The Remarkable Unremarkableness of Joint Speech

Fred Cummins[1]

[1] *University College Dublin, Dublin 4, Ireland*

fred.cummins@ucd.ie

Abstract

Joint speaking, in which many people say the same thing at the same time, is a common vocal practice found in situations of heightened collective significance. In the wild, prosodic stylization is common. In the laboratory, we show that this stylization is not a necessary consequence of the requirement to speak in unison. Speech obtained from groups of 2, 4, 6 and 8 speakers remains relatively unaltered. But if the speech is unremarkable, the act of speaking is clearly not, and there is some behavioral and neuroscientific evidence for emergent phenomena arising in joint speaking that are not present in the speech of single individuals. Consideration of the status of joint speech, and its remarkable absence from contemporary linguistics, suggests that structuralist approaches to language that inform most of modern linguistics oversee much of that which is important about vocal communication.

Keywords: joint speech, choral speech, synchronous speech, unison

1. Introduction

Joint speech is an umberella term that covers those forms of speaking behavior when many people say the same thing at the same time. Joint speech practices are found in every culture, and the study of joint speech production cannot overlook the fact that most situations in which joint speaking occurs appear as overt manifestations of collective intentionality, in which group purposes, group sentiments and group intentions are made vocal.

Collective prayer is the most common form of joint speaking, found in all major religions. Prayers are often short, and repeated many times over, often with the aid of prayer beads. They may have a call and response form, and prayers that are frequently repeated often exhibit highly stylized prosodic forms in which the segmental details are distorted or even radically altered. Fig. 1 shows excerpts from a recording of the recitation of the Catholic Rosary. Four successive iterations are shown of the underlying phrase "Blessed is the fruit of thy womb, Jesus" which is, for this speaker, reliably and invariantly produced as (approximately) /bles.froʊ.θhaɪm.ʤɪz.ʒ̊/.[1] Despite the substantial alteration to the segmental composition of the phrase, it can be seen that the four iterations are produced with considerable stability and consistency.

Prayer and protest make odd bedfellows, but in the chants of protesters we find many similar features: the expression of a collective intentionality through the production in unison of short and frequently repeated phrases. Along with repetition, many protest chants are structured as call and response, not unlike the structure, for example of the Catholic Hail Mary

[1] Thanks to Neasa Ní Chiarain for the recording.

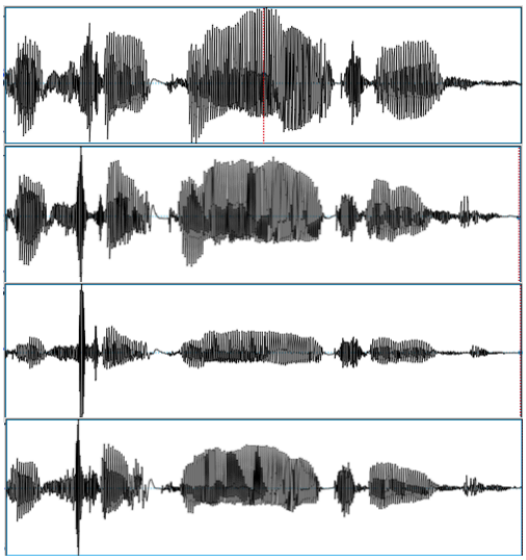

Figure 1: Four successive iterations of the phrase "Blessed is the fruit of thy womb, Jesus".

prayer. In protest too we again find the emergence of stereotypical prosodic forms. Perhaps the best known is the sequence of accents which we might characterize as in Fig. 2. This is the accentual structure of the famous chant "El pueblo, unido, jamás será vencido" ("The people, united, will never be defeated!"), which became globally known after the CIA-led coup that overthrew the government of Salvador Allende in Chile in 1973. What had been a chant associated with his election campaign became a global template for protest. Today the same basic pattern is found in many cultures and languages. A cursory survey of amateur videos from protests throws up examples in English, Greek, Portuguese and Arabic, at least. It is also the basic template for the widespread chant that has come to represent the uprisings across North Africa and the Middle East in the so-called "Arab Spring", where the call of "Ash-sha'b yurīd isqāṭ an-niẓām", or "The people demand the fall of the regime" has been used and adapted in Lybia, Tunisia, Egypt, Sudan, Syria and beyond.

