Zero Mean Lag Communication Over Networks: A Route to Co-Presence?

Fred Cummins[1], *Jonathan Byrne*[1]

[1]University College Dublin

fred.cummins@ucd.ie, jonathan.byrne@ucd.ie

Abstract

We contrast two ways of thinking about communication: communication as message passing, and communication as reciprocal coordination. From the invention of writing to the ubiquity of SMS, speech and language technology has uniformly employed the first model, and thereby done nothing to support, extend, or explore the second model. We suggest that the coordinative approach is better suited to understanding how face to face interactants establish co-presence. The technical challenges of establishing co-presence amounts to achieving synchronisation with a mean lag of 0 ms. We suggest that this goal might be approached through the exploitation of predictive models for behaviours that are inherently constrained, or known to both parties. Although we have not yet succeeded in achieving this goal, we chart a possible route of future exploration, with the distal goal of allowing people to engage in strongly synchronised behaviours such as chanting over networks.

Index Terms: co-presence, reciprocal interaction, liveness

1. Introduction

Communication can be thought of in more than one way. If we view communication as *coordination*, we are emphasising the manner in which your activity and mine become non-independent. If we view it as *message passing*, we focus instead on how my ideas and thoughts can be transferred to you through some medium. The first, coordinative, view emphasises the reciprocity of communication: you affect me and I affect you, simultaneously, and the distance between us is lessened. The second, conduit, view foregrounds the content of the exchange, but portrays the participants as relatively independent.

The coordinative perspective is most clearly demonstrated through the modality of touch. This is the primal mode through which infants first experience bonding with their mother, and it remains the most intimate form of communication, so much so that we ration and regulate who is allowed to touch whom, where, and when. When we touch, there is a two-way continuous connection between us that does not admit of dissection into independent components [1]. The two hands in a handshake cannot properly be understood as separate. In touch, the two subjects are necessarily *co-present* in a very important sense. The reciprocity of communicative touch is the reason why we do not have a recording medium for touch, analogous to our use of pictures or recorded sound. Nevertheless, co-presence is not a merely haptic phenomenon. In each other's presence, we clearly and mutually influence each other's gaze, the sounds we make, the manner in which we move and so on.

This contrasts starkly with communication conceived of as message passing, which has been often taken to be the essence of linguistic communication. Considered in this fashion, communication involves my private intentions and thoughts acquiring some kind of encoding—in sounds, text, images—and then being transferred to you for decoding. The communicating partners are here treated as separate entities and the act of communication is directed: either from me to you, or vice versa, but the act is essentially decomposable into sender and recipient.

This is not to insist that there are two categorically different kinds of communication, but to point out that we can describe, attend to, model, and perhaps facilitate coordinative or message-passing aspects of a given communicative situation. Remarkably, the technological support of speech and language has taken as its object the support of message-passing aspects of communication alone. From the development of writing—a technological breakthrough that set in motion a series of profound cognitive changes—to the most recent forms of messaging through smart phones, communication has been seen as one thing only, and the very possibility of augmenting or facilitating the coordinative aspects to communication has been overlooked [2]. It is to this that we turn our attention.

In what follows we seek to articulate the problem of establishing presence when interactants are only in touch (as it were) over networks. We do not yet have a working implementation of a communicative protocol that can support a sense of co-presence, but we hope it might be of some benefit to lay out the territory, sketch the logical form of a possible solution, and to show how initial exploration of this novel space of technological innovation might proceed.

2. Language and Joint Speaking

The development of many kinds of technological support for message passing has gone hand in hand with a specific view of the nature of language that has occupied the core of the discipline of linguistics for over a hundred years. We might characterise the first half of the 20th Century as the structural era, with the works of Saussure playing a central role, while the second half clearly belongs to the generative grammarians, whose most visible figure was Chomsky. Saussure formalised the study of *langue* over *parole*, thereby emphasising the abstract, systematic, formal aspects of linguistic communication, and self-consciously stepping over the messiness of verbal behaviour (a limitation of which he was painfully aware). Chomsky likewise valorised *competence* over *performance*, seeking to characterise an abstract underlying system that was at some remove from the messy business of speaking. Both programmes viewed speech as just one mode through which language finds expression, and language was understood as the exchange of propositional content encoded in rule-governed sequences.

