Percepta, 1(1), 17–32
©2013 dos autores

ISSN 2318-891X
http://www.abcogmus.org/journals

Joint speech: The missing link between speech and music?

FRED CUMMINS*

Abstract
This paper argues that music and speech are not accidentally related as some have claimed. The argument is made by emphasising the coordinative aspects to speaking and music making, and downplaying the role of message passing with which linguistics has traditionally been concerned. A continuum is identified from silent speech on one hand to full blown music and song on the other. At each point, we see different coordinative forms and opportunities among subjects. Joint speech, where a group of people say the same text at the same time, is identified as an important point on this continuum, linking speech and music. Joint speech is familiar from collective prayer, protest chants, and many other contexts in which group purpose finds expression. An experimental form of joint speech, called synchronous speech, has been studied and some findings are recounted here. However the larger question of how and why joint speaking arises remains to be studied. I present a dynamical systems perspective on the coordination that joint speakers employ and show how it can account for some, but not yet all, aspects of the observed phenomenon.

Keywords: joint speaking, choral speech, synchronous speech, music and speech, chant

Resumo
Este artigo discute que a música e a fala não são acidentalmente relacionadas, como alguns já afirmaram. A discussão enfatiza os aspectos coordenativos de falar e fazer música, e minimiza o papel da comunicação de mensagens (*message passing*) com o qual a linguística tem se preocupado tradicionalmente. Identifica-se um continuum que vai da fala silenciosa, de um lado, até a canção em sua plenitude, de outro. Em cada ponto, vemos diferentes formas e oportunidades coordenativas entre esses objetos. A fala conjunta, quando um grupo de pessoas diz o mesmo texto, ao mesmo tempo, é identificada como uma questão importante nesse continuum, ligando fala e música. A fala conjunta é comum na prece coletiva, em gritos de protesto e em muitos outros contextos em que o propósito do grupo encontra expressão. Uma forma experimental de fala conjunta, denominada fala sincrônica, tem sido estudada e alguns resultados são aqui descritos. No entanto, a questão mais ampla de como e por que surge a fala conjunta precisa ainda ser estudada. Apresento uma perspectiva de sistemas dinâmicos acerca da coordenação empregada pelos indivíduos que falam em conjunto e mostro como ela pode explicar alguns, mas não todos, aspectos do fenômeno observado.

Palavras-chave: fala conjunta, fala coral, fala sincrônica, música e fala, cântico

* UCD School of Computer Science and Informatics - University College Dublin
E-mail: fred.cummins@ucd.ie

Recebido em 17 de junho de 2013; aceito em 24 de junho de 2013.

1 Introduction

From some perspectives, music may appear as an oddity. It serves no obvious function that can be readily described. It is fun, ubiquitous, but to some eyes, of no obvious use. The cognitive psychologist Steven Pinker has famously characterised music as "auditory cheesecake" (Pinker, 1999), and he justifies this by noting that the enjoyment of cheesecake is an epiphenomenon that arises because of our appetites for such energy stores as sugar and fat. The latter, he contends, subserve obvious survival functions and are thus selected for by evolution. Cheesecake, however, is not selected for by evolution. Music, he argues, is similar. It is founded upon, and exploits, auditory functions that themselves serve functions that have survival value. In musical experience, we find a powerful conjunction of stimuli that collectively bring pleasure, but that we could get along without just as well. Cheesecake, artistic expression, and pornography can all be so characterised within this worldview. They push our buttons, but they are useless.

Speech, on the other hand must appear as a very different sort of activity altogether. If we think of speech as the primary vehicle by which language finds expression, and we recognize the centrality of language to the cognitivist view of mind, then we have a phenomenon that is closely associated with such quintessentially human faculties as thought, reason, and intelligence. Staying within the cognitivist framework, speech becomes the observable counterpart to the essence of the human mind. This is far from cheesecake and pornography.

