
Proceedings Paper:
Moore, R.K. orcid.org/0000-0003-0065-3311 (2019) Vocal interactivity in crowds, flocks and swarms : implications for voice user interfaces. In: Dassow, A., Marxer, R., Moore, R.K. and Stowell, D., (eds.) Proceedings of the 2nd International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR 2019). 2nd International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR 2019), 29-30 Aug 2019, London, UK. Ricard Marxer , pp. 94-99. ISBN 9782956202912

Vocal Interactivity in Crowds, Flocks and Swarms: Implications for Voice User Interfaces

Roger K. Moore

Speech & Hearing Research Group, Computer Science, University of Sheffield, UK

ABSTRACT

Recent years have seen an explosion in the availability of Voice User Interfaces. However, user surveys suggest that there are issues with respect to usability, and it has been hypothesised that contemporary voice-enabled systems are missing crucial behaviours relating to user engagement and vocal interactivity. However, it is well established that such *ostensive* behaviours are ubiquitous in the animal kingdom, and that vocalisation provides a means through which interaction may be coordinated and managed between individuals and within groups. Hence, this paper reports results from a study aimed at identifying generic mechanisms that might underpin coordinated collective vocal behaviour with a particular focus on closed-loop negative-feedback control as a powerful regulatory process. A computer-based real-time simulation of vocal interactivity is described which has provided a number of insights, including the enumeration of a number of key control variables that may be worthy of further investigation.

INTRODUCTION

Background

Recent years have seen an explosion in the availability of 'voice user interfaces' (VUIs), initially stimulated by the 2011 launch of *Siri* - Apple's smartphone-based voice assistant - followed in 2015 by Amazon's release of the first 'smart speaker' - *Alexa*. Since then, such smartphone and smart speaker based voice assistants have become almost ubiquitous. For example, *Siri* has had over 40 million monthly active users in the U.S. since July 2017, and smart speaker shipments reached 78 million units worldwide in 2018[1][2]. In the U.K., the number of people who own a smart speaker doubled from one-in-twenty to one-in-ten over a period of just six-months from autumn 2017 to spring 2018[3].

However, setting aside the impressive sales figures, a more critical aspect of such voice assistants is the extent to which they are actually used. For example, a survey conducted in 2015 (i.e. prior to the appearance of the first smart speaker) found that only 26% of the respondents used a voice assistant regularly and the majority of voice assistant users preferred typing to talking (Moore et al., 2016a). A more recent study by Kim (2019) investigating the usage of voice assistants on both smartphones and smart speakers found that over half of the smart speaker owners used their voice assistant several times a day. In contrast, only one-third of smartphone owners used their voice assistants on a daily basis, and half hardly used their voice assistants at all.

These studies also reveal that the majority of users employ quite stylised language, e.g. using simple voice commands to access music playlists, to perform searches using spoken queries, or to set alerts and reminders. Such shallow linguistic interaction is somewhat predictable given the nature of the problems users encounter with contemporary voice-enabled devices. For example, nearly half of the users surveyed reported difficulties with not being understood or simply not being able to do very much.

[1] https://medium.com/swlh/the-past-present-and-future-of-speech-recognition-technology-cf13c179aaf

[2] https://www.canalys.com/newsroom/smart-speaker-market-booms-in-2018-driven-by-google-alibaba-and-xiaomi

[3] https://yougov.co.uk/topics/politics/articles-reports/2018/04/19/smart-speaker-ownership-doubles-six-months