This view of language has an odd historical flavour to it. After all, however we characterise language, it is surely as something which arose uniquely in our species and which differentiates us sharply from our nearest cousins, the great apes. To emphasise the symbolic, mode-agnostic, characteristics of some forms of language use is to adopt a perspective that doesn't care

about the differences between speech and writing. Yet writing has only been around for about 5,000 years, widespread literacy for no more than 500 years. Whatever happened to our species that gave rise to society, culture and human intellectual life is very much older than this, and the voice has been its primary vehicle all along.

There is a common form of speaking, found in every culture on Earth, and central to the affairs of all societies, that is ignored when we view language in this abstract, intellectualised way. This is joint speaking, which is found whenever multiple people say the same thing at the same time [3]. This is the form of speech found in all major religious traditions, frequently built into rituals which play an axiomatic role in establishing a common order. It is also found in situations of protest, when collectives give common voice to common concerns. And it is the mode of speech in which football fans enact a common identity on the terraces. As different as these domains are, joint speech displays some superficial characteristics that transcend the domains and that speak eloquently of the collective subject: The absence of any differentiation between speaker and listener stands in marked contrast to the abstract message passing view of language, as everybody is both speaker and listener, and everybody already knows the text. We also find a continuum of prosodic forms, with no clear distinction between speech and music, the English word "chant" serving double duty in both worlds. We find a central role of repetition, which makes sense only if we acknowledge the performative nature of joint speech: whatever is being achieved through this practice, it is achieved in real time only and through the urgent participation of all concerned.

In joint speaking, a highly charged form of co-presence is brought into being among interactants in a manner entirely unlike a sequenced exchange of messages [4]. Most people have experienced the sense of loss of personal autonomy (or its transference to a group) when taking part in chanting during protest or in support of a team. Choral singers are familiar with the remarkable sense of experiential blending or transcendence that arises when singing in unison. But perhaps the most eloquent illustration of the power of the co-presence that arises in this fashion is given, not by saying anything, but by being silent together. To be silent on one's own achieves nothing, but to be silent collectively, in joint commemoration of tragedy, illustrates viscerally the power of co-presence and the relative unimportance of the lexical content. The frenzy of an angry mob, or the ecstasy of the heavily embodied chant of Sufi *dhikr* likewise reveal the power of joint speech for turning collective activity into something highly charged, something shared, and something that is enacted by doing [5].

The focus on language as message-passing has led to many technologies that allow us to encode and transmit messages of many forms. But there are no technologies that allow us to speak together. This seems odd, and it is not hard to think of uses for such technologies. In what follows, we will first illustrate the problem, and then go on to discuss how the next steps in illuminating this largely unexplored space of potential innovation might proceed.

3. Liveness and Skype

Users of Skype or similar services are frequently aware that there is a slightly unreal feel to the conversation. Some of this, particularly in older implementations, is due to the mismatch between the spatial location of the camera and the position of the eyes of the interlocutor. There is also a small temporal lag, but neither of these two factors is typically an obstacle to carrying on a conversation. VoIP services were developed for the purposes of conversation, and the current standards seek to guarantee a lag of no more than 150 ms end to end [6]. Under these circumstances, we can take turns in a conversation, but we can't chant. If two people try to sing or speak in unison, this seemingly small delay is compounded, leading to an inevitable breakdown in the coordination necessary to synchronise.