But the cognitivist view of mind is no longer the only game in town. It's central concepts of a monolithic executive cognitive system, and a representational domain restricted to a single individual, are under increasing attack from alternative approaches that are finding widespread interest under such banners as ecological psychology (Gibson, 1986; Chemero, 2009), enaction (Stewart et al., 2010), embodiment (Varela et al., 1991), and dynamical systems (Kelso, 1995) approaches to minds, brains, and behaviour. This is a large clash with many ongoing arguments, many outstanding conceptual issues to be resolved, and many consequences for our understanding of our selves and the world we create. Those battles will be fought elsewhere. However it is important to realise that we have available to us views of cognition, mind, and the relationship between brains and behaviour, that are fundamentally different from the assumptions of orthodox cognitive psychology, and that these alternatives may reveal the relation between speech and music in an entirely different

light—a relationship that appears accidental and uninterpretable on the cognitivist view.

In what follows, I will sketch an account of this relationship that is grounded in a dynamical systems framework. I will argue that the relationship between speech and music is far from accidental, and that an examination of this relationship has the potential to reveal much about how humans co-ordinate their activity, and hence their worlds. By adopting a coordinative view of speech, rather than a traditional linguistic view, speech and music appear as poles on a richly populated continuum. Many distinct points on this continuum can be identified, and each of these is deserving of theoretical and empirical study. But some points on this continuum have hitherto been neglected. In particular, I will argue that the activity of speaking in unison, or joint speech, is an important link between speech and music, traditionally considered, and that examination of joint speaking practices can help to illuminate many facets of collective behaviour that demand non-cognitivist, non-representational description. The study of joint speech thus becomes an important empirical battleground within a much larger clash of foundational views of human mind, human activity, and human experience. The stakes are high.

2 Communication



Figure 1: Left: the tube model of communication. Right: the dance model of communication. Thanks to Tom Froese for this appealing visual description.

We begin by distinguishing between two senses of the term "communication". These are illustrated in Figure 1. On the left, we see a view that has been termed the "tube" model of communication (Maturana and Varela, 1987). Here, communication is understood as the passing of encoded messages. A speaker forms an intent that can be encoded as a string of words. These words are translated from one representational form to another until they result in a set of movements that cause a sequence of perturbations to the air known as sound. These perturbations are sensed by a receiver, and the reverse

process ensues by which the sounds are decoded and translated from sensory to semantic representation, through a number of intermediate stages.

On the right we see a contrasting view that highlights the coordinative nature of communication. We might call this the "dance" model. Here, the sounds, and movements of speaking (and speaking is whole body activity!) serve to yoke together the activities of the communicating partners, linking their behaviour and their experience, and resulting in emergent patterns of coordinated behaviour. Importantly, this view of communication makes reference to observables only—coordination among individuals is evident in the non-independence of their joint activity—and it does not depend upon an unknown and unknowable theatre of mind in which meanings originate. At the level of description, two coordinated systems can be described using many less numbers than two similar, but uncoordinated systems, and this parsimony is evidence of their mutual linkage. Where the tube model assumes distinct mental and physical realms, the dance model does not distinguish between an unobservable, "inner", domain and an observable, "outer" one.

The paradigmatic act of speaking within the tube model is the statement of a proposition. Propositions, and propositional content, are indisputably important in human affairs, and the positing of an alternative perspective should not detract from that simple observation. However many, if not most, acts of speaking are not of that nature. The phatic communication at the water cooler, as we exchange ritualised but content-free greetings, makes no sense under the tube model. Nor do the back channels that serve to sustain the conversational linkage between two conversing partners. "Uh-huh", "mmmm" and the like are not well described as encoded messages, but they have an obvious role to play in the orchestration of joint activity among interacting people. Another example that will be of central importance here is the collective recitation of a common text by a group of people. This occurs in prayer, in protest, in swearing public oaths, and in the chants of supporters at football matches. In each case, most, if not all, of the listeners are also speaking the same text, so the tube, or message-passing, interpretation of speaking seems again to be uninformative about these behaviours.

