The Territory Between Speech and Song: A Joint Speech Perspective

Fred Cummins
University College Dublin, Dublin, Ireland

Speech and song have frequently been treated as contrasting categories. We here observe a variety of collective activities in which multiple participants utter the same thing at the same time, a behavior we call joint speech. This simple empirical definition serves to single out practices of ritual, protest, and the enactment of identity that span the range from speech to song and allows consideration of the manner in which such activities serve to ground collectives. We consider how the musical elements in joint speech such as rhythm, melody, and instrumentation are related to the context of occurrence and the purposes of the participants. While music and language have been greatly altered by developments in media technologies—from writing to recordings—joint speech has been, and continues to be, an integral part of practices, both formal and informal, from which communities derive their identity. The absence of joint speech from the scientific treatment of language has made language appear as an abstract intellectual and highly individualized activity. Joint speech may act as a corrective to draw our attention back to the voice in context, and the manner in which collective identities are enacted.

Received: September 18, 2018, accepted September 13, 2019.

Key words: joint speech, song, chant, ritual, enaction

The 2016 award of the Nobel Prize for literature to a singer-songwriter (Bob Dylan) caused some controversy. It seemed to some to ignore a dividing line between literature and song that was no mere administrative nicety. That line separated a serious domain (literature) from something rather more lightweight. Once the boundary between literature and song was even temporarily erased in this manner, there was immediate clamor from other quarters for similar recognition of other singer-songwriters—such as Leonard Cohen—to be given similar consideration. The Nobel Prize committee, it seems, had disturbed a state of affairs that had persisted without question for a long time.

Speech and song are categories of vocal activity that many would regard as distinct. Speech is the currency with which we engage with others; we conduct transactions, inform, debate, address, and cajole others in order to negotiate the difficult social territory that is our primate legacy. Song, on the other hand, does not seem to serve much of a determinate function, except to entertain. The lyrics of a song, when recognized for the first time, may impart textual novelty, but that hardly serves to adequately describe the manner in which songs influence, amuse, and engage us. Most spoken utterances are singular, never to be repeated. Songs are determinate entities that persist through repetition. Speech is but one modality for the serious business of language, and language, so it goes, is the uniquely human innovation that profoundly affected our species, enabling the development of human civilization, and ensuring that we could clearly hold the rather more important line that separates homo sapiens from their nearest relatives, the apes. Language is also copious enough to contain the whole of literature within it. Song too has a home, but it is in the domain of music. By "song," I wish to pick out behaviors uncontroversially describable as singing, or melodic vocalization, and not merely those instances that display the more formal structure of verse and chorus. The domain of music provides some rather more promising parallels between animals and humans than language (e.g., in the chorusing of gibbons, the duetting of songbirds, or the profound whistles and groans of whales). Furthermore, some have argued that music, unlike language, lacks any obvious selection advantage on those who perform or consume it (Pinker, 1999). The rich integration of music making practices in the widest variety of social activity ensures that counter arguments (e.g., about the role of music in group bonding, in mating, and more, are not lacking).

Speech and song are usually treated as categorically distinct behaviors. We will begin below by considering how and why they have been treated as distinct, and contrasting, forms of activity. There are indeed many reasons for researchers to treat the two in this manner, often as part of a larger argument in which their superordinate domains—language and music—are interpreted as qualitatively different activities; such separation suggests that each might provide insight into very different aspects of human sociality and behavior.

Music Perception, volume 37, issue 4, pp. 339–350, issn 0730-7829, electronic issn 1533-8312. DOI: https://doi.org/10.1525/mp.2020.37.4.339

But it is not difficult to enumerate forms of vocal activity that seem to inhabit a poorly mapped territory between these two domains, defying neat categorization. Many of these forms of indeterminate provenance are described in English using the term "chant," which does not clearly belong to either one of the two categories alone.

This conflation of two domains that are normally treated as separate motivates the particular empirical approach to be adopted here, in which we use a simple definition of joint speech as a means of organizing our observations. Joint speech is defined as speech in which multiple people utter the same thing at the same time (Cummins, 2014a). (The term "utter" is used here to emphasize that vocalization is in play, without committing to an interpretation of it as either speaking or singing.) We will use this utilitarian definition of joint speech as an empirical entry point to many behaviors that have hitherto been treated as cultural genres, and thus been largely untroubled by the inquiries of natural science. The simplicity of the definition can be leveraged to bring into focus many highly articulated forms of human practice, with the goal of relating the observed vocal activity to the collective behaviors and purposes of those so engaged. The joint consideration of vocal activity together with the associated rituals, gestures, and practices of collective enunciation may be of use in the delineation of relations between overt form and underlying significance. No claim is made that joint speech is a natural kind, just that where such activities are going on, there is likely to be activity that is relevant to the construction and maintenance of various kinds of social order.

The territory of joint speech is rich and varied, but its geography is largely unmapped (but see Cummins, 2018). Attending to the manner in which joint speech has been integrated into human activity allows us to consider how participation and musical form have developed together, producing many different kinds of relation between text, participants, and collectivities. There is a particularly important strand to be traced in following the use of joint speech within formal rituals, both religious and secular, that may be helpful in understanding the relation between the uttering of a text and the context in which that uttering takes place. This is a story of the voice, of participation, and of the act of uttering; it is not an account of written texts, of performance, or of recordings. It opens a window into the manner in which many kinds of collective subjectivities are grounded, or enacted, through rite, ritual, and collective activity. Finally, we consider how this reframing of the manner in which we view human vocal activity might bear consequences for musicologists, anthropologists, social scientists, and linguists, as they consider phenomena that ground our collective being, but that are not well described within the classical boundaries of these disciplines.

When Speech And Song Are Distinct

If one starts out with the framing assumption that speech and song are categorically different forms of vocalization, then it is easy to design experiments that elicit categorically different forms of utterances—speech here and song there—with clear differences between them, and it is not difficult to draw attention to some empirical differences between the samples one has used. This is the way speech and song are usually treated; it leads to some rather familiar positions which we will visit only briefly and incompletely. We begin with a well-known attempt to highlight the distinctness of language and music most generally.