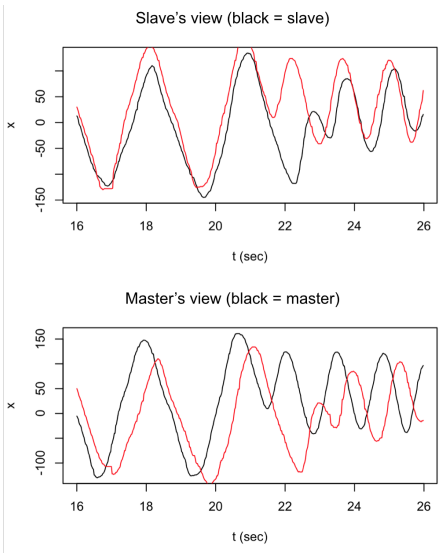

Figure 1: *Time (X) versus horizontal position (Y) for oscillatory finger movements produced in master-slave mode.*

The problem is illustrated in Fig. 1 which shows horizontal finger position against time for a networked application (described below) in which subjects at each end of a network make manual oscillatory movements that are displayed, with a 150 ms lag, on the screen of the other. In the example shown, the lower trace belongs to the "master" who was told to ignore the movements of the other, "slave", participant. The slave, on the other hand, was told to synchronise with the master. While the slave succeeds quite well, the master experiences two traces that are displaced in time by approximately 300 ms. Under these conditions, the equitable and reciprocal form of co-regulation that is possible in the flesh, is rendered impossible.

We frequently speak of "liveness" as if it were clear whether something is live or not. In a world filled with recordings, and edited creations, the distinction seems fairly straightforward. But with a little thought we can show that liveness is not an all-or-nothing affair. Consider a concert in which Jools Holland plays a duet with you. If you make a mistake, that will adversely affect the joint performance. You are very intimately involved with one another in live interaction. Now consider the same concert, but this time you are in the "live" studio audience. You have a sense of co-presence, and you could, at a push, influence things, e.g. by shouting or throwing something, but the manner in which you are involved in the whole scenario is rather different, and your role much smaller, than when playing the duet with Jools. Shift scene, and you are now at home, watching a "live" broadcast of the show. The advertised "liveness" does matter. You have a sense that things could go wrong, the unexpected could happen, and the spectacle is thus somewhat fragile.

But your capacity to influence things is now minimal. Change things just slightly, and you are watching the "live" performance with an hour's, or a year's, delay. There is a meaningful sense in which it is still live: the action is unbroken, the performance is probably less polished and somewhat less predictable than a studio recording, but the manner in which you are involved has now been watered down to homeopathic proportions. Liveness, then, admits of a good deal of variation, but it finds its strongest exemplar when several people are physically co-present to each other, deeply involved in each others' actions.

At first blush, it would seem that this is simply how things must be. If we communicate over networks, there must be a lag, because transmission times are non-zero, always. Synchronisation would seem to demand a zero-mean lag, and this sets an engineering goal that is unreachable in principle. We may be able to reduce lags well below 150 ms (and it is possible to perform a reasonable chant over landlines, as opposed to VoIP), but we cannot eliminate them. But we believe that this kind of thinking is, itself, beholden to the singular view of communication as message passing, and if we adopt instead a coordinative view, an unexplored opportunity for technical exploration opens up.

4. The Mirror Game

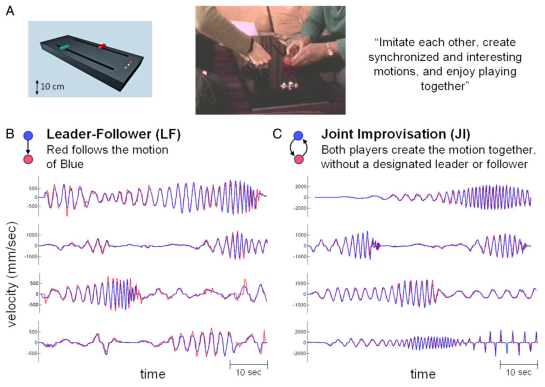

Figure 2: One-dimensional mirror game used in Noy et al. (2011). A: Movement of the sliders is sampled at 50 Hz. B: Sample velocity traces from a leader-follower round. C: Sample velocity traces from a joint improvisation round. From Noy et al. (2011)