A similar dichotomy can be established with respect to music. Music, as Pinker speaks of it, is consumed. The cheesecake metaphor makes this quite explicit. This is one way of looking at some musical activities, and is perhaps particularly appropriate for performative and recorded forms of music. But many forms of music making are very different. Turino helpfully distinguishes between participatory

and performative forms of music making (Turino, 2008). The former refers to music making and dancing that takes place in a shared space in which there is no clear division between producer and consumer. Turino provides numerous illustrations from Zimbabwe, Peru and the USA, but such forms are probably found across the globe. They include many forms of folk music, and typically serve important social bonding functions. Performative music making is rather different, and serves different functions. It is presumably also a newer form of musical activity, as its widespread presence is supported by such cultural developments as concert halls, recording studios, and radio, all of which are fairly recent developments. There is an interesting parallel to be drawn here between participatory and performative forms of speaking, with most attention having paid to the performance and to speech as a product, while the participatory forms of speaking have largely been overlooked.

3 A Continuum from Speech to Music

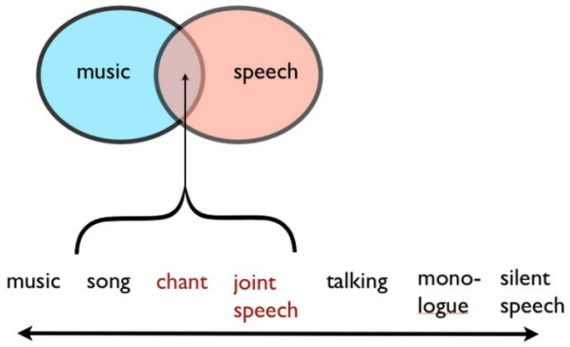

Figure 2: A hypothetical continuum from inner, silent speech to music.

Figure 2 illustrates a hypothetical, and simplistic, continuum from inner, silent speech on the one hand, to music proper on the other. We begin our discussion on the right hand end of the figure, with silent speech.

Silent, or inner, speech bears a direct relationship to overt, public speech. As Vygotsky has pointed out (Vygotsky et al., 2012), young children literally think aloud, and only develop the inhibition required for an inner voice after they have mastered the act of speaking. There has been extensive study of the neural and muscular involvement common to both silent and overt speech (Frackowiak and Frith, 1996). From a coordinative point of view, where we shun any inner/outer distinction, this represents an individualistic extreme as no real-time coordination among individuals is manifest at all.

Between inner speech and conversational speech (mere talking) I have chosen to highlight monologue, which is an asymmetrical form of public speech in which one person speaks while others, typically a

crowd, listens. This form is familiar from preaching, lecturing, and rhetorical displays. An extreme form of this, in which any feedback from listener to speaker is virtually impossible, is found in the transmission of news broadcasts and speeches through radio, TV and other forms of media. When listeners are live, there may well be a great deal of reciprocal coordination, particularly when the speaker is engaging (or provocative) and the audience enthralled (or horrified).

The next point on the continuum refers to the many and varied forms of conversational exchange in which multiple speakers participate. Here, no single speaker dictates the flow of speech. The "floor" passes from one speaker to the other, and meaning emerges in the process of reciprocal exchange. The coordinative nature of speech is more evident now, as each participant is continuously influenced by, and in turn influences, the ongoing activities of the other. There are many and varied kinds of conversations, ranging from the intimate whispers of lovers to the shouts of combatants, and it is apparent that at this point, as at all others, the notion of a continuum is only a guide that can help us perceive some ways in which speech varies, but that should not blind us to the richness and variety of spoken behaviour.

Moving further along, we encounter joint speech. This is speech in which multiple speakers repeat the same text at the same time. Clearly, such speakers are more highly coordinated than speakers in previous forms of speech we have considered. The occasions in which such speaking naturally arises provide strong indication that we are dealing with the expression, not of individual purpose, but of group purpose, group intention, and group sentiment. Collective prayer is one important and highly familiar point of reference. Collective prayer is ubiquitous, and is found in virtually all religious traditions in some form or another, from Tibetan chant to Catholic recital of the rosary. Another, equally familiar, situation in which such speaking happens is in protest and demonstration. Participants in social movement "on the street" readily turn to joint speech to make their demands heard. Both prayer and protest frequently involve not only collective speaking, but the insistent repetition of a short text, often hundreds of times in a row. The chants of sports supporters represent yet another extremely widespread and common example. Other occasions in which joint speech is common are more performative in nature. Groups of citizens taking a collective oath to a secular authority frequently do so in unison. School children are often required to recite selected and valued texts as a single group. In-

deed, joint speech in schools serves several purposes, including performance, memorization, and pronunciation training.

