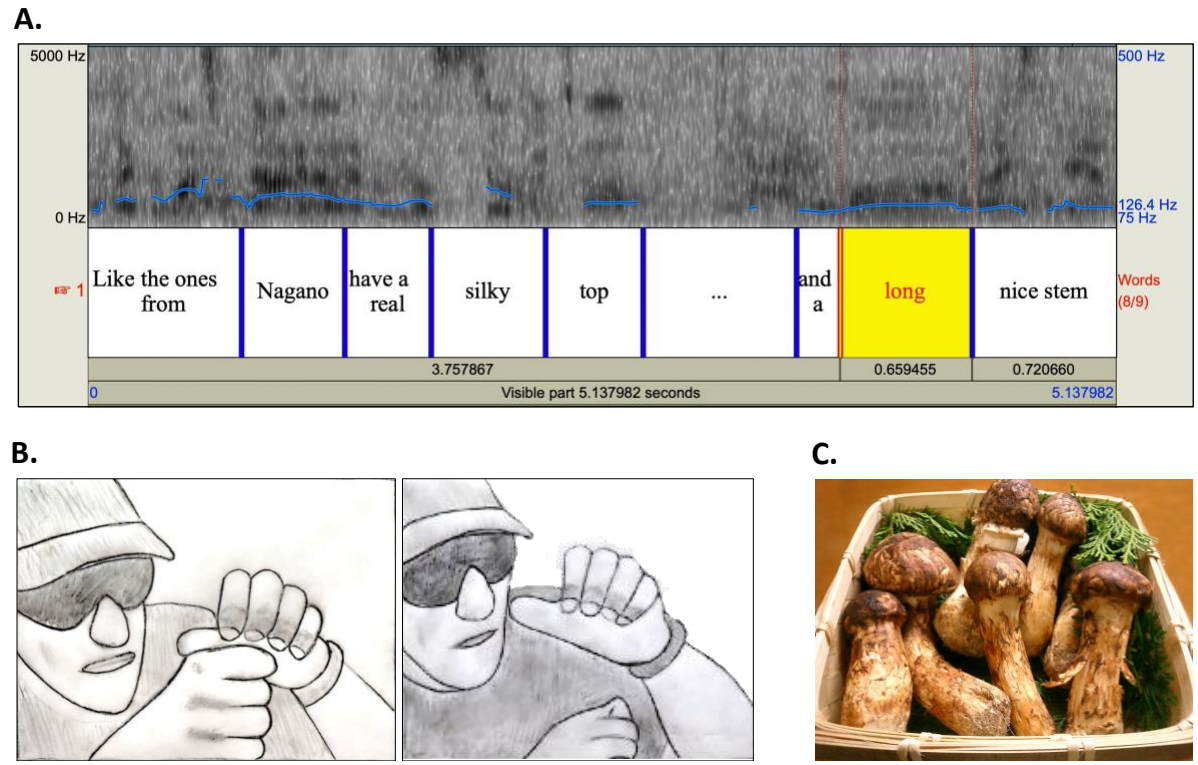

Figure X.2: . A. Screengrab from Praat (cropped), showing a spectrogram of Arora's utterance, along with a text-grid labelling the words. Note that "long" is spoken with the

[5] https:/www.youtube.com/watch?v=VSn3aGGzG1M%2A (at about 4 minutes and 28 seconds)

longest duration in the utterance, although several other words contain more phonemes. B. Drawing of Arora's gesture tracing the "long" length of the stem. C. A basket of matsutake mushrooms. Illustrations by the author. Photo credit: https://commons.wikimedia.org/wiki/File:Matsutake.jpg.

The next example shows the iconic use of phonetic properties of voice to express the tactile property of texture. Consider the retro slogan of the Lay's American-brand Ruffles potato chip (or 'crisp' in British English) (Winter et al. 2022):

(4) R-r-r-uffles have r-r-r-idges.

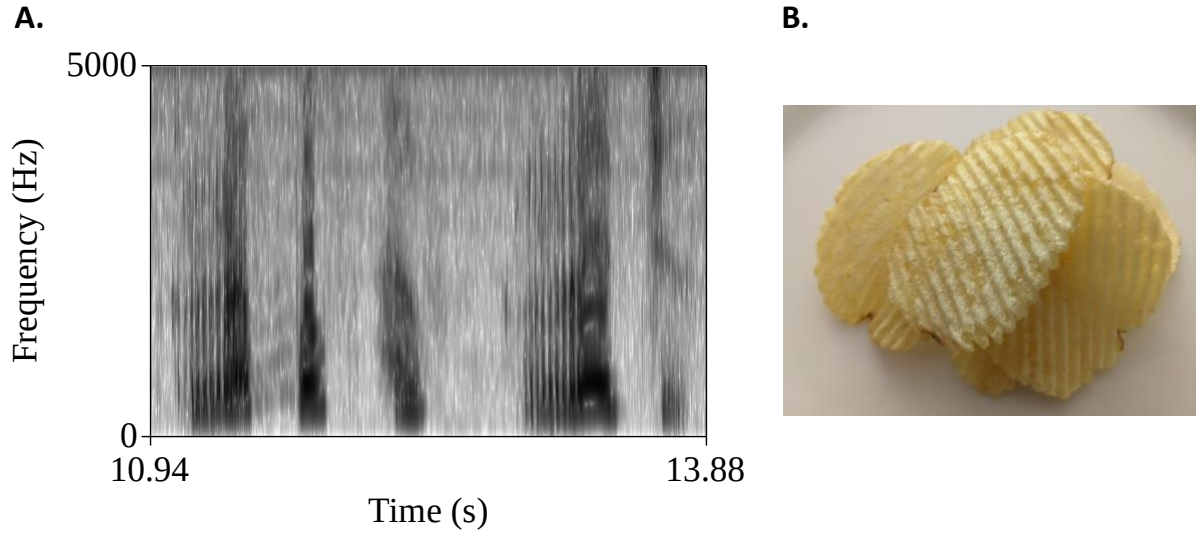

Figure X.3: A. Spectrogram of the potato chip slogan "Ruffles have ridges" as recorded in a commercial. [6] The trills of the /r/s are evident as vertical striations at the start and near the end of the phrase. B. Image of a Ruffles potato chip. Note the visual similarity between the ridges of the chip and the vertical striations of the trilled /r/. Photo credit: Brittany Oakes

The two word-initial trilled segments stand out, not being typical of this variety of English in which rhotic approximants, not trills, would be expected. In context, listeners readily gain the impression that the accentuated pulses of the trill are a depiction of the characteristic corrugations of the potato chip (see Figure X.3). Winter et al. (2022: 5) proposed that the correspondence is "grounded in the acoustically and articulatorily discontinuous nature of trills, which may be associated with the intermittent discontinuity in surface texture".

Finally, in our last example, we see that iconic prosody can be the subject of extensive elaboration and creativity. The following excerpt of speech comes from the television chef Julia Child on her show 'The French Chef'.[7] With her typical flare, Child presents three brioches of different sizes—one small, one medium, and one large—in a brief narration that evokes the story of Goldilocks and the Three Bears:

(5) There was a [little tiny, baby brioche]. And there was the [mother brioche]. And there was the [great big father brioche].