Key challenges

The usage statistics reported above confirm that contemporary VUIs are still a long way from being able to provide the "*conversational interface*" often promoted in the marketing literature for such systems. Indeed, the fact that users are effectively resorting to 'voice button-pressing' suggests that there is a fundamental difference between the richness of everyday human-human spoken language and the simplicity of the voice-based interaction that takes place between humans and machines. It has been argued elsewhere that a 'mismatch' between interlocutors is not only an important obstacle that needs to be explored in a human-machine context (Moore, 2015), but that it may even be an unsurmountable problem (Moore, 2016). In particular, if spoken language interaction is viewed as being based on the co-evolution of two key traits – *ostensive-inferential* communication and *recursive mind-reading* (Scott-Phillips, 2015), then contemporary voice-based systems are essentially only dealing with one aspect – inference. Some of the high-level issues relating to recursive mind-reading have been addressed in Moore and Nicolao (2017), but low-level concerns relating to ostensive vocal behaviour remain an open question. Hence, there seems to be something essential missing from contemporary voice-enabled system in the area of user engagement and interaction – not just what to say, but when to say it and, in a group context, to whom (Moore, 2015).

Potential solutions

Of course, interactive ostensive behaviours are ubiquitous in the animal kingdom, and vocalisation provides a means through which such activities may be coordinated and managed between individuals and within groups (Moore et al., 2016b). Vocalisations are often carefully timed in relation to each other (and other events taking place in an environment), and this may take the form of *synchronised* ritualistic behaviour (such as rhythmic chanting, chorusing or singing) or *antisynchronous* turn-taking (which can be seen as a form of dialogue) (Cummins, 2014; Fusaroli et al., 2014; Ravignani et al., 2014).

Of particular interest here are the *mechanisms* that support the emergence of synchronised vocal interactivity in crowds, flocks and swarms, and the implications of those mechanisms for future voice user interfaces. Hence, this paper presents results from an investigation into such mechanisms using a real-time simulation (i.e. a computational model) of interactive vocal dynamics and synchrony.

COLLECTIVE BEHAVIOUR

The coordinated behaviour of large numbers of independent living organisms has been the subject of scientific enquiry for many years. For example, studies have been conducted into the flocking of birds (Reynolds, 1987), the synchronised flashing of fireflies (Ermentrout, 1991), the dynamics of human crowd movement (Still, 2000), waves of coordinated clapping by audiences (Néda et al., 2000), and spatial sorting in shoals of fish (Couzin et al., 2002). Of particular interest are the transitions from one type of collective behaviour to another, especially in the context of attraction and repulsion between individuals (Katz et al., 2011), and predator-prey interactions (Handegard et al., 2012). Much of the research has involved computational simulation (perhaps the most famous being '*Boids*'[4]), as well as physical implementations in the field of swarm robotics, e.g. Bo et al. (2005).

One important aspect of synchronous behaviours is that they involve parallel coupled simultaneous action, as opposed to sequential action-reaction (Cummins, 2011). Such collective behaviours can thus be viewed as rhythmic entrainment, and thereby constitute a form of accommodation between individuals in a population (De Looze et al., 2014). It has also been posited that such behaviours underpin the links between different modalities, such as between vocalisation and physical movement (Cummins, 2009).

Vocal Synchrony

Whilst there have been a number of studies of vocal synchrony in animals, e.g. male zebra finches (Benichov et al., 2016) and monkeys (Takahashi et al., 2013), what is important here is the synergy with human vocal behaviours. Much of this work has involved 'joint speech', i.e. people speaking in unison (Cummins, 2014), and a more sophisticated view of 'turn-taking' in human dialogue (Heldner and Edlund, 2010). Of particular relevance is evidence that verbal synchrony in large groups of people produces affiliation (von Zimmermann and Richardson, 2016), and that some conversational partners tend to converge their vocal behaviours (Edlund et al., 2009), while others do not (Assaneo et al., 2019).

[4] https://www.red3d.com/cwr/boids/

Mechanisms

With regard to the mechanisms underpinning coordinated collective behaviour, by far the most popular approach is based on *coupled oscillators* (Kuramoto, 1975; Strogatz and Stewart, 1993; Strogatz, 2012), particularly through 'pulsatile coupling' (Mirollo and Strogatz, 1990). Not only has this been a very productive modelling paradigm with real-world implications (such as the simulation of clustered synchrony in electricity distribution networks (Pecora et al., 2014)), but new results are continuing to emerge (Matheny et al., 2019). The coupled-oscillator paradigm is also attractive because of its potential compatibility with known neural mechanisms (Ermentrout, 1991; Matell and Meck, 2000). However, it is only one way of formulating a complex non-linear attractor space, and it overlooks a number of potentially important conditioning variables – e.g. *energetics* (Moore, 2012).