The English term "chant" is ambiguous, as it can refer both to the spoken, repeated, demands of protesters, and to a type of very plain music, in which melody is sparse, instrumental accompaniment is absent or minimal, and meter is not obviously present. By meter, I mean the organization of prominent temporal events into hierarchical repetitive structures such as bars and larger units. There are many varieties of chant, mostly found within religious traditions. Because of the absence of meter, the length of individual phrases is highly variable, and more closely resembles prose than poetry. Sung chant thus represents a point on the continuum at which musical elements appear, but it retains many of the characteristics of speech, and more particularly, of joint speech.

As we move further along the continuum, towards more complex and structured integration of voice and music, we encounter a profuse richness of form, in which the extraordinary flexibility of the human voice is married in every way imaginable to the rhythmic and melodic forms of music, generating such genres as rap, scat singing, throat singing, beat boxing, and the innumerable varieties of song.

Laid out in this fashion, the coordinative perspective serves to emphasise the commonalities that arise in music making and speaking, and to illuminate the variety of ways in which the voice can serve both individual and group purposes. To consider speech as merely message passing is to be blind to this landscape. Moreover, we might note that the relationship between silent, or "inner", speech and overt conversational speech has exercised the minds of the great in many disciplines. It has played a very significant role in the development of the computational theory of mind, providing the substance of Fodor's Language of Thought hypothesis (Fodor, 1975), and making the study of generative grammar central to the study of cognition. We noted the centrality of the relation also in Vygotsky's theory of childhood development. There is a great deal of neuroscience and psychophysical experimentation that has sought to identify commonalities and differences between these two modes of speech-like behaviour.

In contrast, there is almost no scientific field that has addressed the relation between conversational speech and joint speech. There have been occasional works that sought to provide guidance to teachers on how to use joint recitation to improve the pronunciation of school students. There is a small and highly specialised literature devoted to the use of joint speaking (with a tape recording) to stabilise the speech of stutterers (Kalinowski and Saltuklaroglu, 2003), but

there is very little that asks scientific questions of a more general kind about the form of joint speaking activities, and nothing at all that asks scientific questions about the significance thereof. Perhaps this observation might serve to draw our attention to the degree to which modern cognitive psychology has focused almost exclusively on the individual, at the expense of failing to recognise the many ways in which our moment to moment activity is comprehensible only with reference to the activity of others. Psychological theory seems to have no difficulty in attributing unseen and unseeable motivations, intentions, and beliefs to individuals, but it does not encourage any such attribution to groups or collectives, unless the attribution is clearly marked as merely metaphorical. Minds, within latter day psychology, are singular, even solipsistic, domains. Joint speaking, and music making, are unlikely to attract much attention within such a framework. This, in turn, suggests that joint speaking may be a rich and productive domain for scientists to investigate who wish to go beyond or around the limitations of Cartesian and purely individualistic approaches to mind.

4 Synchronous Speech

The term "Joint Speaking" has been introduced here as an umbrella term, covering a variety of forms of speech that have in common the recitation of a single text in unison. Other, more specific, terms used include chant, choral speaking, and recitation. One very constrained form of joint speaking has been introduced in my own experimental work since about 2002. I call joint speech elicited in this experimental context "Synchronous Speech", in order to differentiate it from the other varieties of joint speaking. In Synchronous Speaking, two subjects are provided with a novel text, which they are allowed to read through, silently, first. On a signal from the experimenter, then, they read the text, attempting to remain in synchrony with one another. Subjects typically have no difficulty at all in following these minimal instructions, and they manage to produce highly synchronous utterances (Cummins, 2002; Cummins, 2003; Cummins, 2009).