In his 1999 volume, ambitiously titled "How the Mind Works," Stephen Pinker notoriously suggested that our liking for music is an epiphenomenon, serving no purpose of its own (Pinker, 1999). It arises, he conjectures, because the elements that make up music are elements that we experience as pleasurable for independent reasons:

> Now, if the intellectual faculties could identify the pleasure-giving patterns, purify them, and concentrate them, the brain could stimulate itself without the messiness of electrodes or drugs. It could give itself intense artificial doses of the sights and sounds and smells that ordinarily are given off by healthful environments. We enjoy strawberry cheesecake, but not because we evolved a taste for it. We evolved circuits that gave us trickles of enjoyment from the sweet taste of ripe fruit, the creamy mouth feel of fats and oils from nuts and meat, and the coolness of fresh water. Cheesecake packs a sensual wallop unlike anything in the natural world because it is a brew of megadoses of agreeable stimuli which we concocted for the express purpose of pressing our pleasure buttons. Pornography is another pleasure technology. . . . [T]he arts are a third. (Pinker, 1999, pp. 524–525)

Music, literature and pornography all belong in the bin of the pleasurable but accidental and hence "useless," quite unlike language (and hence speech) which Pinker sees as central to our cognitive capacities. This argument generated considerable negative criticism when published, and Pinker himself has toned down the apparent dismissiveness of the claim somewhat, but it

remains a useful reference point, as it distinguishes between language (selected for, effective, important) and music (an evolutionary "spandrel," accidental, pleasant but inessential) in a way that the discussion herein will interrogate in a novel way (although our focus will not be on evolutionary matters, narrowly considered).

Speech and song have been addressed with respect to differences in their relation to respiration. Treated as different categories of vocalization, song is characterized by a "regular and rhythmic progression from speech sound to speech sound" (Proctor, 2013) which contrasts with the laxness, informality, and unpredictability of everyday speech. Comparing respiration in reciting poems that were speech-like or song-like, Yang observed greater breath volumes for the song-like poems, which also exhibited more frequent exhalations (Yang, 2015). In Binazzi et al. (2006), a direct comparison of speech and song was effected by having participants read or sing the lyrics to "O Christmas Tree" (in Italian). This revealed differences in breath frequency (more frequent in singing), in expiratory duration (longer in singing), in inspiratory flow (greater in singing), and in the number of syllables enunciated per second (smaller in singing).

Neuroscientists have also addressed the speech/song pair, treating the two forms of vocal activity as related, but distinct. Callan, Kawato, Parsons, and Turner (2007) found cerebellar activity in speech and song to be differently lateralized: more left cerebellum activity for song, and more on the right for speech, reversing a pattern familiar from studies of cortical involvement, as detailed, for example, in Callan et al. (2006). Similar contrasts, with similar observations about differential lateralization are reported in Riecker, Ackermann, Wildgruber, Dogil, and Grodd (2000).

Within linguistics, and especially the two domains concerned with the sounds of speech—phonetics and phonology—early analysis of the systematic patterning of speech sounds that is used to encode categorical differences became the bedrock of structural linguistics (Jakobson, Fant, & Halle, 1951). In line with the structuralist approach of Saussure (de Saussure, 2011), the measurable features of speech sound (and by implication, of the underlying articulatory movements) were viewed as either linguistic (if they supported categorical differences, such as a contrast between /bat/ and /pat/) or as extra- or para-linguistic if they seemed to be gradient rather than categorical (such as the amplitude of the voice). Linguistics concerned itself in the first instance solely with the categorical variables, while the continuously varying elements were regarded as someone else's responsibility. From this traditional perspective, there is nothing for linguistics to study in song, as it does not bring further linguistic contrasts to the table.

Interestingly, within the discipline of phonetics, which has the most direct empirical contact with the sounds and movements involved in speech, the last few decades have seen a remarkable shift from attempts to identify these "linguistic" features (phonemes, or as Bob Port has suggested, the ghosts of letters, Port, 2007) to consideration of just those more musical elements of the voice, collectively grouped under the term "prosody," that have largely resisted categorical distinctions. This has brought a rich variety to phonetic exploration of rhythm, melody, voice quality, and phrasing—properties not captured by linguistic models nor represented explicitly in orthography. Empirical considerations thus point towards something more interesting than two separate classes.

An ethnomusicological approach to the commonalities and differences between speech and song is provided by List (1963). His starting point is the desire to implement an objective categorization procedure that will distinguish between speech and song based on acoustic properties, or more specifically, based on melodic variation, alone. However, in recognition of the existence of forms of vocalization that are neither clearly and exclusively one or the other, the procedure to be designed should "make feasible the proper classification of any existing intermediate forms, and should indicate their relations to each other and to speech and song as such" (List, 1963, p. 1). This leads him to consider several examples taken from indigenous cultures including The Nyangumata tribe of Australian Aborigines, Hopi Native Americans, Maori chants from New Zealand, and others. The complex landscape that thereby comes into view inspired List to develop a two-dimensional feature space, in which one axis represents variation between speech-like forms and song-like forms (though how one is to place a given exemplar is not clear), while the second orthogonal axis is supposed to reflect the degree of intonational expansion, or melodic salience (my gloss) of a specific exemplar. I do not know if the proposed representational scheme was ever employed to anybody's satisfaction, but it is worth noticing some relevant degrees of variation that it misses, which nevertheless seem to be essential in any exploration of the territory between speech and song.

Melody—or intonational variation—is the basis of the scheme proposed by List, but it omits all consideration of rhythmicity and, indeed, harmony. In what follows, we will seek to distinguish at least two different forms of rhythmicity: first, the presence of a recurring beat or

periodic regularity, and then the organization of events aligned with such a beat within a metrical hierarchy. As we shall see, these two facets of rhythmicity are separable and the first may be found without the second. Harmony will also be relevant as the introduction of tonal centers and distinguishable voices that interact and respond to each other will be found to be important in understanding the purposes of participants.