As a consequence, the work reported here departs from the standard coupled-oscillator approach. In particular, attention is given to an alternative paradigm for creating a space of behavioural attractors – 'closed-loop negative-feedback control' – a powerful *regulatory* mechanism with roots in 'cybernetics' (Wiener, 1965) and commonly deployed for stabilising engineering systems (DiStefano III et al., 1990) as well as providing a powerful *non-behaviourist* paradigm for modelling the behaviour of living systems (Powers, 1973). The main differences between this approach and coupled oscillators is that the convergence criteria can be made more explicit, thereby offering the potential to gain a deeper understanding of the implications of particular parameters/settings on the emergent collective behaviours. It also offers the advantage that it can, in principle, be generalised to the synchronisation of more complex metrical structures, e.g. as discussed by Fitch (2013).

SIMULATION FRAMEWORK

Basic principles

The basic operation of classic closed-loop negative-feedback control is as follows: (i) a reference signal specifies the *desired* consequences of a system's actions, (ii) the *actual* consequences are sensed/interpreted by the system and compared with the reference target, (iii) the resulting *error* generates a control signal that drives the system in a direction such that the error is minimised. The process continues around the loop causing the system to not only converge to the desired behaviour but, more significantly, to maintain that behaviour in the face of arbitrary disturbances *without* having to sense such disturbances directly.

The tracking behaviour of such a negative-feedback control system is a function of the 'loop gain' of the controller. If the loop gain is too low, then stabilisation may take a long time – an 'overdamped' system. On the other hand, if the loop gain is too high, then the system may overshoot and even oscillate – an 'underdamped' system. The point here is that the loop gain effectively corresponds to the degree of *effort* (energy) applied to a regulatory task, i.e. from a psychological standpoint, it is analogous to *motivation*. An agent that 'cares' about controlling a particular variable would have a high loop gain, whereas a loop gain of zero implies the agent doesn't care at all (i.e. it gives up control). These are important individual differences that are not explicit in the coupled-oscillator approach.

Implementation

The simulations described herein have been implemented in Pure Data[5] – known as "*Pd*" – an open-source object-oriented dataflow programming language that is designed for real-time audio processing (Farnell, 2008). An environment has been constructed in which any number of vocalising (and listening) 'agents' may be connected to each other in arbitrary network topologies. Each agent comprises two feedback-control loops: one to regulate the interaction with other agents and another to regulate the agent's own behaviour. The first of these control loops aims to maintain synchrony between an agent's own vocalisations and those from agents that it can 'hear' (i.e. those to which it is connected). The second control loop attempts to maintain the agent's own preferred vocal rhythm. This arrangement means that each agent has two control parameters that influence the priority given to 'self' versus 'other'.

In addition, each agent has settings for the amplitude and duration of its vocalisations, their phase relation with the rhythmic 'beat', and the agent's preferred rhythm. In principle, these parameters could also be the subject of optimisation using feedback-control, but this was not implemented in the experiments reported here.

[5] `http://puredata.info/`

The sound output from each agent was produced using real-time synthesis of human, bird or insect vocalisations (as selected by the user). Other vocal characteristics for each agent (e.g. pitch frequency) were initialised randomly in order to provide a moderate level of 'individuality'.

Figure 1 illustrates a particular configuration of the simulation environment.