Before briefly recounting the findings of these studies, it is worth noticing some differences between this experimental situation and other, ethologically situated, forms of joint speech. First and foremost, the texts used bear no significance for the speakers. They do not express any joint belief or purpose. Texts used are often of the kind used by phoneticians in other studies, such as the famous "North Wind and the Sun" or the "Rainbow" passages. When observed "in the wild", texts found in joint speaking are inevitably

emotionally effective strong expressions of group purpose or belief. A second major difference lies in the requirement to remain in synchrony. Informal observation of joint prayer in churches and temples suggests that speakers have a great degree of tolerance for imprecise temporal alignment. Speakers are often only loosely synchronized, and the strongly reverberant characteristics of the surrounding architecture (domes!) may even exaggerate the acoustic imprecision, making it difficult to differentiate between individual voices, and thereby creating a communal acoustic blur, in which the individual is lost at the expense of the collective. This common architectural feature of spaces of worship is probably a design feature, rather than an accident. In the experimental situation, tight synchrony is an explicit goal, and the degree of mutual coordination observed is considerably higher than found in the wild.

In studying the process of synchronization, we have found that speakers are not only excellent at remaining in tight synchrony with one another, they do not even get notably better with practice (Cummins, 2003). When we compare well defined points in two parallel speech waveforms, we can estimate the average asynchrony between speakers. Typically, we find an asynchrony of approximately 40 ms within a phrase. This value rises to about 60 ms at the onset of a phrase after a pause, suggesting that pauses are of somewhat indeterminate duration. When subjects cannot see each other, this increases uncertainty at phrase onset by about another 20 ms.

In English, the speech produced synchronously is not noticeably different from speech produced by a single speaker reading the same text. The prosody, i.e. the pattern of intonation and of relative timing, seems to be largely unaffected. We have observed that speakers of Mandarin Chinese alter their prosody substantially under similar circumstances, producing a list-like form of speech, in which individual syllables are exaggerated, yielding a regular recurrence (Cummins et al., 2013). We suspect that the phonological differences between the languages, particularly with respect to syllable structure and the process of vowel reduction, may underlie this phenomenon. Speaking with equal emphasis on each syllable is an option in Chinese, where there is no categorical difference between stressed and unstressed syllables, no large difference between full and reduced vowels, and syllables are simple CV or CVC sequences. The recurrent, list-like, pattern may serve to stabilise the joint performance, thus enhancing synchrony. English, on the other hand, exhibits alternation between stressed and unstressed forms, with a small set of greatly reduced vowels in unstressed syllables, and full vowels in stressed syllables. Individual syllables vary greatly in complexity, from single vowels

(V) or sonorants to complex multi-consonantal syllables such as "strengths" (CCCVCCC). Falling back on syllabic regularity is thus not an option for English speakers.

A second observation leads us to believe that the regularised reading observed for Mandarin speakers serves to stabilise the joint performance. When reading jointly, and consciously attempting to stay in close synchrony with a co-speaker, we sometimes observe a type of speech error that is unique to this elicitation condition. Once one speaker becomes somewhat uncertain, e.g. after a speech error by either speaker, both speakers will sometimes stop speaking abruptly and at the same time. This remarkable error type recurs with some regularity for English speakers. In a small experiment designed to encourage errors though the occasional use of mismatched texts, we induced 25 such errors with 142 cases of mismatched texts (Cummins et al., 2013). The same manipulation on Mandarin speakers induced only 3 such errors, demonstrating a much greater stability of the dyadic reading than in English.

There are many other questions one might have about the form of synchronous speech. One study currently under way seeks to explore the manner in which the number of speakers affects the character of speech and synchronization; another explores the role of speaker familiarity in synchronization. Much work remains to be done in investigating the way that synchronization varies with phonological systems, to extend the inquiry beyond just English and Mandarin Chinese; this can easily be done by using a paradigm that induces speech errors through mismatched texts. There remains much to explore in the form of synchronous speech, but there is more again to understand when we ask how and why do people synchronize, and how should we understand this ubiquitous behaviour?

5 A Dynamical Perspective and a Puzzle

The voice is a highly expressive instrument. The way in which we speak is sensitively influenced by the context in which we speak, our relation to our co-speaker(s), our purposes, ambient noise, and much more besides. Despite this capacity for variable realization, speakers have no obvious difficulty in constraining their speech to remain in synchrony with another. The relationship between two synchronous speakers is not one of leader/follower, as found, e.g. in the task of speech shadowing (Marslen–Wilson, 1973). Rather, the two speakers seem to be mutually coupled, or entrained, forming a dyadic level of organization that persists as long as they speak together.