The second major property that List's scheme misses will be the primary focus here. It omits any consideration of the broader context in which the vocalizing happens that frames and structures the manner in which anyone might participate. Thus, when one considers, as List does, the warblings of a livestock auctioneer, and compares these to a Maori ritual chant designed to drive away unwanted flocks of birds, it is surely absolutely essential to include the fact that the auctioneer is necessarily speaking as an individual in a highly structured form of exchange that is transactional in nature with different roles for bidders and auctioneer, while in the latter, multiple people perform the chant simultaneously in the context of a formal ritual with a specific purpose that arises in response to particular circumstances. Concentration on the physical characteristics of the acoustic signal alone serves to divorce the activity from the purposes and identities of the participants.

The examples adduced here could be multiplied without difficulty. Importantly, any experimental approach to studying speech and song that begins with the pre-theoretical assumption that there are two distinct categories is unlikely to establish commonalities or continuities that threaten that assumption. Such work will rather serve to reinforce the categorical separation between speech and song.

Joint Speech as an Organizing Frame

Human vocalization is riotously varied. If one could adopt the disinterested perspective of a nonlinguistic, interplanetary observer studying our species as a naturalist might observe a songbird, the conclusion would have to be drawn that vocal signaling and vocal coordination is an inalienable part of the behavioral repertoire of the species. We vocalize in public and in private, in dyads, throngs, and even when alone. Vocalizations manifestly structure the many forms of interpersonal coordination that give rise to all societies. Much of the patterning of vocal interaction bears the stamp of local communities; indeed, one of the ways in which we might identify such communities in the first place is precisely the relations of inclusion and exclusion that arise based on shared vocalization repertoires.

Our interplanetary observer might not immediately identify anything as abstract as "language." There is a large leap from manifest patterns of vocal exchange to the reification of an abstract sociopolitical domain such as "English," or "Yoruba." Similarly, the (to us) important link between writing (in all its heavily mediated forms) and speaking might not be obvious either. It would be clear that humans use their voices in all kinds of situations, and it would likewise be plain that the use of the voice plays a very important role in just about all collective activity.

The rich panoply of vocal behaviors would not be easily interpretable. Far from exhibiting two kinds of vocalization (speech and song) there would be a virtually unlimited number of different kinds of vocalization employed in an equally plural array of social settings. As plastic as the voice is, it would also be clear that there is an important association between certain kinds of vocalization and specific kinds of collective activity. The regulated spoken exchanges of a committee meeting look nothing like the shouting in a sports bar during a match. The playful dialogue between a mother and infant does not resemble the use of the voice by the same mother when she argues on the telephone.

The conventional manner of taming this diversity by the recognition of speech as a specific mode of language encourages us to identify and categorize patterns that are independent of the social context in which they occur. Linguists ponder what a sentence such as "John kicked the ball" is made of, without insisting that there be an actual person called John, who has, or does not have, a ball, and they certainly do not feel any professional obligation to discuss the circumstances of the kicking. Even the phoneticians, who have the most direct interest in the physically instantiated activity of speaking, will conduct most of their work in laboratories, using sound-treated anechoic chambers, where paid participants repeat meaningless sentence lists ("I say heed again; I say hood again; I say who'd again," etc.) into a microphone. Speech is understood to be something that can be studied without reference to the social situation in which it occurs. And of course as competent language users, we have no difficulty in maintaining a clear distinction between the worlds of language and music, and hence between speech and song.

Our notional alien observer does not have such categories at hand, and is forced to try to understand human vocalization by observing and interpreting. There is one possible route to systematizing such observations that the observer might adopt, and that we will pursue here. While vocalization occurs in many constellations among many kinds of participants, there are readily

observable occasions in which multiple people produce the same sounds in synchrony, vocalizing in unison. From a strictly empirical point of view, this kind of vocal activity must surely be more obvious, more easy to recognize, than even such conventional staples as words or sentences. The contexts in which such activities occur are highly constrained and an understanding of the role of the voice in such situations cannot ignore the embedding of the act of speaking in a specific kind of context. It might be that selective attention to such joint speech (as we will call it) could be informative about the manner in which the voice is used, in a way that is clearly different from the insights to be gained using our received categories of language and music.

A focus on joint speech will present us with a spectrum of kinds of vocalization that range from examples we would consider to be clearly spoken (e.g., when a group of people stand to swear a collective oath) to others we would perceive immediately as musical (e.g., in unison choral singing), with many identifiable intermediate points along this continuum. The traditional elements of musical theory, melody, rhythm, and harmony, will make distinct entrances as we move from clearly spoken to clearly sung, and this will allow us to pay attention to the different roles played by these different elements.[1]

To the classically trained linguist, it might appear odd to focus on this strange hybrid territory in which canonical speech or music are not represented. She will find it odd that there do not appear to be clearly differentiated roles for speakers and listeners, which she will assume to be a prerequisite for understanding speech as the passing of encoded messages. The musicologist who specializes in Western art music will be far from the familiar structures of performances and audiences as we encounter many other forms of collaboration and participation. (The broader field of ethnomusicology is urgently concerned with many and varied kinds of socially embedded vocal and embodied practices that refuse this default framing.) By adopting an empirical stance that does not rely on such conventional distinctions, it is to be hoped that new insights may arise.

As well as allowing us to ignore, for present purposes, any categorical division between language and music, joint speech has the remarkable property of erasing or transcending boundaries between distinct spheres of collective activity that we rather naturally think of as distinct. One domain of activity in which joint speech plays a central role is the domain of ritual, which we will have to delineate somewhat generously to include liturgy, collective prayer, secular celebration, and ceremony, but also the domestic ritual of singing Happy Birthday together. The more obvious examples we can adduce in this domain suggest a seriousness of purpose, as in the collective enunciation of a shared credo, but not all rituals need be so solemn. Defining ritual is a difficult business, and different scholars have attempted to identify a variety of features that might be taken to single out rituals, but there has been little consensus here. With joint speech as our topic, such definitional issues need not concern us, and we may use our simple empirical definition to single out those activities to be considered together.