[6] e.g., https://www.youtube.com/watch?v=gL3b58Ibw5U
[7] https://www.youtube.com/watch?v=5vzo6q3epC4 (at about 7 seconds)

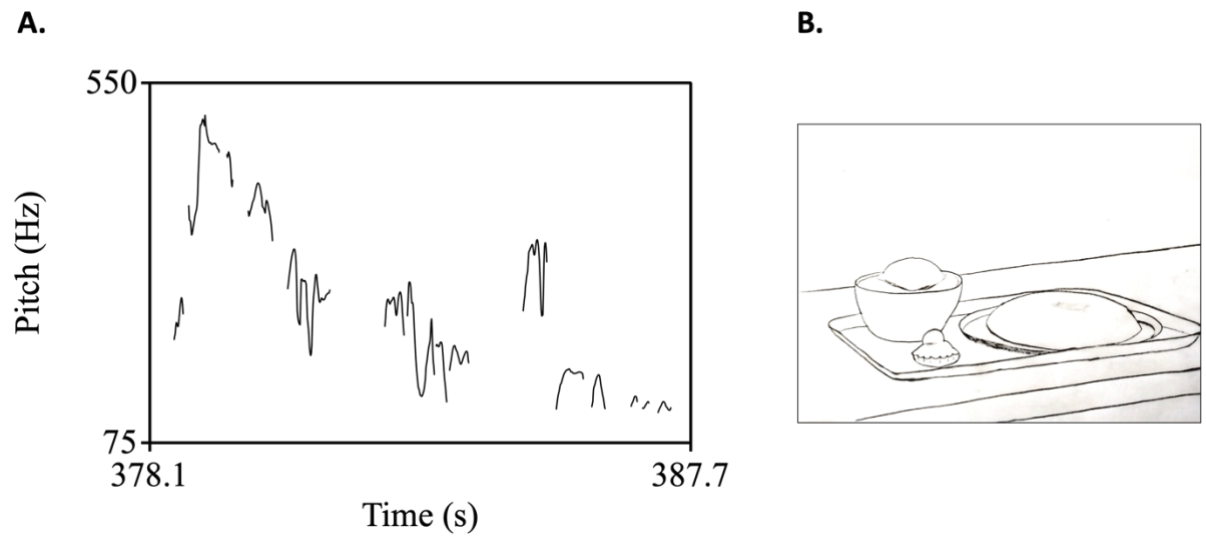

Figure X.2: A. Plot of Julia Child's pitch as she refers to the three different-sized brioches. Her pitch incrementally decreases as she refers first to the small brioche, then to the medium-sized one, and finally to the large one. The small jump in pitch before the final noun phrase corresponds to "and there was a". B. Drawing of the three different-sized brioches. Illustration by Brittany Oakes.

From the transcription, we see that Child imaginatively anthropomorphizes the pastries, inviting the audience to view them as three human-like bears. Yet, the transcribed words miss a vital part of Child's creative presentation – her voice (see Figure X.4). Listening to Child, one hears her bring the brioches to life through the use of iconic prosody. The plot in Figure X.4A shows the pitch of her voice across the respectively sized noun phrases, showing how she distinctly presents the brioches in order of size with her pitch: a very high, child-like voice for the small one, a middle-ranged motherly voice for the medium-sized one, and a deep fatherly voice for the large one. Child's vocal modulations are clearly for dramatic effect, far outside of her usual pitch range and the declination typical of English utterances. Notably, this example is not just a matter of simple mapping between size and pitch. Rather, Child's voice takes on a sort of character-viewpoint perspective as if she were, to an extent, impersonating the bears, even as her words describe them in the third person.

In these examples, we glimpse just a few of the various ways that speakers use their voice, not just for the articulation of words, but – like in the use of iconic gesture – for the depiction of meaning. In what follows, I make the case that phenomena like these, grouped together as iconic prosody, are widespread and common in spoken communication. I begin in Section X.2 by defining iconic prosody against the backdrop of traditional studies of prosody, emphasizing the long observed connection between prosody and gesture. In Section X.3, I discuss the pioneering theories and research in the early study of iconic prosody, focusing especially on Bolinger's (1986, 1983) study of emotional expression and Ohala's (1995, 1984) study of magnitude and the frequency code. Section X.4 reviews more recent experimental studies showing that iconic prosody can be detected and measured within a controlled laboratory setting and also demonstrating people's ability to create iconic (non-linguistic) vocalizations to communicate a wide array of meanings – thus showing the considerable semantic potential for iconic prosody in speech. Section X.5 examines research focused on studying iconic prosody in the wild, where it features especially in the multimodal

context of quotation, the use of ideophones, and in infant-directed speech. Finally, in Section X.6, I conclude with some key questions for future research on iconic prosody, ultimately arguing that an understanding of iconicity prosody is crucial to a grand theory of language and its evolution.

X.2 Defining iconic prosody

Although the term iconic prosody has recently been used a few times in the research literature (e.g., Ćwiek and Fuchs 2019; Murgiano et al. 2021; Perlman et al. 2015a), it is not one that is well established or clearly defined. In fact, and probably not coincidentally, the more general term 'prosody' – which is the subject of a much larger body of research – is itself not all that well defined. As it has been traditionally studied, prosody often serves as a kind of catch-all label for the variable properties of speech that relate to *how* a speaker articulates a given utterance, or portion thereof, as opposed to *what* they say (Wagner and Watson 2010). Prosody typically includes properties like pitch and intonation, rhythm, tempo, loudness, and voice quality, which are sometimes described as the 'musical' aspects of speech, implying that they can be analyzed separately from the particular words that are spoken (or the 'tune' vs. the 'text'; Roettger and Grice 2019). Research on prosody – which has often focused on intonation or the non-lexical use of pitch – is largely concerned with suprasegmental properties of the speech stream that are spread across words and phrases (e.g., 'prosodic units', 'intonational phrases'). However, in actuality, while prosody might be analyzed at the suprasegmental level, it must be realized in the phonetic affordances of spoken words and the phonological segments that comprise them. As Wagner and Watson explained (2010: 2): "at the signal level, there is no separation of prosodic and segmental information. Both use the same channel and encode information by the same phonetic correlates, e.g., fundamental frequency, duration, and intensity".

Traditional research on prosody shows that speakers vary their voice and the way they speak for a wide range of reasons (Cole 2015; Wagner and Watson 2010). For example, the prosody of an utterance can depend on factors like the speaker's internal emotional and physiological state (e.g., Scherer 1986) and attitude (e.g., Bryant and Fox Tree 2002). It can also depend on the prominence of words and phrases within the utterance and discourse, such as when speakers use pitch accent and other cues to indicate topical focus or the information status of referents (e.g., Ladd and Arvaniti 2023). Another well studied function of prosody is its role in segmenting boundaries of words, phrases, and clauses, serving to demarcate the intended syntactic structure of an utterance (e.g., Ferreira 1993). And at a higher level of structure, prosody can distinguish different speech acts, such as a question from an affirmative statement (Pierrehumbert and Hirschberg 1990).

Scholars of prosody have often noted a tight connection between speech prosody and certain types of gestures. Bolinger, for example, observed that prosody bears a deep likeness to gesture, proposing even that they are essentially the same phenomenon, just expressed through the different modalities of "speech and limb", [8] which, he suggests, may be an artificial separation to begin with (1983: 159–160; also Kendon 1980). More recently, experimental studies show that the use of prosody is often 'audiovisual', particularly in functions like indicating pitch accent and focus, where it often acts in conjunction with other bodily movements such as manual (beat) gestures, as well as eyebrow raises, head nods, changes in eye gaze, and postural adjustments (Krahmer and Swerts 2009).

[8] A more neutral way of stating this might be 'between *voice* and limb', which does not assume that one modality carries language and not the other.