The notion of coupling or entrainment (I will use the terms synonymously here) is familiar from the interaction among oscillatory systems. Metronomes, pendulum clocks and other systems that are characterised by a periodic motion of their own, based on an energy source, will subtly alter their motion when allowed to weakly interact, so that the resulting collective motion is simpler than that of the systems considered individually (Pikovsky et al., 2001). This property of oscillating systems was first noticed by Christian Huygens, the Dutch polymath and inventor of the pendulum clock. Two clocks mounted on a common housing were found to fall into a stereotypical pattern in which one pendulum started its cycle just as the other was half way through its cycle. Today we would call this an antiphase coordination. Huygens himself called it an "odd sympathy". Since then, coupling or entrainment among dynamical systems with periodic behaviour has been documented in very diverse fields, from the motion of planets and their satellites, to the joint flashing of fireflies, the synchronous waving of the claws of fiddler crabs, and, of course, the tendency of interacting metronomes to become synchronized. The dynamical principles involved are now well known to be not tied to any specific physical substrate, so that we find similar processes arising from the interaction of animate and inanimate systems. The mathematical treatment of such interacting systems is well developed. Recommended sources include Pikovsky et al. (2001) and Kelso, (1995).

In many respects, two synchronized speakers bear similarities to synchronised oscillators. The absence of a leader/follower relation, along with the sustained temporal coincidence of highly intricate sequential behaviour suggests that there is a bond between the speakers that we might think of as coupling. The unique form of speech error encountered, when two speakers abruptly and simultaneously stop speaking, is further evidence that the dyadic level constitutes a level of emergent systematic organization that is, to some extent, independent of the individual component speakers, as it is the dyad, rather than each speaker separately, that responds to the perturbation induced by a speech error. Elsewhere, I have characterised this as similar to the bond that exists between runners in a three-legged race, who likewise are prone to catastrophic collapse if one person makes an error (Cummins, 2012).

But the synchronization of two simultaneous speakers differs from the synchronization of the planets or fireflies in one crucial respect: the behaviour is not periodic. Speaking is a complex activity in which repetition is rare, and seldom sustained. Most naturally occurring speech, and all speech used in the above synchronous speech

experiments, is simply not periodic. Stresses in English are irregular, and there may be no, one, two, or many unstressed syllables between successive stresses. The syllables in turn will vary greatly in their complexity, and their duration. Any brief predictability will last no more than a few syllables, at best. This does not seem to be an impediment to synchronization.

In light of this, it is perhaps worthwhile comparing the act of synchronous speaking to several other synchronized behaviours humans are capable of. In doing so, we will use a rather strict definition of synchronization, and require that individuals perform the same thing at the same time. Many group behaviours are exquisitely coordinated across individuals, but would fail to meet this strict definition of synchronization. A couple dancing a tango, for example, are clearly yoked together in a joint performance as if driven by the same clock (which is the underlying meaning of the word "synchronization"), but as the man and the woman do different things at any given interval, we omit them from our survey here.

This no doubt unduly strict definition leaves us with relatively few synchronized behaviours to review. The clockwork marching of soldiers is a notable example that does meet the definition. Some sports seek to exaggerate synchronization, as in synchronised swimming, diving, trampolining, and similar. Rowing is likewise highly synchronous, although this seems to be a necessity, rather than an aesthetic choice. An informal but extended review of such activities reveals the following:

- Spectators are tolerant of a considerable degree of asynchrony. Thus synchronization among swimmers is clearly less pronounced that synchronization among divers, but this is accepted by viewers, perhaps as a limitation of the genre.
- Many forms of synchronized behaviour rely upon the presence of a perceptible pulse that establishes a temporal grid. This is true of rowing, unison music making, synchronized dancing, and, to some extent, synchronized trampolining.
- Many forms of synchronization, including those that appear to be most rigorous, involve strong physical coupling between the actor and her environment. Thus in rowing, the rower is yoked together with the large oar, which in turn is vastly constrained by the properties of the water through which it moves. In trampolining and diving, the elasticity of the launch surface, and the non-negotiable influence of gravity all serve to constrain the action and thereby to promote synchronization. Indeed, in synchronized trampolining and diving, there is little or no reciprocal interaction among the actors. They are both caught within a

physical mesh of elasticity and inertia that scaffolds the whole action.