A second domain jumps out at us as we seek examples of people chorusing in unison; this is the more raucous and improvised world of protest. Wherever crowds gather to object, protest, and insist, they chant in unison, frequently augmenting the chants with drums, fist pumping, and the like. From prayer to protest is quite a leap, but one that arises by using the empirical definition of joint speech to focus our observations. And with that, a third domain comes into view that is yet again qualitatively different: In sports and similar forms of quasi-tribal activity, chanting is frequently adopted as a means of expressing, or, better, enacting the collective identity of the participants. Not all sports have chanting traditions. Soccer does, and chanting and singing may alternate during a single match, while rugby does not typically engender chant, even though it has a unique singing tradition quite its own. Chant is foreign to tennis, snooker, and cricket, but is at home with American football, baseball, and ice hockey. Joint speech arises in many other situations as well, of course, but these hastily sketched domains ensure that we will be confronted with a wide variety of activities that have global extent, varying from place to place, but found in some form or another in every human culture. And so we might recommend to our extra-terrestrial observer who is interested in vocalization that this could be one way to begin to understand what all the chattering, shouting, grunting, and whispering is about.

For us too, there may be some utility in using joint speech, defined as simply as synchronized uttering, as a means of coming at the voice afresh. In refusing the received category boundaries between language and music, and between prayer, protest, and sports, joint speech may point us to a novel way of observing our own activities. It is perhaps telling that the language, human, and social sciences have, with painfully few exceptions, not identified joint speech as a substantive topic in its own right. There are no specialized conferences or journals on chanting, broadly construed, and

[1] As pointed out by a reviewer, many other kinds of speech may display quasimusical properties, as in the beat induced by reading a list aloud. Joint speech is not the only kind of musically tinged speech.

the specializations that do exist are drawn within, rather than across, the boundaries we are here willfully ignoring. There are a few individual studies that are relevant. Von Zimmerman and Richardson (2015) have conducted a study in a social psychological context that suggested that synchronized chanting before a group activity may improve collective performance on a task while boosting group affiliation, and Heaton (1992) conducted interesting observations of the tonal precision of the "air ball" chant in basketball. Cummins (2018) provides a more complete overview of the few cases in which joint speech has attracted the attention of researchers within the sciences.

The Musical Phonetics of Joint Speech

We here review some selected examples that can help to map out the territory between the more speech-like and more sung-like kinds of joint speech. In each case, we will naturally be concerned with the objective properties of the voice, such as pitch, timing, and so on. But we will be especially interested in drawing links between such properties and the kind of situation in which they occur. That is, joint speech will require of us that we attend closely to context, usage, and purpose, all of which are frequently factored out in linguistic or musicological analysis.

A first example will serve to define a speech-like end to the continuum. This is well illustrated by the purely instrumental activity of collectively swearing a formal oath. This is done, for example, when becoming a new citizen of many countries. A concrete example of this will be taken, as it occured at the naturalization ceremony performed on March 11, 2015 in Dublin.[2] In a solemn gathering, potential new citizens are gathered who are called upon to stand and to recite the following text in unison:

> I, <speaker's name>, of <speaker's address>, having applied to the Minister for Justice and Equality for a certificate of naturalization, hereby solemnly declare my fidelity to the Irish nation and my loyalty to the state. I undertake to faithfully observe the laws of the state and to respect its democratic values.

Each speaker inserts their own name and address in the corresponding slot (the similarity to a printed form is not accidental). Even though each person has the text written in front of them, a leader walks them through the recitation, calling out each phrase in turn, before the crowd echoes it back. The acoustic blur that results is not intelligible, least of all when everyone speaks a different name or address, but intelligibility is not a reliable, or necessary, characteristic of joint speech.

This ceremony is instrumental in character. Everyone participating will do so once and once only, and after participating, their legal status will be changed. The recitation of the oath is thus performative in the narrow sense introduced by Austin (1962). Nobody is terribly familiar with the text, and there is certainly no expectation that it be spoken from memory. The prosody of the speech is a consequence of these constraints. There is no musicality at all to the voice. Each short phrase is spoken as a unit, slowly. Intonation contours are labored, and there is no beat. This is the speech end of the speech-song continuum. In general, obviously instrumental use of joint speech in such transformative rituals will feature singular (non-repeated) utterances with a very speech-like prosody, and no overt musical elements, and the resulting sound will have a characteristic lack of strong synchronization.

A second example illustrates a feature of joint speech so ubiquitous that its absence in the first example stands out: repetition. Many chants, from all three principal domains of human activity, are short and are repeated very many times. In this example, disgraced U.S. General Michael Flynn joins in a spontaneous chant of "Lock her up!" as it notoriously was featured during Donald Trump's 2016 election campaign. The three beats of the three words are repeated within a four-beat phrase, with a one-beat rest after each one. The enthusiastic crowd emphasized the beats through manual gestures, by shaking signs held aloft or pumping fists in the air. There was no apparent melodic exaggeration of the basic call though. Each word was produced with a constant pitch level.

Staying in the world of repetitive protest chants, it is worth visiting any one of innumerable examples of chanting during the so-called Arab Spring, in which a wave of popular opposition to autocratic regimes erupted about 2011. The example we will look at is from Cairo, Egypt, in which the cry "Ash-**sha'b** yur**īd** is**qāṭ an**-ni**zām**" (with bold font indicating the placement of stresses) or "The people demand the fall of the regime" played a prominent role. This chant, and several variants thereon, became the clarion call of the popular uprising, and was repeated across the Arab world. The chant is produced with a familiar rhythm that has been used at least since the seminal chant of "¡El **pue**blo

[2] This and subsequent illustrative examples may be found at jointspeech.ucd.ie, where an archive of recordings of joint speech is being assembled. Specific examples used in this text may be found by searching using the tag "speech-to-song." It is to be hoped, though, that the specific examples considered here are sufficiently representative of the practices they illuminate to allow them to be readily substituted for by other examples matching the descriptive frame.