There is a form of collective exercise known as the Mirror Game with origins in improvisation theatre and movement therapy [7]. In this, two or more participants improvise sequences of movements in one of two modes. In leader-follower (LF) mode, one person dictates the movement sequence while the others try to follow. In the second mode of joint improvisation (JI), nobody has the assigned role of leader, and synchronised activity must emerge spontaneously. Noy et al. created an experimental variant of this game in which movement is confined to the horizontal movement of a slider [8]. They found that patterns created in the two experimental conditions were equally complex, but that synchronisation was, on average, somewhat better in the JI condition. In particular, in JI rounds they would sometimes find periods of co-confident motion in which the two horizontal traces remained in lock step with no appreciable jitter (Fig: 3). A follow up study [9] they found that such co-confident motion displayed curvature that was qualitatively different from motion in which one player dominated or led. In [8], the authors introduced a reactive-predictive control model in which each participant used a simple model to predict future trajectories of the other.

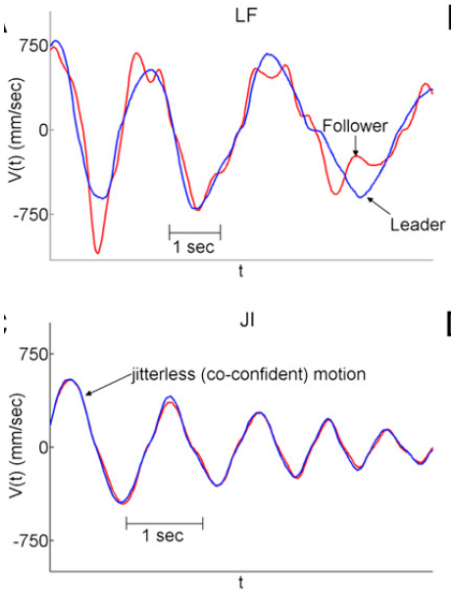

Figure 3: Example of co-confident motion (bottom) contrasting with more jittery trajectories found in LF-mode. From Noy et al. (2011)

No networks are involved here, of course. Players are synchronising in real face-to-face live interaction. However, the signature of co-confident motion illustrated in Fig. 3 may provide an important target that a comparable setup over networks might aim for. The mirror game thus provides us with a potential empirical index of reciprocal coupling among actors.

5. Towards zero mean lag over networks

In order to see how one might approach zero mean lag over networks it is useful to recall again the fundamental difference between communication as message passing and communication as coordination. In the first case, there is uncertainty about the message that will be transmitted. One must therefore wait until a signal has been received before one can know what it is. In the latter case, we are frequently dealing with a situation in which all parties know the sequence of movements, or words, that are to be performed. If that is the case, then for each subject, the future actions of the partner are largely predictable from the past. We can exploit this predictability in the behaviour of the interacting parties.

Let us consider two exemplary implementations of a notional zero-mean-lag system: long-distance chanting, and remote playing of the mirror game. In the case of chanting, predicability arises because the same text is repeated many times over. If the text is not known at the start, then it is available after a single iteration. Chanting thus is inherently predictable. In the case of the mirror game, trajectories along the rail are greatly constrained by physical contingencies. They will be

continuous, and at any given moment, knowledge of the position and velocity of the marker at the last several time steps, $t-1, t-2\ldots$ will allow confident prediction, within some margin of error related to the step size of discretisation.

Assuming that the signal can be discretised into equally spaced samples, and assuming a minimum transmission lag of one time step at each point, we simply ensure that what one person hears/sees/encounters at any time is the best possible prediction, based on recent values of the incoming signal. In what follows, $A(t)$ refers to the signal generated by person A at time t, and $\hat{B}(t)$ to the prediction of where B is most likely to be at time t. For simplicity, we assume prediction based on a single previous time step, but the approach should generalise to a sliding window of prediction.