These commonalities among synchronized actions have not previously been noted. Speaking in unison differs interestingly from all of these activities in that (1) there is no periodic referent or beat, and (2) the coordinated action takes place largely without scaffolding by the physical environment. Most movements of speech occur in a protected space, behind the lips, and relatively unaffected by the surfaces and gravitational constraints that surround the speaker. Despite this, sustained and highly accurate synchronization is possible.

This then is the puzzle, and I suspect it is a deep puzzle, and not just a matter of detail. In speaking together, we achieve synchrony without the principal external supports (beat, inertia, gravity) that scaffold all other synchronous behaviours. This has as a consequence that the direct modelling of the temporal coupling between speakers is not immediately possible, as no speaker constitutes a plausible oscillatory system, and the movements that together make up speaking are not simply repetitive. The synchronization we see among speakers is interestingly different from the joint flashing of fireflies, the simultaneous waving of fiddler crab arms, and the resonant orbits of the planets. The difference has something to do with the knowledge that speakers share, knowledge that forms the basis of being a speaker of this or that language.

There is a long-standing debate within phonetics and phonology about the relation between speech production and speech perception. While some accounts would see the two processes as separate, there are both theoretical and empirical reasons to believe that they are inseparably linked. An old hypothesis, known as the Motor Theory of Speech Perception argued that common representations must underlie the two sides of speaking (Liberman and Mattingly, 1985). This theoretical move has found empirical support, of a sort, in the recognition of the intricate linkage between perception and action found within the so-called “mirror system” (Rizzolatti and Craighero, 2004) and in neuroscientific studies that have demonstrated subliminal activation of the tongue and facial muscles in listeners that would be appropriate for speaking (Fadiga et al., 2002). In a recent dynamical model of speech production, we attempted to make the link between perception and production explicit, such that speech was the product of both processes at once (Simko and Cummins, 2011). In doing so, we were formalising the notion that speech results from a tradeoff between production constraints and perceptual constraints, most directly expressed in Lindblom's H&H theory (Lindblom, 1990). All of these theoretical and empirical strands point towards an intricate in-

terweaving of speech perception and production that still needs full explication.

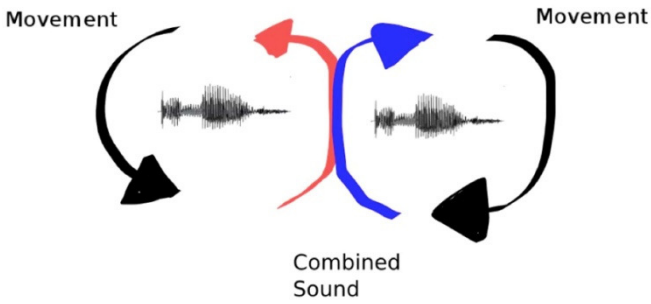

Figure 3: The tight intertwining of speech among synchronous speakers.

The synchronization we observe among speakers represents a further, and novel, contribution to this debate. Figure 3 illustrates one way to think of the coupling among speakers. The physical medium of coupling must be sound, and in the figure, we show the sound produced by the two speakers as additively combined. Each speaker is therefore speaking under a condition in which the normal association of own-movement + own-sound is augmented by the sound of the co-speaker, so that the sound of each speaker plays a constitutive role in the production of speech by both speakers. On this account, the speakers are jointly responsible for the speech that results. The joint speech of a dyad is thus not simply the combination of the speech of one with the speech of the other, just as a handshake is not simply the movement of one hand and the movement of the other.