Yet, while research on prosody spans a wide range of functions and has often drawn connections to gesture, it mostly neglects cases, such as those in our introductory examples, in which speakers use their voice like *iconic* gesture – as a means for depicting meaning. *This* is the territory of iconic prosody – the "affecto-imagistic dimension" of language (Kita 1997: 380) – often described with terms like 'semantic', 'conceptual', 'referential', or 'propositional', as distinct from other so-called 'pragmatic' aspects of meaning (Nygaard et al. 2009; Shintel et al. 2006).

X.3 Evolutionary roots of iconic prosody

Some scholars of prosody and intonation have, in various ways, suggested that the roots of iconic prosody (even if it was not labeled as such) can be traced back to aspects of our evolved physiology (Bolinger 1986; Gussenhoven 2002; Ohala 1984), motivated by 'universal' factors that are considered to exist, more-or-less, across cultures and languages (Ladd 2001). A pioneer of this approach, Bolinger (1986), described the intonation of speech as a form of gesture reflecting the primitive expressions of emotion and physiological arousal that humans share, in part, with other animals. By this view, the expression of emotion through the voice is rooted deep in our vertebrate ancestry (Bryant 2021) – an observation that goes back to Darwin (1871) in the *Descent of Man*. This work traced the origins of human vocal emotional expression to the emergence of air-breathing tetrapod vertebrates, which put in place the basic blueprint for our vocal tract.

Emotional vocalizations are understood to arise from the evolutionary process of ritualization. Emotions are connected to particular physiological states and correlated with their degree of arousal (e.g., symptoms such as elevated heart rate and blood pressure, increased muscle tension), which can affect the actions involved in vocal production and drive changes in the voice. Over the course of evolution, these mechanically induced changes of the voice have become ritualized into vocal signals – genetically specified, stereotyped and stylized – functioning to communicate particular information about the vocalizer's internal state and plans for action (e.g., to attack or flee). This heritage enables us, as humans, to recognize the level of arousal of vocalizations from across diverse clades of air-breathing vertebrates, from frogs to alligators to various mammals like elephants and panda bears (Filippi et al. 2017).

According to Bolinger (1986: 194), this evolutionarily ancient code of arousal underlies much of the variation in the rise and fall of speakers' pitch, which is driven by a "primitive drive mechanism that raises pitch as tension rises and lowers it as tension falls" . Speakers naturally follow the principle of going "up on what arouses you", which hearers, in turn, interpret as 'important', 'focusing', 'surprising', and 'new'. Going even further, Bolinger (1986: 202) recognized the importance of metaphor in expanding the conceptual scope of emotion and arousal, building on "an upness of effort and tension [of the voice] associated with getting up, lifting, reaching a high place, getting the upper hand, and in general escaping the pull of gravity that decrees that states of rest must somehow be down".

Bolinger (1983: 156) saw this sort of variation in speech as fundamentally "iconic", observing that these elements of intonation reflect, in some way, "their own inner nature". There is a direct and intuitive correspondence between the acoustic properties of a speaker's prosody and their felt emotions and physiological states. Yet, as Bolinger recognized, the claim that prosody is truly iconic in this way is not entirely straightforward. On the one hand, iconicity might be dismissed as "mere symptomization of physical states" (Bolinger 1983: page), outside of speakers' intentional control, and fundamentally different from the depictive gestures that accompany speech, which are typically conceptual in nature (cf. McNeill 1992).

Hinton et al. (1994: 2) called this "corporeal sound symbolism": the sound pattern of prosody "only has 'meaning' in that it directly reflects an internal state of the body or mind". However, Bolinger (1983: 156) argued that to dismiss all emotion in the voice as mere physiological symptom is to "ignore our skill at dissimulation—in pretending, for example, to be keyed up when we are not". As a case in point, consider the finding from a recent meta-analysis that people can successfully communicate, across cultures, twenty-five different emotions through the sound of their voice, either in the form of single words or short sentences, or non-linguistic vocalizations (e.g., screams, laughter, sighs) (Laukka and Elfenbein 2021). Notably, the stimuli of the analyzed experiments were not emitted spontaneously under natural circumstances, but rather, they were all produced by actors – either by profession or just for the purpose of the experiment – tasked to communicate specific emotions. Presumably they did this, to considerable cross-cultural success, by simulating the experience of the emotion and imagining the sound they would produce in that circumstance.

On the other hand, Bolinger (1986) also noted that it is possible to dismiss iconicity, not as a symptom, but for what is, in a sense, the opposite reason – because it is highly conventionalized. In other words, any iconic correspondences that appear to the eye of the analyst, such as in the use of heightened pitch for topical focus, could be fossilized relics of symptoms past – effectively, arbitrary conventions to the synchronic viewpoint of speakers. In analogy to how emotional vocalizations emerged in terrestrial vertebrates through phylogenetic ritualization, the expression of emotion through the voice can become conventionalized through cultural evolution. Through repetition, the natural expression of arousal can evolve into a systematic way to mark certain words in order to highlight them and signal their prominence in discourse, as in the codification of pitch accent. On this issue, Bolinger (1986) made an important point: such patterns of intonation are only actually 'arbitrary' to the extent that they are not felt by speakers in the moments of speaking. Moreover, while it is certainly true that prosody is, in many ways, highly conventionalized in spoken languages, and often produced automatically without much underlying feeling of the speaker, it is also true that much of the variation in speech is spontaneous and tuned to the particulars of the speaking context, both inside and outside of the speaker. Whatever the case – whether a given instance of prosody might be spontaneous or conventionalized – Bolinger argued that most of it, ultimately, was rooted in the iconic expression of emotion and arousal.

Bolinger's theory is compelling, but there is more to the story of how iconic prosody is rooted in our vertebrate physiology. Another and generally complementary idea is that much of prosody stems from the innate expression of size and strength, and relatedly, sex and dominance relations – rooted in what Ohala called the size 'frequency code' (Ohala 1995; also see Pisanski et al. 2016). Larger bodied animals, with larger vocal tracts, tend to produce vocalizations with lower frequencies compared to smaller animals with smaller vocal tracts. Through evolution, the frequency code has become ritualized into the vocal displays of different animals. For example, a survey of the close-range agonistic signals of mammals and birds found that the more threatening, more confident aggressor produces vocalizations that are lower in pitch, compared to submissive whines and yelps, which are higher in pitch (Morton 1977).[9] Ohala's proposal was that, through our biological inheritance as air-breathing vertebrates, this frequency code has also come to underlie many features of human vocal communication, including a range of prosodic phenomena in speech, from the

[9] Ohala (1995) also proposed a multimodal aspect to the frequency code. Observing that smiling (lip-corner retraction) – particularly in contrast to an aggressive 'o-face' – has the effect of raising the resonances of vocalizations, he suggested that the original motivation for smiling may have been the expression of submission and making oneself sound small. If true, this would seem to connect grins and smiles with sound-symbolic phenomena in speech, such as the lip spreading of the diminutive-sounding high front vowel /i/.

intonational contour of questions (high-pitched) versus statements (low-pitched), to how the speaking voice conveys, through high versus low pitch, social messages such as deference versus assertiveness, politeness versus authority, submission versus aggression, uncertainty versus certainty, and female versus male. Across these cases, the frequency code is argued to structure a motivated ('sound-symbolic') connection between the shape of the intonational pattern and its meaning or function in the utterance.