At time t,
$\ldots A$ sees/hears $\hat{B}(t)$
$\ldots A$ says/does $A(t)$
$\ldots A$ receives $B(t-1)$
$\ldots A$ updates predictive model based on $B(t-1)$

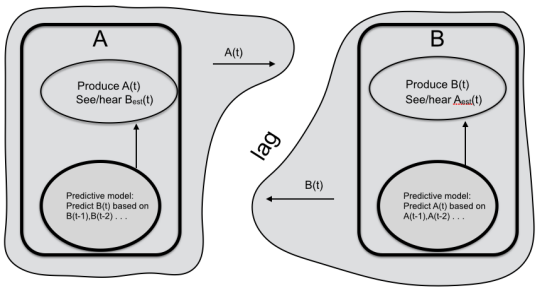

Figure 4: Basic architecture for exploiting prediction. Although there are lags in transmission from A to B, at any moment in time, A hears/sees/encounters a signal without lag, with fidelity that is proportional to the predictability of the signal.

This way of framing the problem and its solution is wholly generic. As long as the behaviour of one system is, to some extent, predictable by the other, a solution of this form can be implemented. The confidence with which the value of the signal at t can be predicted based on the last n values seen will be an important determinant of the strength of the mutual entrainment that can be achieved.

We are implementing a pilot system for exploring this idea (Fig. 5). We use two networked computers equipped with Leap Motion controllers [10]. In our initial implementation, each player sees a marker on screen corresponding to their own hand position in the vertically-oriented X-Y plane, along with another marker corresponding to a *prediction* of the other person's position, based on a sliding window of previous observations. Biological motion is constrained by the requirement that it be continuous, that motion be physically plausible, etc, so that the position of the co-player's hand can be predicted. For testing purposes, we include the option of introducing a specific fixed lag between packets exchanged over the network.

Fig. 6 illustrates the situation without any prediction. It shows asynchrony for trials run in master-slave mode, where the master ignores the slave, while the slave tries faithfully to synchronise with the master. As the transmission lag increases, so the asymmetry between the views of the two subjects becomes more extreme. The slave manages fairly constant performance across conditions, while the master becomes increasingly asynchronous with respect to the slave.

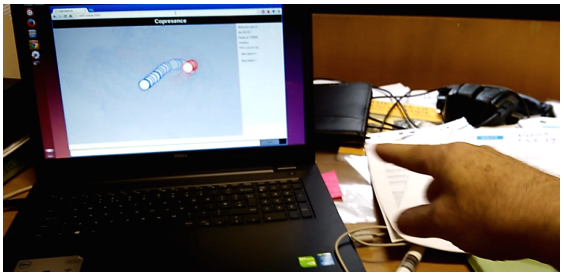

Figure 5: Snapshot of a pilot system under development

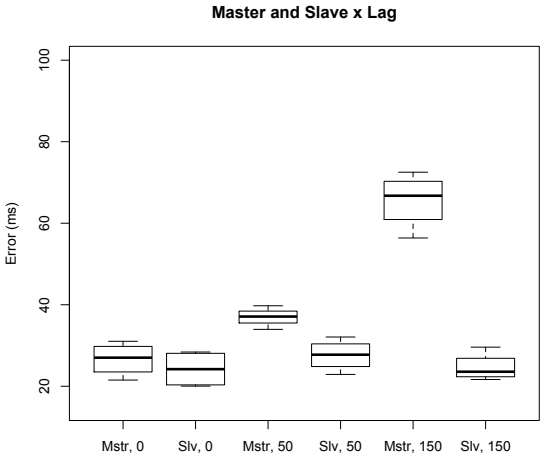

Figure 6: Measured asynchrony for both master and slave where the master ignores the slave, while slave synchronises with master, with no predictive model. Lags of 0, 50 ms and 150 ms were used.

We have tried several predictors so far, including a weighted linear, and a polynomial spline. At this point, however, a performance increase over the baseline has not yet been achieved. The system is intended to act as a crucible for refining questions about how a predictive system might overcome the challenge of zero mean lag synchronisation. In the remainder of this article, therefore, we consider the challenges ahead.