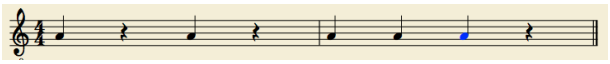

Figure 2: Common accent pattern found in protest chants in many languages.

The basic elements of repetition, prosodic stylization, (sometimes) call and response, and the vocal demonstration of

a collective identity are also found in the chants of sports fans. Beyond these domains, joint speech is used in educational settings for diverse purposes such as rote learning, pronunciation training, and performance. In the latter case, the term "choral speaking" is conventionally used.

Despite its ubiquity, and despite the embedding of joint speech practices into situations of heightened collective significance, there has been little or no scientific treatment of joint speech practices to date. A laboratory variant of joint speaking, dubbed Synchronous Speech, has been studied in some detail (Cummins, 2003; Cummins, 2009). In this paradigm, pairs of subjects are presented with unseen texts to be read in synchrony with one another on a go-signal from the experimenter. This task is surprisingly easy for subjects, and the temporal alignment attained has been estimated to be characterized by a mean asynchrony of approximately 40 ms (Cummins, 2003). Anecdotal observations of synchrony in collective prayer suggest that synchrony in a synchronous speech task is much tighter (less asynchrony) than that commonly found in the wild. Unlike the ritual repetition of prayer or protest chant, utterances produced in a synchronous speech experiment are not prosodically marked, and indeed speaking in synchrony has become an experimental constraint that can be used to reduce inter-subject variability in phonetic experiments (Cummins and Roy, 2001; Cummins, 2004; Krivokapić, 2007; Kim and Nam, 2008).

In a recent study, some prosodic alteration, in the form of an exaggerated syllable-timing, was observed when Mandarin speakers read in synchrony (Cummins et al., 2013). In follow up work, we have been unable to reliably replicate this effect. We now suspect that the apparent change was a perceptual effect that is properly attributable to the known slower speech rate of synchronous speech, rather than any substantial prosodic reorganization. And so we now confront the observation that both prayer and protest chants reliably exhibit highly stylized prosodic forms, while synchronous speech in the laboratory apparently does not. We therefore conducted a small experiment to see whether collective speaking in groups larger than the dyad resulted in substantial prosodic reorganization. Representative results will be provided here.

2. Synchronous speech is unremarkable

2.1. Methods

Eight native English speakers took part in a single recording session. Five texts with very different accentual structures were employed, including a list of 8 trochees, a poem with some lines in a duple and some in a triple meter, and the second part of the Hail Mary prayer. Subjects stood in a single room with approximately one meter distance between neighbouring individuals. On the instruction of the experimenter, all texts were read in varying combinations of 1, 2, 4, 6 and 8 speakers at a time, with group membership chosen randomly. Each subject wore a head-mounted microphone to minimize cross-channel bleed.

In order to examine the coarse temporal structure of utterances, we look at the intervals between prominent onsets. For each utterance, the onset of prominent syllables was calculated using the algorithm introduced in Cummins & Port (1998), which provides a working estimate of the perceived onset, or P-center of the syllable (Morton et al., 1976; Scott, 1993). Synchrony among more than two speakers has never been quantitatively examined, and so a quantitative estimate of pairwise asynchrony was also computed for each speaking pair, by estimating the degree of temporal warping necessary to map one utterance onto the other, using the computational method introduced in Cummins (2009). For dyadic readings this generates a single score, while for readings with 8 speakers, this produces 28 dyadic scores. In each case, a high score indicates lack of synchrony, and a score of zero would imply perfect synchrony.

2.2. Results

In what follows, we have very different numbers of observations for the different conditions of n=1, 2, 4, 6, and 8. We therefore eschew analysis by ANOVA and look instead for obvious qualitative features that might index the number of speakers. Our principal target is any evidence of wholesale changes to macroscopic temporal structure as a direct result of the synchronization task. Beyond that, we can ask whether adding more speakers alters pairwise synchrony in any obvious fashion.