More recently, Gussenhoven (2016) proposed three additional 'biological' codes which, he argued, motivate the expressive use of pitch in complement to Ohala's size-frequency code. Like the frequency code, these codes derive from anatomical and physiological aspects of the speech production mechanism that affect the rate of vocal fold vibration. First is the 'effort code'. More careful pronunciation typically results in wider and more precise pitch excursions, and thus these speech characteristics have come to signal qualities like emphasis and cooperativeness. Second is the 'respiratory code'. Fundamental frequency naturally declines over the course of a breath of speech as a result of diminishing subglottal air pressure, which speakers exploit, for example, when using high pitch to signal a new topic and low pitch to signal continuation of a topic. Third is what Gussenhoven called the 'sirenic code'. He speculated that breathy voice came to be associated with feminine sexiness and related meanings (e.g., relaxed, intimate, friendly, timid) because of changes in the lubrication of women's larynxes that result from hormonal conditions (e.g., pregnant and pre-menstrual states) and arousal. Considering how these codes are realized during speech, Gussenhoven made a similar point to that of Bolinger's on pretense, noting that "communication by means of the codes does not require that these physiological conditions are actually created. It is enough to create the effects", which "are not automatic, but have been brought under vocal control" (Gussenhoven 2002: 48).

Seen in complement to each other, the theories of Bolinger, Ohala – and related hypotheses like those of Gussenhoven – carve out a swath of inter-related domains integral to human experience, which arguably bear some influence on the prosodic form of a wide range of utterances. These authors were also keenly aware that these core mappings, through conceptual processes like metaphor and metonymy, are readily extended and elaborated in various ways. Yet – Bolinger's emphasis on prosody as a form of gesture notwithstanding – there is a crucial gap in these complementary theories. They overlook the homology between many manifestations of prosody and 'iconic gesture', and thus they vastly underestimate the prevalence and scope of iconic prosodic phenomena.

X.4 Investigating iconic prosody in the psycholinguistics laboratory

As we saw in the introductory examples, the use of iconic prosody can be fairly obvious in many cases. Anecdotal examples abound, and it is often evident from casual observation that a speaker is somehow modulating their voice to depict an aspect of their meaning. In scientific practice, however, it can be challenging to measure iconic prosody in a verifiable way (Perlman et al. 2015a). When a speaker produces a gesture with their hands, the movement is clearly distinct from the speech signal. There is no question that the speaker produced a gesture, and if the analyst has a plausible intuition of how its form appears to reflect an element of the speaker's verbal meaning, we are ready to accept it as an iconic gesture. In contrast, iconic prosody is expressed through the same vocal channel as speech, which is also known to vary for a host of factors that are traditionally treated as separate from meaning (e.g., syntax, focus, attitude). Take the description of the matsutake mushroom in (3) above, for instance. That the speaker uses his hands to create an iconic representation of the stem is clear and indisputable; that he also depicts the stem through the extended duration of

his voice is intuitive, but difficult to verify. Consequently, it can be challenging to identify and measure the 'semantic' part of prosody – that is, the part that would reveal its full homology with iconic gesture.[10] As a result, there remain major questions regarding the prevalence and scope of iconic prosody in ordinary day-to-day speech and how it takes shape in spontaneous utterances.

Although experiments on their own cannot fully answer these questions, some recent studies have sought to investigate iconic prosody in the psycholinguistics laboratory as a proof of the concept that prosody can indeed reflect the meaning of an utterance. For example, Shintel et al. (2006) conducted a series of experiments to investigate the production and comprehension of what they called 'analog acoustic expression' – that is, when speakers produce analogical variation in the acoustic properties of speech in such way as to directly convey conceptual or propositional information about external referents. In a first experiment, participants in one condition watched dots on a computer monitor moving either up or down, and stated aloud for each one whether "It is going up" or "… going down". Analysis, which controlled for the phonetics of the target words, showed that participants tended to raise the pitch of their voice when speaking the word "up" and lower it when saying "down". In a subsequent experiment, participants watched dots moving horizontally across the screen at different speeds, and described the direction of the dot ("It is going left" or "… right"). Even though the speed of the dot was not relevant to the task, participants spoke the phrases with a shorter duration when describing the direction of the dot in the fast-moving trials compared to the slow-moving trials. Other studies have shown similar effects of speed under (somewhat) more naturalistic conditions, for example, when freely describing short video clips of fast and slow events (Perlman 2010), or when reading stories that varied in the speed of action (Perlman et al. 2015a). Experiments also confirm that listeners are sensitive to the iconic use of speech rate (Shintel et al. 2006), particularly when it is contextually relevant (Shintel and Nusbaum 2008).

Another study, this one examining a wider set of semantic dimensions, looked to infant-directed speech as a more ecologically valid context to elicit iconic prosody in an experimental setting, particularly as this style of speech is often characterized by exaggerated intonation. Nygaard et al. (2009) asked three adult female speakers – all with extensive experience with young children – to make imaginary requests in infant-directed speech using sentences like "Can you get the *blicket* one?" Each phrase contained a novel word (e.g., "blicket"), assigned in each instance to a different meaning from six antonymic pairs of adjectives: *happy*/*sad*, *hot*/*cold*, *big*/*small*, *tall*/*short*, *yummy*/*yucky*, and *strong*/*weak*. Analysis revealed that each meaning, in contrast to its antonym, displayed a distinct acoustic signature. For example, *strong* was expressed with a larger amplitude than *weak*; *happy* with higher pitch, higher amplitude, and a shorter duration than *sad*; and *tall* with a longer duration than *short*. Differences between antonymic meanings were evident both in how the novel carrier word was pronounced and also in the pronunciation of the full sentence. Follow-up experiments confirmed that listeners were sensitive to the semantic specificity of the acoustic profiles: listeners showed an advantage for identifying the specific meaning of the carrier word that went beyond simply associating it with another meaning of the same positive/negative emotional valence. Nygaard et al. (2009: 127) concluded that these findings demonstrating the semantics of prosody call for a reconceptualization "of traditional distinctions between linguistic and nonlinguistic properties of spoken language" and the "role of prosody in communication" (2009: 142).

These latter results especially point to the possibility that iconic prosody can be expressive of a varied range of meanings. Yet, the research literature has been riddled with

[10] This methodological challenge is parallel to studying the use of gesture by signers (e.g., Emmorey 1999).

speculations and assumptions about the limited potential to express meaning with the voice beyond the use of conventionalized words (Perlman 2017). Indeed, this myth was baked into the foundation of linguistics. Saussure (1983 [1916]) described the vocal channel of speech as serial and thus one-dimensional, an idea that was reflected many decades later in Hockett's (1978: 275) reasoning that, in comparison with sign, "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 fact, the voice is a rich medium for the iconic representation of meaning. From a semiotic standpoint, far from being one-dimensional, sound production with the voice involves a complex orchestration of multiple articulators, including the lungs and diaphragm, vocal folds, throat, lips, jaw, nose, tongue, and teeth. Humans exercise exquisitely coordinated control over these articulators, which can be finely tuned to affect acoustic qualities of the voice like its fundamental frequency and frequency spectrum, temporal qualities like rhythm, duration, and tempo, loudness, and various other aspects of voice quality. Critically, this vocal material is loaded with expressive potential – clay to be molded into all kinds of meanings that a speaker may have in mind.