6. Extension to speech

If coordinated finger wiggling over networks, is challenging, one might opine that synchronising speech is even more so. We believe there are grounds for cautious optimism, however. One important aspect to chanting is to note that the text of the speech or song is known to both parties, and so does not, itself, need to be transmitted. Rather, because the sequence is known, the participants need only to keep track of where they each are in the sequential unfolding of that sequence over time.

A standard linear predictive encoding of speech produces a single vector for each windowed frame of the speech signal, and

good intelligibility can be obtained with a transmission rate of about 50 frames per second. This is a relatively sparse encoding, but still poses a significant challenge for prediction. For a known text, however, each participant can have a model sequence of LPC vectors, so that prediction becomes the easier task of estimating where, within the known sequence, the other participant is now. The known reference sequence could be generated using text to speech models, or, in an iterative process, the previous actual enunciation of the standard text might be employed. This remains as territory to be explored.

7. Conclusions

We have attempted to outline what a future technology of co-presence would be like. It would allow remote participants to establish genuine synchrony for known, or predictable, patterns of behaviour, including both hand movements and speech. It is our understanding that rich reciprocal coupling is a necessary prerequisite for engendering a sense of co-presence, and that this can be approached through the restricted domain of hand movement, and then extended to joint speech. The space of whole body synchronisation might be approached once progress has been made in these rather more restricted domains. Despite the initial difficulties, we believe that the development of prototypes and pilot systems may be a good way of teasing out the complexities of the field, and may be the starting point for a genuinely different form of communicative technology.

We foresee potential application in domains of human activity that have not yet benefitted from technological support, and we would suggest that one such domain is the participation in rituals, which frequently demands synchronisation of both speech and gesture. Participation in rituals is an important means by which cultural and religious identities of various kinds are maintained, and the enormous numbers of migrants, both voluntary and involuntary, suggests that there is a very large potential user base for any such application.

8. Acknowledgements

The prototype implementation is supported by a UCD seed funding grant to the first author.

9. References

[1] M. Ratcliffe, "Touch and situatedness," *International Journal of Philosophical Studies*, vol. 16, no. 3, pp. 299–322, 2008.

[2] W. J. Ong, *Orality and literacy*. Methuen & Co. Ltd., 1982.

[3] F. Cummins, "The remarkable unremarkableness of joint speech," in *Proceedings of the 10th International Seminar on Speech Production*, 2014, pp. 73–77.

[4] ——, "Voice, (inter-) subjectivity, and real time recurrent interaction," *Frontiers in Psychology*, vol. 5, 2014.

[5] ——, "Towards an enactive account of action: speaking and joint speaking as exemplary domains," *Adaptive Behavior*, vol. 21, no. 3, pp. 178–186, 2013.

[6] *Recommendation ITU-T G.114 One-Way Transmission Time*, Int'l Telecommunication Union Std., 1996.

[7] R. Schechner, *Environmental Theater*. New York: Applause Theatre and Cinema Books, 1994.

[8] L. Noy, E. Dekel, and U. Alon, "The mirror game as a paradigm for studying the dynamics of two people improvising motion together," *Proceedings of the National Academy of Sciences*, vol. 108, no. 52, pp. 20 947–20 952, 2011.

[9] Y. Hart, L. Noy, R. Feniger-Schaal, A. E. Mayo, and U. Alon, "Individuality and togetherness in joint improvised motion," *PloS one*, vol. 9, no. 2, p. e87213, 2014.

[10] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler, "Analysis of the accuracy and robustness of the leap motion controller," *Sensors*, vol. 13, no. 5, pp. 6380–6393, 2013.

[11] J. Laroche, A. M. Berardi, and E. Brangier, "Embodiment of intersubjective time: relational dynamics as attractors in the temporal coordination of interpersonal behaviors and experiences," *Frontiers in psychology*, vol. 5, 2014.