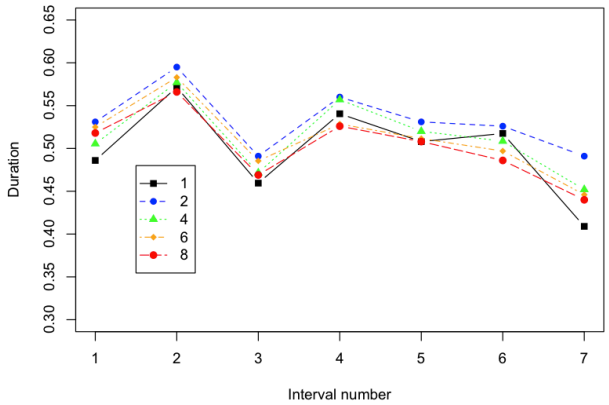

Figure 3: Inter-onset intervals for lists of 8 trochees.

We first consider timing measurements based on the list of 8 trochees ("Borrow, Dancer, Butter, Dagger, Boiler, Doggie, Body, Deeper"). Fig. 3 plots the median interval duration observed for the seven intervals defined by the eight word onsets, with separate plots for different numbers of speakers. No time normalization has been applied. Because only a single set of words was employed, there is little to be deduced from the slight pattern of alternation from one interval to the next, which is heavily influenced by the contingent segmental content of these specific words. For $n = 1$, the final interval is shorter than for all other values of n. But there is no visible effect whatsoever as one goes from 2 to 4, 6, and then 8 speakers. This is evidence that no substantial reorganization of macroscopic interval timing arises as a function of the number of speakers for this series of maximally regular words.

Turning now to a text with considerably more complex metrical structure, we examine the intervals between stressed syllable onsets in the short poem, the text of which reads (with apologies to cat lovers): *Kill a cat, kill a cat, Bash its brains in with a bat. Its nine lives expire, When tossed in a fire, So kill a cat today.* Fig. 4 shows the distribution of interval durations between stressed syllable onsets (underlined) for one and eight speakers. The shift from duple to triple meter between the seventh and eighth intervals is very obvious, and the metrical regularity is augmented by a considerable degree of lexical and segmental variability. However, when we examine the succession of median interval durations, there is no apparent qualitative difference observable as we move from one to 8 speakers. Examination of n=2, 4 and 6 confirms the absence of any sub-

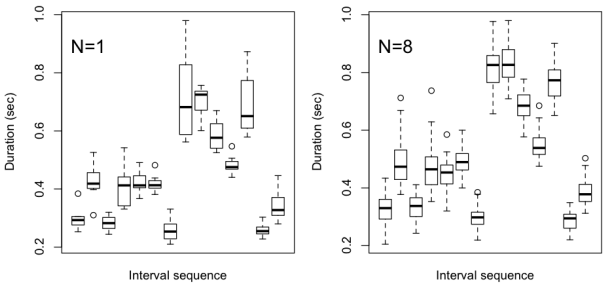

Figure 4: Inter-onset intervals for the *Kill a Cat* poem. Onsets correspond to underlined letters in the text of the poem.

stantial effect of group synchronization on macroscopic interval durations.

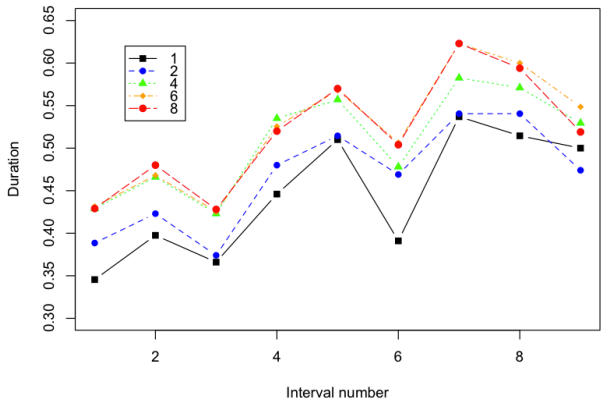

Figure 5: Inter-onset intervals for the second half of the Hail Mary prayer. Onsets correspond to underlined letters in the text.

Finally, we consider a text with less pronounced metrical structure, but whose inclusion is warranted as it is frequently recited collectively (Fig. 5). We use the second half of the Hail Mary prayer: *Holy Mary, Mother of God, Pray for us Sinners, Now, and at the Hour of our Death, aMen*. As with the trochees, we plot median interval durations as a function of the number of speakers. In this case, there is a clear rate effect, as recitations with 4, 6 and 8 speakers are all slower than those with one or two speakers. With the exception of the sixth interval, the pattern of successive durations appears relatively invariant. The sixth interval is that between *Sinners*, and *Now*, which includes a major syntactic break. Previous work has established that speaking in synchrony greatly reduces inter-subject variability in pause placement (Cummins and Roy, 2001), though no work has identified either lengthening or shortening as an overall effect of synchronization.