As evidence, consider a series of experiments in which participants played, under different circumstances, what was effectively a game of 'vocal charades' (Perlman et al., 2015b; Perlman and Cain 2014; Perlman and Lupyan 2018). Asked to express different meanings with the sound of their voice alone (no words permitted), the participants in these studies proved to be impressively good at the task. Even as they sometimes showed hesitation or uncertainty in the process, the acoustic properties of the sounds they produced were remarkably consistent, reflecting their similar intuitions about how to use their voice to express various concepts from domains like texture (smooth vs. rough, dull vs. sharp), space (e.g., down vs. up, near vs. far), number (one vs. few vs. many), size and magnitude (long vs. short, big vs. small), appraisal (good vs. bad, nutritious vs. poisonous), movement (e.g., fast vs. slow), luminance (bright vs. dark), and various others.

Notably, many of these vocalizations can be understood by listeners from diverse linguistic backgrounds. Perlman and Lupyan (2018) conducted a contest (the 'Vocal Iconicity Challenge') in which people were invited to submit recordings of non-linguistic vocalizations designed to communicate 30 different meanings including various kinds of animate and inanimate entities, actions, properties, quantities and demonstratives. Figure X.5 shows oscillograms of two vocalizations taken from the winning submission (i.e., the set of vocalizations that were guessed most accurately by naïve listeners). Figure X.5A shows the sound for *water*, making visible the four repeated drops, and Figure X.5B shows the wispier, more continuously fluctuating sound that was made for *fire*. In a follow-up study, Ćwiek et al. (2021) played the vocalizations from the contest – which were all created by English speakers – to listeners who were speakers of 28 different languages from 12 different language families. The results showed that listeners from each language group were able to guess, to some degree, the intended meaning of the vocalizations for all of the 30 meanings tested. Thus – whatever researchers might argue about any absolute advantage for iconic expression in gesture over vocalization (e.g., Fay et al. 2013; Macuch Silva et al. 2020) – it is evident that the voice offers a wealth of potential for iconic expression across a wide expanse of semantic space.

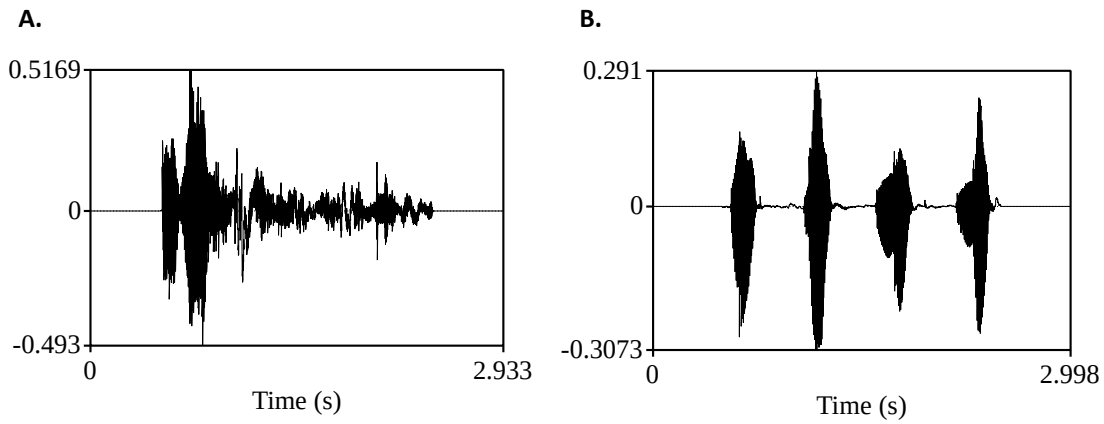

Figure X.5: A. Oscillogram (pressure in Pascals) of a vocalization for water, showing four successive drops. B. Oscillogram of a vocalization for fire, depicting a continuous noisy blaze.

These studies of vocal charades raise the possibility that iconic prosody could be far more prevalent than researchers have realized. Whatever it is that a person is talking about – whatever the meaning they are conjuring to mind and relating through words – it could conceivably find expression in vocal iconicity. It simply is not true that iconicity is squeezed out of speech "perforce" as Hockett (1978: 275) suggested; there is plenty of potential for iconicity in speech. But how prevalent is iconic prosody? How often does it actually happen that the meaning of our message affects how we say it? And how does iconic prosody manifest in real utterances? The value of experimental rigor aside, the most compelling evidence for the abundance of iconic prosody comes from observing it in the wild, where even though it has not yet been precisely quantified, it clearly runs rampant in speech.

X.5 Studying iconic prosody in the wild

In the previous section, we saw how some recent psycholinguistic experiments have managed to capture iconic prosody in the laboratory, providing a verifiable demonstration of its occurrence and hinting at its potential for a wide ranging semantic scope. Yet the design of these experiments – in part resulting from limitations of experiments, in general – may also give some distorted impressions of iconic prosody that do not entirely fit with observations of the phenomenon as it occurs naturally in the wild. Specifically, there are at least three issues with many psycholinguistic experiments on iconic prosody.

First, these experiments have tended to treat prosody as something that is independent of the words that are spoken, often referred to – in line with the traditional view of prosody – as a separate channel of communication that is overlaid on top of the phonological forms of words. This may aptly characterize some cases of iconic prosody, such as (5) above where Child modulates her pitch over strings of multiple words – enacting the voices of the different-sized brioches, as it were. But in other cases, there appears to be more fine-grained, dynamic interaction between the phonetic features of particular words and the specific ways they are iconically modified in an utterance. This is evident in (3) above, for instance, where the word "long" (/lɔŋ/) readily affords lengthening by virtue of having all sonorant consonants. (4) shows another example in the latent potential of a rhotic approximant to take form as an elaborative trill to depict the texture of a potato chip. And in (1) and (2), we saw how the speaker actually transforms onomatopoeic words ("bang"), and even more standard words ("arc"), into free-flowing iconic representations of explosions and sparks. These

examples demonstrate how iconic prosody, rather than being a separate channel overlaid onto speech, can manifest as an outgrowth of the phonetic affordances of words. Indeed, the line between iconic prosody and iconic phonetics is not one that is clearly defined in nature. This aspect of iconic prosody poses challenges for psycholinguistic experiments aiming to draw generalizations across multiple items, such as by using a small set of semantically contrasting target words, or by using a nonsense word as a carrier for contrasting concepts.

Second, many psycholinguistic experiments on iconic prosody – mainly in their design rather than their theoretical persuasion – have treated it as a unimodal phenomenon, failing to take gesture into account, and in many cases even suppressing it by the experimental tasks used. Yet, when we look at natural conversation, the more typical case may be for iconic prosody to occur in multimodal complexes of speech and gesture. For instance, in the description of the "long stem" in (3), we saw tight coordination between the lengthened word and the manual tracing of the stem's length, which appear fused together into a single multimodal representation (of a mostly visual property). The fusion of voice and body is especially striking in the recount of the car accident in (1) and (2), in which the speaker integrates vocal effects, gestures, facial expressions, and bodily posturing into full-bodied re-enactments of the events that transpired.