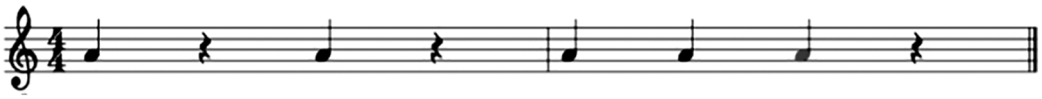

FIGURE 1. Rhythmic pattern used in many protest chants, and traceable back to at least Chile of the 1970s.

u**ni**do, jam**ás** se**rá** ven**ci**do!" ("The people, united, will never be defeated!") that arose in Chile in the 1970's. Here we have not only a simple repeating structure, but a metrical structure, as shown in Figure 1.

When a short phrase is repeated over and over, and that repetition is accentuated by gestures such as fist pumping, it is unsurprising that the beats find organization into larger metrical units. These simple structures are easy to recognize and to join in with. Good leaders of political protests will ensure that their specific demands are expressed within this kind of template as, for example, in "What do we want? Free education! When do we want it? Now!" where the slot filled by "free education" may, of course, be arbitrarily substituted for.

We also find beat without meter, but to do so, we must move from the raucous world of improvised chanting in the street to the more austere world of Gregorian chant, or more broadly, plainsong. This is a style of singing associated with monastic orders and Christian liturgies, with a history extending back to the earliest formative centuries of Christianity. Although several stylistic subtypes may be identified, they have in common the use of a single melody without counterpoint, and minimal or no use of instrumental accompaniment. In most forms, there is a one-to-one mapping between syllables and notes, and the length of individual phrases is determined by the underlying text, though we might note that the question of whether a beat is present, and exactly how notes relate to an underlying score have remained active sites of discussion over many years (Apel, 1990). For the most part, this means that the sequence of accents is somewhat irregular, and there is no overt organization into bars and larger metrical units. Gregorian chant employs a distinct form of musical notation, using four lines rather than five, but notably without overt subdivision into bars.

Repetition may also serve to shift the perception of the intonation pattern of the repeated phrase from a speech-like form to that of melody, a perceptual shift identified by Deutsch, Henthorn, and Lapidis (2011).

Football chanting is often song-like, and fully song-like elements may alternate with shorter assertive chants that more closely resemble the protest forms noted above. Not infrequently, chants repurpose well-known phrases from popular songs. For example, the bass line of the song "Seven Nation Army" by The White Stripes has been used as the template for very many localized chants in soccer, baseball, and American Football (at least).

The quasi-musical nature of the familiar Happy Birthday ritual provides an illustration of the strong link between a specific vocal form, and the attendant interpersonal context in which it occurs. To regard the singing of Happy Birthday as a kind of musical performance would be both curiously insensitive to musical considerations and would miss the fact that participation in the ritual provides its own justification. The melody is frequently sung in multiple keys simultaneously, without regard for aesthetics or virtuosity. It would be folly indeed to record most recitations of the "song." Its purpose and form speak instead to the collective articulation of a shared perspective—in this instance with a singular focus on celebrating a landmark in the life of the birthday celebrant.

Choral forms that include multiple voices or rich harmonic accompaniment by multiple instruments seem to belong firmly at the musical end of things. Choirs are often large ensembles of singers, allowing many people to participate at once, but once all these strongly musical elements are in place, we are more likely to find a distinction between musicians/singers and audience, and the activity acquires the characteristics of a performance, rather than the enactment of a collective purpose.

The organizing framework of joint speech thus presents us with examples of vocal activity that extend broadly between speech/language and song/music. Different dimensions of musical organization such as beats, meter, melody, harmony, and instrumentation may be combined with the voice in very many combinations, but such combination is usually specific to the social or intersubjective context in which a particular kind of organizing activity takes place. Neither linguistic nor musicological categories seem to be of much help in understanding this intertwining of form, activity, and context. Rather, it seems, to understand joint speech we need to recognize that participatory ritual activities—whether venerable and fossilized in liturgy or improvised in revolution—provide access to means by which collective identities are made manifest. We have already spoken several times of the enactment of an identity, and the technical vocabulary of enaction, as

one current theoretical strand within embodied approaches to behavior and cognition, may indeed provide many relevant concepts that might be brought to bear on these foundational activities (Cummins, 2014b; Froese & Di Paolo, 2011).

Historical Considerations

Chant—or joint speech in all its forms and attendant rituals—has largely evaded the scrutiny of objective inquiry. But it has never been absent from human intercourse, nor did it emerge recently. The foundational scriptural texts of many major religious traditions are routinely chanted and have been chanted for millennia. Even today, the authoritative version of the Vedas—the founding scriptural texts of many Indian religious traditions—is provided by the chanted, not the written version. Vedic chanting extends back over 3,500 years. Initiates are taught not only the sequence of words, but complex combinatorial recombinations of the constituent syllables, separating meaning from form so that in collective repetition, any error must be immediately evident. The tradition of Vedic chanting is inscribed on the Representative List of the Intangible Cultural Heritage of Humanity by UNESCO (2008). Zoroastrian and Hebrew scriptures are also routinely chanted, while the Quran is regarded as the *spoken* word of Allah and is thus untranslatable. Indeed, once one views scripture through this lens, it becomes clear that the Bible is unique in its freedom from the occasion of uttering, and in its relation to writing. Among these scriptural texts, only the Bible can be freely translated. This circumstance has had a recursive influence on the very conceptual definition of language itself, as much scientific work that seeks to identify and document individual languages has been carried out by a faith based organization[3] for whom a language is precisely that which can admit of a biblical translation. Were we not from a Biblical tradition, it might be easier to observe the pervasive role of collective utterance in the foundation of many human societies.

The development of writing in the Mesopotamian realm about 3,000 BCE began with the use of labels, tally marks, and lists. The first texts in this tradition that might be called literature date to approximately 2,600 BCE, and include a remarkable liturgical lyric called the Kesh Temple Hymn. The text of this has been found on tablets such as that shown in Figure 2 that extend over the following thousand years, making this a remarkably stable and widespread text. When we examine the text, we find a familiar structure, despite its antiquity. It consists of verses of matched length, and each verse ends with an identical set of phrases, shown in italics in the following excerpt:

FIGURE 2. Kesh Temple Hymn (Image courtesy of Walters Art Museum, Baltimore, Maryland).