We estimated the asynchrony among all possible pairs of speakers. Asynchrony estimates are in units of area under a time warping curve that maps one utterance of a pair onto the other, as reported in Cummins (2009). Estimates have been log transformed, which produces more nearly normal distributions, and have then been converted to standard scores. Fig. 6 shows the distribution of asynchrony scores as a function of the number of speakers for the Hail Mary text. Similar results were obtained for other texts. It is clearly not the case that adding more speakers leads to greater dyadic synchrony, nor asynchrony. There is

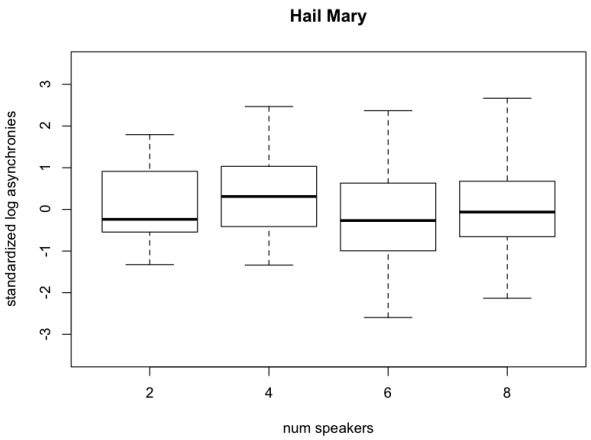

Figure 6: Distribution of pairwise standardized long asynchrony scores for different number of speakers.

thus no manifest effect of the number of speakers on either the macroscopic temporal structure of an utterance, nor on the synchrony obtaining among subjects, with the one exception found that a pause at a major syntactic break led to convergence upon a relatively long value for $n > 1$ speakers. In line with previous results, it seems then that synchronous speech is relatively unmarked and unremarkable, even as we increase the number of speakers from 2 to 8. The exaggerated prosodic stylization that we regularly find in collective prayer and protest is thus not due to the demands of synchronization among speakers.

3. But synchronous speaking is special

If synchronous speech is unremarkable, the same can not be said for synchronous speaking. We have established two sources of evidence that suggest that the act of speaking together needs to be understood as a collective act, in which the speakers become mutually entangled, or coupled, in a manner analogous to the way in which two runners in a three-legged race are physically coupled.

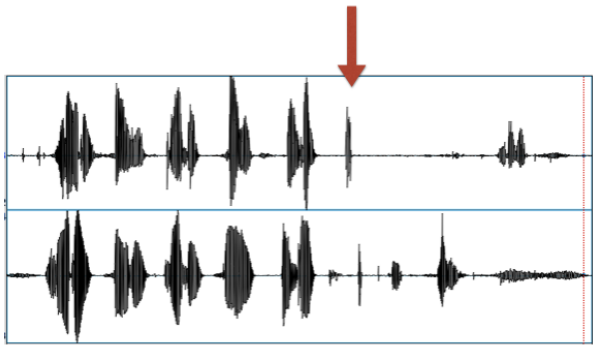

Figure 7: Waveforms illustrating synchrony interrupted by abrupt and simultaneous cessation of speech in mid-syllable. The arrow indicates the point of cessation.

The first source of evidence arises in the observation that there is a distinguished class of speech error, common when two people speak in synchrony, and unknown otherwise. This typi-

cally arises when one speaker either makes an error, or displays some degree of uncertainty. What is then observed is that the two speakers abruptly stop speaking, often simultaneously, at which point laughter typically ensues, but without further coordination across the individuals. Just like the runners in a three-legged race, the dyadic system that accomplishes the collective task is somewhat brittle, and an error by one person can bring the coordinative unity tumbling down. Fig. 7 illustrates waveforms from one trial in which subjects read lists of trochees. A small speech error on the part of one in the fifth word leads to the abrupt and almost simultaneous cessation of speech in both speakers. The point of cessation is indicated by an arrow, and it lies part way through the onset syllable of the sixth word. Other speech error behaviors also occur, e.g. one speaker may continue while the other stops. But abrupt cessation happens frequently and when it happens virtually simultaneously, it is unique to the synchronous speaking condition.