Third, the design of psycholinguistic experiments of iconic prosody has tended to neglect the considerable flexibility in iconic mappings that a speaker might instantiate. Statistical analyses assess the consistency of a particular pattern of response, but there are often many iconic alternatives available to speakers beyond a single, straightforward one-to-one mapping between form and meaning, and sometimes different intuitive mappings can even be in opposition to each other. One source of this variability is the potential for an aspect of meaning to be iconically represented by different material features of the signal. Take the vocal expression of size, for instance. According to the frequency code, iconic vocalizations created to refer to larger objects ought to be lower-pitched – as in Julia Child's description of the brioches in (5). But in a study in which Chinese children played a vocal charades game, they used higher-pitched sounds to refer to larger objects (e.g., a *big* ball vs. a *small* ball) – likely because of the natural correlation between pitch and intensity as more forceful sounds tend to be higher in pitch (Perlman et al. 2022). Even more consistently than the use of pitch, the children – including both those who were hearing and deaf – used the intensity and duration of their voice to distinguish the size of the referent. They referred to big items with louder, longer sounds. A second source of variability in iconic mappings is what Winter et al. (2021) called 'pluripotentiality' (or 'plurisignificance'; Sidhu and Pexman [2018]) – the potential for a given form to iconically represent different meanings. For example, in a charades experiment with English speakers, high and/or rising pitch was used to distinguish not just small from big, but also good from bad, bright from dark, up from down, sharp from dull, fast from slow, female from male, yes from no, and surprising from predictable (Perlman and Cain 2014).

In part because of these limitations, experimental studies have, so far, provided just a narrow window onto iconic prosody and its full scope in spoken communication. To gain a fuller understanding of iconic prosody, it is necessary to study the phenomenon on its own terms, taking direction from its natural occurrence in the wild. Indeed, in the realm of real-world discourse, we find some promising leads.

One of these is research on the use of quotation (such as through the use of quotatives like *say*, *go*, or *be like*) and related constructions that speakers use to recount and depict events (Clark 2016; Clark and Gerrig 1990). Although quotation is stereotypically associated primarily with the reproduction of speech, these constructions are used far more broadly than this. Even just within the realm of speech, quotations can mimic a wide array of aspects of the quoted speech, from the speaker's intonation, voice quality, and emotional state, to

aspects of their voice related to their age, sex, and gender, to their accent and various other idiosyncratic qualities such as particular affects or manners of speaking.

Quotations are also regularly used to depict various noises and sound effects outside of speech, often with onomatopoetic words. Clark and Gerrig (1990) noted that English has a number of such sound-imitative expressions (see also Rhodes 1994), which span a wide range of sounds from animals and inanimate objects, often serving to depict acoustic qualities such as rhythm, timing, loudness, and pitch. Importantly, these words are not, in many instances, spoken just in their standard phonological form but are elaborated in iconic ways through the use of prosody. For example, consider the quotation from Clark and Gerrig (1990: 781):

(6) The car engine went [brmbrm], and we were off.

Here the quotative "went" introduces the representation of a noise with "[brmbrm]", what could be called a *wild* onomatopoeia (Rhodes 1995) that appears to be a modified variant of the somewhat tamer and more standard "vroom". One can imagine the gravelly tone of voice in which the word might be spoken as an imitation of the sound of the revving engine. We saw a similar phenomenon in the car accident in (1) and (2) above with "[bam]", but in that case, produced without a lexical quotative device (sometimes called a zero quotative), similar to the use of constructed action in sign languages. This example also illustrates how, when producing quotations, speakers often combine their voice and body together to depict the sights and sounds of different events, in many cases fusing these channels together into a single multimodal construction (Blackwell et al. 2015).

Clark and Gerrig (1990) and Clark (2016) concluded that the use of depictive quotative constructions is a common and ordinary part of discourse that needs to be taken into account in psycholinguistic and linguistic theories of language use (. While their data consisted mainly of examples from English, their conclusion finds further support in analyses of a diversity of spoken languages showing that similar kinds of depictive constructions are used with classes of words called ideophones (Dingemanse and Akita 2017). Going well beyond just the sound-based iconicity of onomatopoeia, ideophones – a marked lexical class of depictive words that are widely occurring in most, if not all, of the world's spoken languages – express an array of vivid sensory and motor-related meanings, for example, *hirahira* 'fluttering', *kibikibi* 'brisk, energetic', and *gokun* 'gulping', in Japanese. Ideophones are special in their fluid, "gradient use of material, where changes in form analogously correspond to changes in meaning" (Akita and Dingemanse 2019: 3). Japanese ideophones (traditionally called 'mimetics') – and probably ideophones in general – are often articulated with expressive prosody and are observed to commonly occur with iconic gestures (Kita 1997) and facial expressions (Akita 2019), which work together to express aspects of 'affecto-imagistic' meaning.

Observations of ideophones in ten diverse languages confirm that these patterns are widespread, and further work on a spoken corpus of Japanese indicates that there is often an inverse relationship between the expressiveness of ideophones and their morphosyntactic integration within a linguistic utterance (Dingemanse and Akita 2017). Specifically, when these words are used in quotative and collocational constructions, which are more grammatically independent, they are more often marked by expressive intonation (e.g., markedly higher or lower pitch, intonational pauses) and phonation (e.g. breathy voice, growl, creaky voice, voicelessness, whisper) and by aspects of expressive morphology that are also related to prosody (e.g., reduplication, stem repetition, vowel lengthening, consonant gemination). Ideophones in this expressive context are also used more often with iconic gestures, which typically depict aspects of meaning that are closely related to the meaning of the word (Nuckolls 2020). This trade-off is demonstrated in (2) above, in which the word

"arc" is articulated in highly imitative fashion within a quotation-type construction, and then spoken again in the following sentence, this time integrated grammatically and articulated in a more standard form bearing the progressive suffix *-ing* (see Figure X.1C).

Dingemanse and Akita (2017: page number) suggested that the "performative foregrounding" of ideophones through marked prosody and expressive morphology serves the pragmatic function of marking "some stretch of the speech signal as a depiction". They present, for example, a Japanese utterance with the ideophone [↑zabuun-zabuun↑] 'splaash-splaash', spoken in the upper part of the speaker's pitch range, in which reduplication, the marked pitch, and elongated vowels serve, arguably, to foreground the utterance as a depiction. It is not clear, however, to what extent the marked prosody in this case is iconic and actually contributes to that depiction beyond just drawing attention to it. As we saw in the previous section, people are remarkably good at using their (non-linguistic) voice to express different kinds of meanings, but the question remains whether speakers draw on these intuitions when articulating ideophones and other iconic words.

To date, there is just some limited experimental evidence that they do, at least to an extent. For example, English speakers were better able to guess the meanings of Japanese words when they were spoken with an expressive, conversational quality of voice compared to monotone (Kunihira 1971). A more recent and comprehensive experiment tested whether Dutch listeners could guess the meanings of ideophones from five different languages (Dingemanse et al. 2016). The stimuli were presented under different conditions, including the original recording, a full computer-generated resynthesis, and a resynthesis with either prosody or with phones only. Guessing accuracy across these conditions was mostly above chance, although somewhat modestly so. But critically, guessing was above chance in the prosody only condition, and highest in the original and full resynthesis conditions, indicating that the *way* the words were spoken directly contributed to interpreting their meaning. It remains to be seen how iconic ideophones might become when produced under the fully animated conditions that are typical of their use.

Notably, ideophones and onomatopoeia are commonly used in discourse with young children (Fernald and Morikawa 1993; Laing 2019; Perry et al. 2018), and it is theorized that iconicity in this context facilitates early word learning, helping children to connect spoken forms with their meanings (Imai and Kita 2014). This suggests another scenario in which iconic prosody might be prevalent, where it could enhance the iconicity of certain depictive words and provide children an additional clue to word meaning (e.g., Nygaard et al. 2009). Laing (2019), for example, described the use of iconic prosody in exchanges between a mother and one-year-old infant that focused on animal sounds. Their productions – initiated by the mother and repeated by the child, sometimes with the "correct" prosody and sometimes not – included onomatopoeia such as [hʌhuhuhu], spoken in a high pitch, for an owl, and a low-pitch [i i] for a frog. Although these prosodic effects are, presumably, recognizably iconic for the mother, Laing casts some doubt on whether they are iconic for the child, and proposes instead that their main purpose is to mark the salience of words and help children segment the speech stream.[11] This is an issue similar to that raised above regarding the use of ideophones more generally: to what extent is the expressive prosody that is so typical of ideophones *iconic* as opposed to just serving as an index to mark special words? With the more expansive view of iconic prosody laid out in this and the previous sections, this will be an important question for future research.