> **Verse 3:** House, great enclosure, reaching to the heavens, great, true house, reaching to the heavens! House, great crown reaching to the heavens, house, rainbow reaching to the heavens! House whose diadem extends into the midst of the heavens, whose foundations are fixed in the abzu, whose shade covers all lands! House founded by An, praised by Enlil, given an oracle by Mother Nintur! House Keš, green in its fruit! *Will anyone else bring forth something as great as Keš? Will any other mother ever give birth to someone as great as its hero Ašgi? Who has ever seen anyone as great as its lady Nintur?*
>
> **Verse 4:** House, 10 šar at its upper end, five šar at its lower end; house, 10 bur at its upper end, five bur at its lower end! House, at its upper end a bison, at its lower end a stag; house, at its upper end a wild sheep, at its lower end a deer; house, at its upper end a dappled wild sheep, at its lower end a beautiful deer! House, at its upper end green as a snake-eater bird, at its lower end floating on the water like a pelican! House, at its upper end rising like the sun, at its lower end spreading like the moonlight; house, at its upper end a warrior mace, at its lower end a battle-axe; house, at its upper end a mountain, at its lower end a spring!

[3] The Summer Institute of Linguistics, and its offshoot, the Ethnologue database.

> House, at its upper end threefold indeed! *Will anyone else bring forth something as great as Keš? Will any other mother ever give birth to someone as great as its hero Ašgi? Who has ever seen anyone as great as its lady Nintur?*

The use of a repeated chorus at the end of each verse of a hymn integrated into a liturgical structure is strong circumstantial evidence that joint speech played a central role in the ritual basis of this ancient society too.

The advent of writing profoundly changed those societies in which it occurred. Widespread literacy and an abundance of easily reproduced texts are two relatively recent developments that lead us to single out some characteristics of our vocally structured social lives, and to distinguish them from others. Notably, those elements that admit of symbolization and removal from the context of uttering come to be seen as constituting a separate, linguistic domain. Many of the cognitive and social consequences of this transition have been documented by McLuhan, Ong, Olson, and others (McLuhan, 1994; Olson, 1996; Ong, 2013). It is important to recognize that writing and literacy induced such changes, but it is also essential to note that many forms of vocalization persisted that never made it onto the page. Joint speech surely may lay claim to being an integral part of language broadly conceived, yet it does not find expression in writing, but in the collective participatory uttering. It persists in the most highly technologized societies as well as the least. In this respect, at least, the contrast drawn between so-called "oral" and "literate" societies in the work of Ong, McLuhan, and others appears rather too crisp.

For music too, the media landscape has been dramatically altered by the development of recording, broadcasting, and playback technologies that have conspired to change the very notion of music from an activity in which one participated, to a kind of product that can be indifferently packaged, sold, and stored, irrespective of the context in which it originates. Participation remains an obvious characteristic of music that is integrated into ritual and rite, from Protestant hymn singing to the ecstasies of Hindu kirtan. Participation demands some sort of familiarity, especially if one is to join in singing or chanting. Participation through dancing or coupled gestures is a relatively simple affair if music has a strong metrical structure. Indeed, tapping along with a beat is frequently an unconscious response. This draws our attention to an interesting way in which the musicality of an activity might be related to its subjective earnestness. When we administer a formal oath, or recite a solemn credo, musical elements are largely lacking. Participation thus demands knowledge of the words spoken, and a responsibility to vouch for those words arises in the act for the speaker. As musical elements are gradually introduced, through repetition, rhythmic exaggeration, metrical phrasing, and instrumental accompaniment, participation may become more lightweight, and the associated activities may allow for the use of music and participation in less formal circumstances. To mouth the words of a pop song on the dance floor is not to acquire any commitment to the sentiments they express.

Along with the development of such mediated forms of music, the meaning of the term itself has changed, giving the word "music" a specific sense in popular discourse that includes sounds generated by machine and inscribed directly into digital files for broadcast, without ever having being touched or produced by actual humans, but that bears a tenuous relation, at best, to specific, culturally local forms of participatory practice, such as the Hakka of New Zealand or the Happy Birthday of Europe and America. It is worth noting that the austere religious forms of Islam that disapprove of music making in general distinguish clearly between music for entertainment's sake and melodic unison chant. In the strictest forms of Wahabi Islam, found among the most conservative elements in Saudi Arabia, the call to prayer—or *adhan*—continues to be melodically intoned from the minarets, even as instruments are burned. Islamic State propaganda routinely uses unison chant as its aural backdrop, yet the same group professes to despise music and to persecute those who sing and dance for the mere pleasure of it. The boundaries of music are no more determinate than those of speech and song.

Many of the entrenched categorical distinctions that serve to drive a wedge between the domains of language and music may thus be viewed in a novel manner by using joint speech as an empirical focus. There is a much larger story to be told here, as considerations of chanting and associated practices are brought to bear upon the very terms with which we think of language and the establishment of those bonds that ground specific communities. The media landscape is evolving in accelerating fashion, and with this seething change, the very notion of any monolithic community identity may be largely a quaint anachronism. Under these circumstances, the recognition of the continuous thread of joint speech—which reaches back before even writing—may be a useful corrective to existing theories of language which have been informed by conceptual accounts drawing on the metaphor of message passing between individuals rather than the grounding of collectives.

Speaking, Singing, and Gesturing in Ritual

Joint speech is a reliable feature of ritual, whether that be austere liturgy, the tribal experience of a local football match, or a formal pledge of allegiance to a secular authority. In such contexts, the unison uttering is typically paired with stylized and synchronized gestures of various kinds. Those rituals that bear repetition typically span the range from purely spoken joint speech, through rhythmically accentuated repeated elements to fully melodic chant. Given the centrality of ritual throughout history, one may reasonably suggest that such practices have been foundational in most, if not all, human societies. It is particularly odd, then, that the scientific and academic treatment of language has ignored this kind of vocal activity completely.