The second strand of evidence comes from a recent fMRI study conducted by Sophie Scott and Kyle Jasmin (mspt. in preparation) in which subjects spoke sentences in a variety of conditions: rest (silence), listening to a sentence, speaking alone, speaking together with the experimenter (different sentences), speaking in synchrony with the experimenter, and speaking in synchrony with a recording. Two comparisons of regional cortical bloodflow subsequent to speaking are of particular interest here. Firstly, when a comparison is made between speaking in synchrony with the experimenter and the conjunction of speaking alone and listening alone, there is a marked increase in activity in the primary auditory cortex and in the anterior and posterior auditory processing streams bilaterally (Scott and Johnsrude, 2003). This demonstrates that speaking in unison with others is not merely speaking+listening. In the second comparison, cortical activity when speaking in synchrony with an experimenter was markedly different from that observed when speaking in synchrony with a recording of the same experimenter—a contrast of which the subjects were unaware, not having been informed that recordings were to be employed at all. Subsequent debriefing confirmed that no subjects were aware of the distinction. A detailed account of these findings is in preparation, but the fact that real time live synchronization with another speaker produces cortical activity that is different from speaking+listening, and that is sensitive to the reciprocal interaction among live persons, suggests that synchronized speaking, as opposed to synchronized speech, is special indeed, in the sense that it exhibits properties that are not derivable or predictable from the mere conjunction of speaking activity by more than one person.

4. Discussion

Joint speaking appears to be a rather bizarre activity if we view speech as a specialization of language, and language as the means by which we exchanged encoded propositions. In joint speech, everybody is speaking, and nobody is listening. Utterances are repeated over and over, frequently with associated emphatic body movements, but it is unclear who, if anybody, is being addressed. Yet we are all familiar with joint speech practices, and they appear to be ubiquitous, old (Vedic chanting practices go back about 3,500 years), and to bear an important role in certain forms of highly charged collective activity. Joint speech itself, regarded as structured sound or movement, is relatively unremarkable, as we have seen. Yet the act of joint speaking seems to be imbued with a great deal of significance for practitioners. Utterances that are issued collectively seem to be performative, rather than referential in nature, as something is accomplished by the very act of speaking collectively (Austin, 1975; Meijers, 2007). The behavioral and scientific evidence suggests that the phonetic and phonological properties of joint speech are intact, but that something else is going on at the collective level that is unique to the realtime reciprocal linkages between joint speakers. An understanding of what this behavior is, why it occurs, and how it acquires such significance requires us to adopt a rather different perspective on speech and language.

The scientific study of language that has arisen since the structuralist approach of de Saussure has emphasized some aspects of language, notably the combinatorics of finite, discrete elements in symbolic structures, that must then be decoded by a listener. In this view, the roles of speaker and listener are utterly distinct, and much of the behavior we are familiar with that attends speaking is simply not addressed. This is a thoroughly conventional view of what "language" is, and of the kind of message-passing activity that it facilitates. More recent extrapolations of this approach have attempted to narrow, rather than enlarge, the set of "linguistic" phenomena proper, so that on one influential view, "language" is to be viewed as a modular faculty whose defining (perhaps only?) property is the support of recursion (Hauser et al., 2002). This extremely narrow focus will not serve to understand joint speech, nor will it serve to understand most human languaging behaviors.

The multifarious ways in which speaker/listeners become coupled during a conversational exchange, with the rich intertwining of facial and manual gestures, with backchannels that support, encourage and nudge the flow of speech, with the careful regulation of gaze, all these, because specific to face-to-face vocal communication, are excluded from the science of language so construed (Richardson et al., 2007; Wagner et al., 2014). Within the descendants of the structuralist tradition, even the sound itself is to be partitioned into those elements that are found to support the demarcation of discrete sound categories (phonemes) and everything else which is consigned *en masse* to the miscellaneous drawer of "prosody". Prosodists, then, spend most of their time vainly attempting to demonstrate how non-segmental aspects of speech sounds can be retro-fitted into a symbolic framework that has aquired the status of an immovable authoritative object.