[11] In infant-directed signing, deaf mothers modified iconic signs more than other signs, and in this case the modifications were observed to be specifically related to iconic features of the signs (Perniss et al. 2018).

X.6 Conclusion

Throughout this chapter, we have seen the potential for spoken utterances to come alive as our meaning animates our voice, from the minutest phonetic details to the intonational contours that span across words and phrases. Looking at a range of examples, we have seen how these various manifestations of iconic prosody bear a close relationship and deep likeness to iconic gesture (cf. Kendon 2004; McNeill 1992). Indeed, Bolinger recognized early on the essential similarity between spoken prosody and gesture, even identifying an iconic basis for much of prosody, which he suggested was rooted in the evolutionarily ancient expression of emotion and arousal. Yet, Bolinger's theory – and other complementary ideas put forward, such as Ohala's size frequency code – fail to appreciate the full iconic potential of the voice, and consequently, the full extent of iconic prosody in ordinary, day-to-day discourse. Charades experiments demonstrate how people are able to use their voice to express a wide range of meanings, and psycholinguistic experiments show proof of the concept – at least within a few semantic domains such as speed and vertical motion – that iconic prosody can be incorporated, sometimes fairly automatically, into ordinary speech. But the full richness of iconic prosody is perhaps most fully on display in the linguistically widespread use of onomatopoeia and ideophones and the kinds of freelancing constructions – like quotation and constructed action – in which they commonly occur.

The recognition that iconic prosody is a common part of speaking has profound implications for psycholinguistic and linguistic theories of language. Spoken language is, potentially, far more iconic than most researchers have previously realized. This raises a number of exciting questions for ongoing research. What does iconic prosody, and its close relationship to iconic gesture, reveal about the cognitive processes that underlie the use, and especially, the production of language? How do iconic actions of the vocal tract come to be part of the way we conceptualize and talk about our experiences? What is the full extent of iconic prosody, and how is it extended through creative conceptual processes like metaphor and metonymy? And what is the relationship between iconic prosody in spoken languages and the kinds of iconic modulations that occur in signed languages? Ultimately, a complete understanding of iconic prosody will be crucial to a developing a grand theory of language, spoken and signed.

Acknowledgments
Thanks to Ray Gibbs, Timo Roettger, an anonymous reviewer, and the editors Kimi Akita and Olga Fischer for their helpful comments on the manuscript. Thanks also to Brittany Oakes for her help with illustrations and photography.

References

Akita, Kimi (2019). Mimetics, gaze, and facial expression in a multimodal corpus of Japanese, in Kimi Akita and Prashant Pardeshi (eds), *Ideophones, Mimetics and Expressives*, Amsterdam: Benjamins, 229–47.

Akita, Kimi, and Mark Dingemanse (2019). Ideophones (mimetics, expressives), in Kimi Akita and Mark Dingemanse (eds), *Oxford Research Encyclopedia of Linguistics*, Oxford: Oxford University Press, 1-18.

Blackwell, Natalia L., Marcus Perlman, and Jean E. Fox Tree (2015). Quotation as a multimodal construction, *Journal of Pragmatics* 81: 1–7.

Bolinger, Dwight (1983). Intonation and gesture, *American Speech* 58: 156–74.

Bolinger, Dwight (1986). *Intonation and Its Parts: Melody in Spoken English*. Stanford: Stanford University Press.

Bryant, Gregory A. (2021). The evolution of human vocal emotion, *Emotion Review* 13: 25–33.

Bryant, Gregory A., and Jean E. Fox Tree (2002). Recognizing verbal irony in spontaneous speech, *Metaphor and Symbol* 17: 99–119.

Clark, Herbert H. (2016). Depicting as a method of communication, *Psychological Review* 123: 324–47.

Clark, Herbert H., and Richard J. Gerrig (1990). Quotations as demonstrations, *Language* 66: 764–805.

Cole, Jennifer (2015). Prosody in context: A review, *Language, Cognition and Neuroscience* 30: 1–31.

Ćwiek, Aleksandra, and Susanne Fuchs (2019). Iconic prosody is rooted in sensori-motor properties: Fundamental frequency and the vertical space, in Jim Davies and Evangelia Chrysikou (eds), *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*: 1572–78.

Ćwiek, Aleksandra, Susanne Fuchs, Christoph Draxler, Eva Liina Asu, Dan Dediu, Katri Hiovain, Shigeto Kawahara, et al. (2021). Novel vocalizations are understood across cultures, *Scientific Reports* 11: 10108.

Darwin, Charles (1871). *The Descent of Man: Selection in Relation to Sex*. New edition. London New York, NY Camberwell, Victoria Toronto, Ontario: Penguin Classics.

Dingemanse, Mark, and Kimi Akita (2017). An inverse relation between expressiveness and grammatical integration: On the morphosyntactic typology of ideophones, with special reference to Japanese, *Journal of Linguistics* 53: 501–32.

Dingemanse, Mark, Will Schuerman, Eva Reinisch, Sylvia Tufvesson, and Holger Mitterer (2016). What sound symbolism can and cannot do: Testing the iconicity of ideophones from five languages, *Language* 92: e117–33.

Emmorey, Karen (1999). Do signers gesture?, in Lynn Messing and Ruth Campbell (eds), *Gesture, Speech, and Sign*, Oxford: Oxford University Press, 133-59.

Fay, Nicolas, Michael Arbib, and Simon Garrod (2013). How to bootstrap a human communication system, *Cognitive Science* 37: 1356–67.

Fernald, Anne, and Hiromi Morikawa (1993). Common themes and cultural variations in Japanese and American mothers' speech to infants, *Child Development* 64: 637–56.

Ferreira, Fernanda (1993). Creation of prosody during sentence production. *Psychological Review* 100: 233–53.

Filippi, Piera, Jenna V. Congdon, John Hoang, Daniel L. Bowling, Stephan A. Reber, Andrius Pašukonis, Marisa Hoeschele, et al. (2017). Humans recognize emotional arousal in vocalizations across all classes of terrestrial vertebrates: Evidence for acoustic universals, *Proceedings of the Royal Society B: Biological Sciences* 284: 20170990.