One of the few authors to recognize the centrality of synchronized speech and gesture in ritual is Rappaport (1999) who chose to define ritual as, "the performance of more or less invariant sequences of formal acts and utterances not entirely encoded by the performers" (Rappaport, 1999, p. 24). In noting that the texts that are recited are typically not authored by those present, he has put his finger, I believe, on the principal reason why such a ubiquitous and foundational human behavior has managed to hide in plain sight, as it were, for so long. Ritual, as Bloch, Rappaport, and others have pointed out, resists interpretation in symbolic terms, even though such an approach has been attempted by many (e.g., Bettelheim, 1954; Turner, 1970, both cited in Bloch 1974). Symbolic communication has long been seen as a form of message passing, based on processes of encoding and decoding. This metaphor (and it is a metaphor) underlies almost all of the scientific treatment of language, serving to drive a clear wedge between language and music. Placing message passing at the heart of language means that there must be senders and receivers, or, in a face to face situation, speakers and listeners. Yet in joint speech, these distinctions fail to apply. Everybody is both. Everybody knows the text (for it was authored elsewhere), and there is no obvious recipient. If one strains and suggests that a deity or supernatural entity is being addressed, that would still fail to account for the fact that such acts are repeated over and over again, thus nullifying any supposed informational value.

McGraw clearly articulates a distinction between a symbolic reading of ritual and an enactive one (Cummins, 2013; McGraw, 2016). In an enactive account, participation in collective activity is one manner in which collective identities are brought into being. He quotes Houseman (2008) thus: "Rituals do not tell stories; they enact particular realities." One might go further. If rituals can bring realities into being, and thereby create and sustain collective identities, it is but a small step to the observation, by Durkheim, that, " . . . there are rites without gods, and indeed rites from which gods derive" (Durkheim, 1912/1976).

Synchronized activity, in which everybody does the same thing at the same time, is but an extreme form of coordination and collaboration. If a group of people work closely together to hunt a mammoth or build a village, there will be differentiation in the tasks that arise, but the shared goal will suffice to ensure that all those who participate experience themselves as part of a collective, and not as lone workers. In ritual, this participation seems, itself, to be a significant goal, and the stereotyped activities serve as proxy goals, not to be achieved through work and effort, but to allow participation itself. This altered state of affairs is well described by McGraw as a "displacement of intentionality" and it may lead to an altered relation to the notion of authorship that is conducive to altered affective states.

It would be difficult to overestimate the efficacy of such apparently pointless activities. Through the enactment of a collective, the common ground is established from which the world will thereafter be addressed. Relations of trust necessarily arise. Of the many speech acts that joint speech can achieve, lying seems to be a virtual impossibility, precisely because there is no transfer of information from one knowledgeable party to another, who might be duped. If participation in ritual establishes who "we" are, for some value of "we," it must also bring into being the other, who is differentiated from us precisely by non-participation. This essential intertwining of alterity and religious experience generally considered has been noted by some (Csordas et al., 2004).

Speculative considerations about the pre-history of joint speech and collective vocalization is necessarily on shakier ground. In a recent article, Knight and Lewis (2017) argued that choral singing—as found, for example, among the BaYaka pygmies of central Africa—may have its origins not in the construction of symbolic propositions, nor in the pursuit of entertainment, but in the need to project an auditory signal to ward off predators, while simultaneously establishing a form of coalition that establishes who is within and who outside the group.

Concluding Remarks

Defining joint speech is remarkably easy, and although there may be borderline cases (e.g., stylized synchronized breathing patterns in Sufi *dhikr* ritual), it is trivially easy to assemble many uncontroversial examples of people saying or singing the same thing at the same

time. When we do so, we have objectively picked out a suite of activities that are central to the enactment and sustenance of collective identities of many kinds, of a tribe or people, of a congregation, of team supporters, or of political allies, for example. Such activities, I have argued, are worthy of study as a class, and have much to offer the attentive observer.

A remarkable feature of joint speech is the manner in which it erases any principled boundary between language and music, or more narrowly, between speech and song. The various ways in which rhythm and melody arise as a function of repetition, gesture, and stylization demonstrate a non-arbitrary relation between form and behavior, even though such vocal behavior stubbornly resists interpretation within a symbolic or message passing framework. In accord with the enactive account tentatively suggested here, repetition and rhythmicity may play distinguished roles in facilitating participation in the attendant activities that bring into being the commonality of the group. These considerations might readily prompt the attentive observer to combine accounts of chanting along with other context-bound features such as synchronized gestures, to augment the more usual symbolic indices used to characterize the formal activities that ground collectives.

The observations made here and in related works (Cummins, 2013, 2014a, 2014b, 2018) represent a starting point from which a great deal remains to be explored. We know, for example, that not all crowds who chant in the streets are driven by the same forces. Where some crowds want to overthrow the authorities, others are paid by them. Does this affect the chanting? We do not know, but could empirically investigate, the relation between spontaneous collective chanting and the subsequent occurrence of violence in situations of public disorder. A great deal remains to be studied in the use of joint speech in educational contexts, from madrassas to primary schools, where teachers employ chanting for a wide variety of purposes, including memorization, pronunciation training, and the simple marshaling of collective attention. Joint speech offers a novel means of framing many empirical research questions of broad applicability. There is work to be done.

But the considerations that arise in the study of joint speech in context suggest something further. The language sciences have existed at a peculiar remove from the social and human sciences broadly considered. The manner in which "language" has been framed has isolated an abstract system from the complexities of situated occurrence among collectives of many kinds. This approach has found particular traction in an intellectual landscape characterized by a strong belief in the autonomous individual, most notably expressed in the models of contemporary scientific psychology. When language is framed in this manner, it appears as something categorically removed from the participatory activities that give rise to music and ritual. Joint speech draws our attention back to the use of the voice in context. It raises the tantalizing possibility that much of the efficacy of language in forming and maintaining collectivities of various kinds may be revealed if we re-insert the utterance into the world in which it functions. The apparent invisibility of joint speech in the study of speech and language may be readily explained by the failure of such activities to match a template in which individuals are hermetically closed islands, communicating by encoded messages.