If we pull back our field of vision and examine what a science of language is asked to provide an account of, we find that there has been a veritable industry born of the construction and defence of theories of how language has evolved, how it gives rise to the construction of a shared world, enables the development of the whole of human culture and technology, how it facilitates complex cognitive processes, and more (Deacon, 1997). All these great feats are unhesitatingly attributed to something called "language", and the scientific position on what *that* is finds its acknowledged authority in the fields of syntax and semantics. Remarkably, theories of syntax and morphology, along with the whole of formal semantics, can all be constructed, tested, and established without distinguishing in any meaningful way between written and (transcribed) spoken utterances at all. From some perspectives, it appears as if the object of academic linguistics might be better viewed as the code underlying *writing*.

If we recognize this, we might begin to reassert and recognize some of the power of the voice (Connor, 2000). Writing is much younger than language. The introduction of writing introduced wholesale changes to how we think, how we communicate, how we argue, reason, and situate ourselves in a shared

world (Olson, 1996; Ong, 1982). It is an extension of the vocal tradition that preceeded it by dozens of millenia, and the regularities of the written code are elaborations of, and transformations of, the regularities that are to be found in vocal utterances. To dismiss the productions of the voice as mere performance, and insist that they derive from an underlying formal system of rule based symbolic manipulation, is to deny the power of the voice, to ignore its position as the principal form of linguistic behavior, or languaging, which gave rise to what we recognize as modern humanity. It is to miss the very business of languaging.

For writing is not the only descendent of the much older phenomenon of vocal behavior. Many of the attributes of joint speech have likewise become codified, elaborated and transformed, but this has typically happened in the development of liturgy and ritual. The improvised and repeated gestures of spontaneous chant become encoded in practices of kneeling, solemn walking, head bowing, bead twirling, marching, and more. Hidden away in rituals consigned to the sphere of contingent cultural practice, such codifications have been overlooked by the sciences of language. Here, then, is the origin of the phonetic distortions, the prosodic stylizations, and the recurring patterns that arise in practices of prayer and protest. The characteristics of joint speech are invisible if we accept the received view of language, but if we can recognize the remarkable breadth of language behaviors, and the efficacy and power of the voice in structuring our collective practices, we can see so much more.

5. References

Austin, J. L. (1975). *How to do Things with Words*, volume 1955. Oxford University Press.

Connor, S. (2000). *Dumbstruck: A Cultural History of Ventriloquism*. Oxford University Press.

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

Cummins, F. (2004). Synchronization among speakers reduces macroscopic temporal variability. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, pages 304–309.

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

Cummins, F., Li, C., and Wang, B. (2013). Coupling among speakers during synchronous speaking in English and Mandarin. *Journal of Phonetics*, 41(6):432–441.

Cummins, F. and Port, R. F. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics*, 26(2):145–171.

Cummins, F. and Roy, D. (2001). Using synchronous speech to minimize variability in pause placement. In *Proceedings of the Institute of Acoustics*, volume 23(3), pages 201–206, Stratford-upon-Avon.

Deacon, T. W. (1997). *The Symbolic Species: The Co-Evolution of Language and the Brain*. WW Norton & Company.

Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579.

Kim, M. and Nam, H. (2008). Synchronous speech and speech rate. *Journal of the Acoustical Society of America*, 123(5):3736.

Krivokapić, J. (2007). Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics*, 35(2):162–179.

Meijers, A. (2007). Collective speech acts. In *Intentional Acts and Institutional Facts*, pages 93–110. Springer.

Morton, J., Marcus, S., and Frankish, C. (1976). Perceptual centers (P-centers). *Psychological Review*, 83:405–408.

Olson, D. R. (1996). Towards a psychology of literacy: On the relations between speech and writing. *Cognition*, 60(1):83–104.

Ong, W. (1982). *Orality and Literacy: The Technologizing of the Word*. TJPress, London.

Richardson, D., Dale, R., and Kirkham, N. (2007). The art of conversation is coordination. *Psychological Science*, 18(5):407.

Scott, S. K. (1993). *P-centers in Speech: An Acoustic Analysis*. PhD thesis, University College London.

Scott, S. K. and Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences*, 26(2):100–107.

Wagner, P., Malisz, Z., and Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, 57:209–232.