Gussenhoven, Carlos (2002). Intonation and interpretation: Phonetics and phonology, in Bernard Bel & Isabelle Marlien (eds), *Proceedings of the Speech Prosody 2002 Conference*, Aix-en-Provence: Laboratoire Parole et Langage, 47-57.
Gussenhoven, Carlos (2016). Foundations of intonational meaning: Anatomical and physiological factors, *Topics in Cognitive Science* 8: 425–34.
Hinton, Leanne, Johanna Nichols, and John Ohala (1995). Introduction: Sound-symbolic processes, in Leanne Hinton, Johanna Nichols, and John J. Ohala (eds), *Sound Symbolism*, Cambridge: Cambridge University Press, 1–12.
Hockett, Charles F. (1978). In search of Jove's brow, *American Speech* 53: 243–313.
Imai, Mutsumi, and Sotaro Kita (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution, *Philosophical Transactions of the Royal Society B* 369: 20130298.
Kendon, Adam (1980). Gesticulation and speech: Two aspects of the process of utterance, in Mary Key (ed), *The Relationship of Verbal and Nonverbal Communication*, Berlin: De Gruyter Mouton, 207-26.
Kendon, Adam (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.
Kendon, Adam (in press). *Three Modalities of Languaging: Speaking, Gesturing, Signing: Selected Essays 1972-2022*. Amsterdam: Benjamins.
Kita, Sotaro (1997). Two-dimensional semantic analysis of Japanese mimetics, *Linguistics* 35: 379–416.
Krahmer, Emiel, and Marc Swerts (2009). Audiovisual prosody—Introduction to the special issue, *Language and Speech* 52: 129–33.
Kunihira, Shirou (1971). Effects of the expressive voice on phonetic symbolism, *Journal of Verbal Learning and Verbal Behavior* 10: 427–29.
Ladd, D. Robert (2001). Intonational universals and intonational typology, in Martin Haspelmath, Ekkehard König, Wulf Oesterreicher and Wolfgang Raible (eds), *Language Typology and Language Universals: An International Handbook*, Berlin: De Gruyter Mouton, 1380-90.
Ladd, D. Robert, and Amalia Arvaniti (2023). Prosodic prominence across languages, *Annual Review of Linguistics* 9(1): 171-93.
Laing, Catherine (2019). A role for onomatopoeia in early language: Evidence from phonological development, *Language and Cognition* 11: 173–87.
Laukka, Petri, and Hillary Anger Elfenbein (2021). Cross-cultural emotion recognition and in-group advantage in vocal expression: A meta-analysis, *Emotion Review* 13: 3–11.
Macuch Silva, Vinicius, Judith Holler, Asli Ozyurek, and Seán G. Roberts (2020). Multimodality and the origin of a novel communication system in face-to-face interaction, *Royal Society Open Science* 7: 182056.
McNeill, David (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University of Chicago Press.
Morton, Eugene S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds, *The American Naturalist* 111: 855–69.
Murgiano, Margherita, Yasamin Motamedi, and Gabriella Vigliocco (2021). Situating language in the real-world: The role of multimodal iconicity and indexicality, *Journal of Cognition* 4: 38.
Nuckolls, Janis B. (2020). "How Do You Even Know What Ideophones Mean?": Gestures' contributions to ideophone semantics in Quichua, *Gesture* 19: 161–95.
Nygaard, Lynne C., Debora S. Herold, and Laura L. Namy (2009). The semantics of prosody: Acoustic and perceptual evidence of prosodic correlates to word meaning, *Cognitive Science* 33: 127–46.

Ohala, John J. (1984). An ethological perspective on common cross-language utilization of F_0 of voice, *Phonetica* 41: 1–16.
Ohala, John J. (1995). The frequency code underlies the sound-symbolic use of voice pitch, in Leanne Hinton, Johanna Nichols, and John J. Ohala (eds), *Sound Symbolism*, Cambridge: Cambridge University Press, 325–47.
Perlman, Marcus (2010). Talking fast: The use of speech rate as iconic gesture, in Fey Parrill, Vera Tobin, and Mark Turner (eds), *Meaning, Form and Body*, Stanford: CSLI Publications, 245–62.
Perlman, Marcus (2017). Debunking two myths against vocal origins of language: Language is iconic and multimodal to the core, *Interaction Studies* 18: 376–401.
Perlman, Marcus, and Ashley A. Cain (2014). Iconicity in vocalization, comparisons with gesture, and implications for theories on the evolution of language, *Gesture* 14: 320–50.
Perlman, Marcus, Nathaniel Clark, and Marlene Johansson Falck (2015a). Iconic prosody in story reading, *Cognitive Science* 39: 1348–68.
Perlman, Marcus, Rick Dale, and Gary Lupyan (2015b). Iconicity can ground the creation of vocal symbols, *Royal Society Open Science* 2: 150152.
Perlman, Marcus, and Gary Lupyan (2018). People can create iconic vocalizations to communicate various meanings to naïve listeners, *Scientific Reports* 8: 2634.
Perlman, Marcus, Jing Paul, and Gary Lupyan (2022). Vocal communication of magnitude across language, age, and auditory experience. *Journal of Experimental Psychology: General* 151: 885–96.
Perniss, Pamela, Jenny C. Lu, Gary Morgan, and Gabriella Vigliocco (2018). Mapping language to the world: The role of iconicity in the sign language input, *Developmental Science* 21: e12551.
Perniss, Pamela, Robin L. Thompson, and Gabriella Vigliocco (2010). Iconicity as a general property of language: Evidence from spoken and signed languages, *Frontiers in Psychology* 1.
Perry, Lynn K., Marcus Perlman, Bodo Winter, Dominic W. Massaro, and Gary Lupyan (2018). Iconicity in the Speech of Children and Adults', *Developmental Science* 21: e12572.
Pierrehumbert, Janet B., and Julia Hirschberg (1990). 'The meaning of intonational contours in the interpretation of discourse', in Philip R. Cohen, Jerry Morgan, and Martha E. Pollack (eds), *Intentions in Communication*, Cambridge, MA: The MIT Press, 271-311.
Pisanski, Katarzyna, Valentina Cartei, Carolyn McGettigan, Jordan Raine, and David Reby (2016). Voice modulation: A window into the origins of human vocal control?, *Trends in Cognitive Sciences* 20: 304–18.
Rhodes, Richard (1995). Aural images, in Leanne Hinton, Johanna Nichols, and John J. Ohala (eds), *Sound Symbolism*, Cambridge: Cambridge University Press, 276–92.
Roettger, Timo B., and Martine Grice (2019). The tune drives the text: Competing information channels of speech shape phonological systems, *Language Dynamics and Change* 9: 265–98.
Saussure, Ferdinand de (1983). *Course in General Linguistics*. La Salle, IL: Open Court.
Scherer, Klaus R. (1986). Vocal affect expression: A review and a model for future research, *Psychological Bulletin* 99: 143–65.
Shintel, Hadas, and Howard Nusbaum (2008). Moving to the speed of sound: Context modulation of the effect of acoustic properties of speech, *Cognitive Science: A Multidisciplinary Journal* 32: 1063–74.

Shintel, Hadas, Howard C. Nusbaum, and Arika Okrent (2006). Analog acoustic expression in speech communication, *Journal of Memory and Language* 55: 167–77.
Sidhu, David M., and Penny M. Pexman (2018). Five mechanisms of sound symbolic association, *Psychonomic Bulletin & Review* 25: 1619–43.
Wagner, Michael, and Duane G. Watson (2010). Experimental and theoretical advances in prosody: A review, *Language and Cognitive Processes* 25: 905–45.
Winter, Bodo, Grace Eunhae Oh, Iris Hübscher, Kaori Idemaru, Lucien Brown, Pilar Prieto, and Sven Grawunder (2021). Rethinking the frequency code: A meta-analytic review of the role of acoustic body size in communicative phenomena, *Philosophical Transactions of the Royal Society B: Biological Sciences* 376: 20200400.
Winter, Bodo, Márton Sóskuthy, Marcus Perlman, and Mark Dingemanse (2022). Trilled /r/ is associated with roughness, linking sound and touch across spoken languages, *Scientific Reports* 12: 1035.