Author Note

I owe a particular debt of gratitude to the editors of *Music Perception* for being willing to consider a creative and constructive way of dealing with an article that relies so much on anecdote and informal observation, without a formal empirical structure. Their willingness to adopt a novel format of commentary and response greatly enriches the topic in a way that is very well aligned with my intentions in writing it. My thanks also to Frank Russo and two anonymous reviewers whose observations served to improve the present text and to kick-start what I hope will be productive and extended discussions around the topic of joint speech.

Correspondence concerning this article should be addressed to Fred Cummins, UCD School of Computer Science, University College Dublin, Belfield, Dublin 4, IRELAND. E-mail: fred.cummins@ucd.ie

References

Apel, W. (1990). *Gregorian chant*. Bloomington, IN: Indiana University Press.

Austin, J. L. (1962). *How to do things with words*. Cambridge, MA: Harvard University Press.

Bettelheim, B. (1954). *Symbolic wounds: Puberty rites and the envious male*. London, UK: Thames and Hudson.

Binazzi, B., Lanini, B., Bianchi, R., Romagnoli, I., Nerini, M., Gigliotti, F., et al. (2006). Breathing pattern and kinematics in normal subjects during speech, singing and loud whispering. *Acta Physiologica*, *186*(3), 233–246.

Callan, D. E., Kawato, M., Parsons, L., & Turner, R. (2007). Speech and song: The role of the cerebellum. *The Cerebellum*, *6*(4), 321–327.

Callan, D. E., Tsytsarev, V., Hanakawa, T., Callan, A. M., Katsuhara, M., Fukuyama, H., & Turner, R. (2006). Song and speech: Brain regions involved with perception and covert production. *Neuroimage*, *31*(3), 1327–1342.

Csordas, T., Bowie, F., Dupr, M., Hauser, B., Lambek, M., Littlewood, R., et al. (2004). Asymptote of the ineffable: Embodiment, alterity, and the theory of religion. *Current Anthropology*, *45*(2), 163–185.

Cummins, F. (2013). Towards an enactive account of action: Speaking and joint speaking as exemplary domains. *Adaptive Behavior*, *13*(3), 178–186.

Cummins, F. (2014a). The remarkable unremarkableness of joint speech. In *Proceedings of the 10th International Seminar on Speech Production* (pp. 73–77). Cologne, DE: ISSP.

Cummins, F. (2014b). Voice, (inter-)subjectivity, and real time recurrent interaction. *Frontiers in Psychology*, *5*(760).

Cummins, F. (2018). *The ground from which we speak: Joint speech and the collective subject*. Newcastle upon Tyne, United Kingdom: Cambridge Scholars.

de Saussure, F. (2011). *Course in general linguistics*. New York: Columbia University Press.

Deutsch, D., Henthorn, T., & Lapidis, R. (2011). Illusory transformation from speech to song. *Journal of the Acoustical Society of America*, *129*, 2245–2252.

Durkheim, E. (1976). *The elementary forms of the religious life*. Abingdon, United Kingdom: Routledge. (Original work published 1912)

Froese, T., & Di Paolo, E. A. (2011). The enactive approach: Theoretical sketches from cell to society. *Pragmatics and Cognition*, *19*(1), 1–36.

Heaton, C. P. (1992). Air ball: Spontaneous large-group precision chanting. *Popular Music and Society*, *16*(1), 81–83.

Houseman, M. (2006). Relationality. In J. Kreinath, J. A. M. Snoek, & M. Stausberg (Eds.), *Theorizing rituals* (pp. 413–428). Leiden, Netherlands: Brill Publishers.

Jakobson, R., Fant, G., & Halle, M. (1951). *Preliminaries to speech analysis. The distinctive features and their correlates.* Cambridge, MA: MIT Press.

Knight, C., & Lewis, J. (2017). Wild voices: Mimicry, reversal, metaphor and the emergence of language. *Current Anthropology, 58*(4), 435–453.

List, G. (1963). The boundaries of speech and song. *Ethnomusicology*, *7*(1), 1–16.

McGraw, J. J. (2016). Doing rituals: An enactivist reading of Durkheim's elementary forms. *Intellectica*, *63*(1), 37–48.

McLuhan, M. (1994). *Understanding media: The extensions of man.* Cambridge, MA: MIT Press.

Olson, D. R. (1996). *The world on paper: The conceptual and cognitive implications of writing and reading.* Cambridge, United Kingdom: Cambridge University Press.

Ong, W. J. (2013). *Orality and literacy.* Abingdon, United Kingdom: Routledge.

Pinker, S. (1999). *How the mind works*. New York: W. W. Norton & Company.

Port, R. F. (2007). The graphical basis of phones and phonemes. In O. Bohn & M. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 349–365). Berlin, Germany: De Gruyter.

Proctor, D. F. (2013). *Breathing, speech, and song.* Berlin, Germany: Springer Science and Business Media.

Rappaport, R. A. (1999). *Ritual and religion in the making of humanity* (Vol. 110, Cambridge Studies in Social and Cultural Anthropology). Cambridge, United Kingdom: Cambridge University Press.

Riecker, A., Ackermann, H., Wildgruber, D., Dogil, G., & Grodd, W. (2000). Opposite hemispheric lateralization effects during speaking and singing at motor cortex, insula and cerebellum. *Neuroreport*, *11*(9), 1997–2000.

Turner, V. (1970). Symbols in Ndembu ritual. In D. M. Emmet (Ed.), *Sociological theory and philosophical analysis* (pp. 150–182). New York: Macmillan.

Von Zimmerman, J., & Richardson, D. C. (2015). Verbal synchrony in large groups. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 2523–2528). Austin, TX. Cognitive Science Society.

Yang, F. (2015). A study on the features of chest and abdominal breathing between reciting and chanting Chinese poetry. *Journal of Chinese Linguistics*, *43*(1), 399–410.