6 Final Thoughts

Speech is typically not periodic. Most music, however, is. By laying out a continuum between speech and music above, I have tried to suggest that there are many points between the non-coordinative inner speech of the silent thinker and the highly coordinated joint production of musicians and singers. At different points we can see the introduction of elements that facilitate coordination among individuals. In repetitive chant-like speech, we already have rhythmic form emerging, without a consistent musical meter or a sustained beat. In plainsong and similar chant forms, we have speech modified by the introduction of the simplest of melodies. There are many ways in which music and speech combine, and some lean more heavily than others on the coordinative possibilities of each.

For both speech and music represent forms of coordination. Each provides possibilities for the expression of joint understanding. In so doing, each also provides a platform for the expression of individual-

ity. The soloist can be as expressive as she is, precisely because she is backed by an orchestra that provides a unified framework. A single speaker can hold forth in a display of individuality, but it will be effective only if listeners, and potential co-speakers, remain in coordinative bond with the speaker, through the devotion of attention and the provision of feedback, vocal or otherwise.

These commonalities between speech and music help to make sense of the many and various forms in which voice and instrument can be combined. But they demand a coordinative, rather than a psychological, view of the action of both speakers and listeners. I have tried to employ dynamical systems theory as a vocabulary suited for expressing this reality. In the bibliography, significant points of reference for the interested reader are prefixed with an asterisk (*). As we develop scientific vocabularies and frameworks that come at human behaviour and experience from many different viewpoints, we gain a plurality and richness to our understanding of our own lives that enriches us all.

References

*Chemero, A. (2009). *Radical Embodied Cognitive Science*. The MIT Press.

Cummins, F. (2002). On synchronous speech. *Acoustic Research Letters Online*, *3*(1), 7–11.

Cummins, F. (2003). Practice and performance in speech produced synchronously. *Journal of Phonetics, 31*(2), 139–148.

Cummins, F. (2009). Rhythm as entrainment: The case of synchronous speech. *Journal of Phonetics, 37*(1), 16–28.

Cummins, F. (2012). Periodic and aperiodic synchronization in skilled action. *Frontiers in Human Neuroscience, 5*(170).

Cummins, F., Li, C., & Wang, B. (2013). Coupling among speakers during synchronous speaking in English and Mandarin. *Journal of Phonetics*. Submitted.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *European Journal of Neuroscience*, *15*(2), 399–402.

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.

Frackowiak, R. S. J., & Frith, C. D. (1996). Functional anatomy of inner speech and auditory verbal imagery. *Psychological Medicine, 26*, 29–38.

*Gibson, J. J. (1986). *The Ecological Approach to Visual Perception*. Psychology Press.

Kalinowski, J., & Saltuklaroglu, T. (2003). Choral speech: the amelioration of stuttering via imitation and the mirror neuronal system. *Neuroscience & Biobehavioral Reviews*, *27*(4), 339–347.

*Kelso, J. S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior*. The MIT Press.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1–36.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In W. J. Hardcastle, & A. Marchal (Eds.), *Speech Production and Speech Modelling* (pp. 403–439). Kluwer Academic, Dordrecht.

Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature, 244*(5417), 522.

Maturana, H., & Varela, F. (1987). *The Tree of Knowledge: The Biological Roots of Human Understanding*. New Science Library/Shambhala Publications.

Pikovsky, A., Rosenblum, M., & Kurths, J. (2001). *Synchronization: A Universal Concept in Nonlinear Sciences*. Number 12 in Cambridge Nonlinear Science Series. CUP.

Pinker, S. (1999). *How the Mind Works*. W. W. Norton.

Rizzolatti, G. and Craighero, L. (2004). The mirror-neuron system. *Annu. Rev. Neurosci., 27*, 169–192.

Simko, J., & Cummins, F. (2011). Sequencing and optimization within an embodied task dynamic model. *Cognitive Science, 35*(3), 527–562.

*Stewart, J. R., Gapenne, O., & Di Paolo, E. A. (2010). *Enaction: Toward a New Paradigm for Cognitive Science*. MIT Press.

Turino, T. (2008). *Music as Social Life: The Politics of Participation*. University of Chicago Press.

*Varela, F. J., Thompson, E. T., & Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. MIT press.

Vygotsky, L., Hanfmann, E., & Vakar, G. (2012). *Thought and Language*. MIT press.