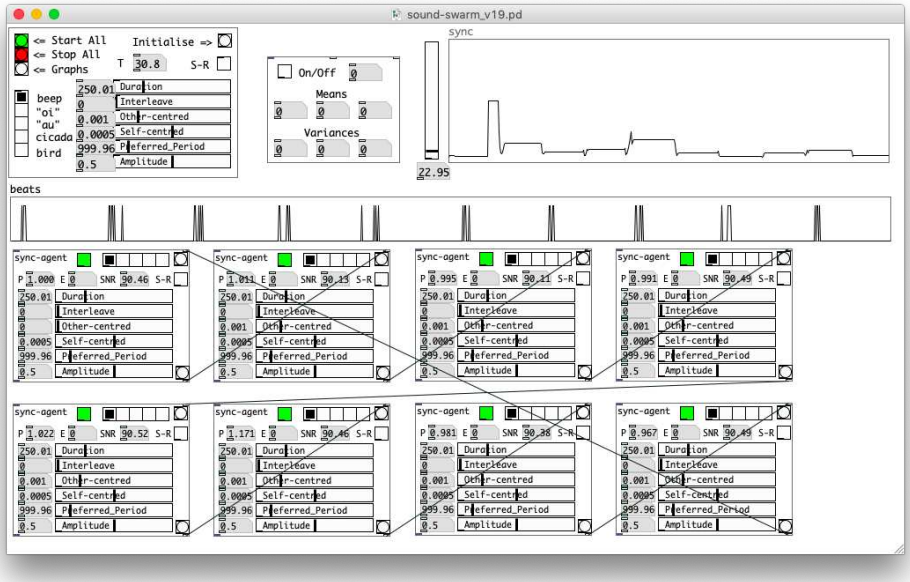

Figure 1. *Screenshot of the user interface for the* Pd*-based vocal simulation environment. The main control panel is shown in the top-left corner; buttons and sliders allow a user to specify global parameters for the population of agents, such as the type of vocalisation (human, insect, bird), duration, loudness etc. The lower half of the user interface facilitates the creation of an arbitrary number of agents, and the specification of which agent is listening to which other agent(s). In the example shown, eight agents have been configured in a loop topology (agent-2 is listening to agent-1, agent-3 to agent-2, agent-4 to agent-3, . . . , agent-1 to agent-8). As can be seen, sliders on each agent allow a user to set parameters individually if required. The graph shown at the top-right of the interface displays a timeline of the overall vocal synchrony between the agents, and the graph shown across the middle displays the individual rhythmic 'beats' from each agent (bunching indicates a degree of synchrony).*

Experiments

A range of experiments has been conducted based on varying numbers of interacting agents, different interconnection topologies, and alternative parameter settings. There is not space here to report all the findings. So what follows is a selected highlight.

One of the overarching research questions is concerned with the relationship between the topological connections between agents (i.e. the 'ostensive' relationships) and the emergent collective behaviour of the agents. In this context, one particular configuration is a *chain* with a 'master' (pacemaker) agent and a sequence of 'slave' agents. Figure 2 illustrates the outcome of simulating such a configuration with a chain of eight agents. As can be seen, on average, all of the agents in the feedback-control configuration maintained synchrony, but the agents further down the chain exhibited less stable rhythms. In contrast, agents in an action-reaction configuration maintained the rhythm, but the agents further down the chain were increasingly out of sync.

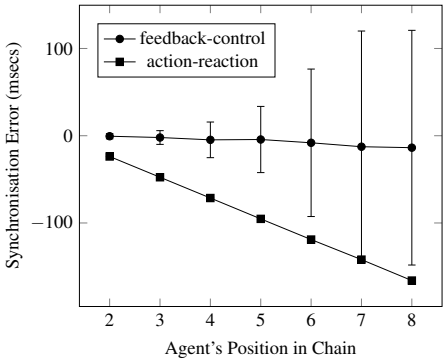

Figure 2. *Relationship between a 'slave' agent's position in a chain and its vocal synchronisation error with respect to the 'master' agent at the head of the chain.*

DISCUSSION & CONCLUSION

As a result of this research, it is possible to draw some conclusions about the control variables that are worthy of investigation with respect to vocal interactivity. These are summarised in Table 1.

Table 1. *Dimensions of vocal interactivity identified in this study. The left-hand column specifies the relevant control variables, and the right-hand column gives an indication of the expected range of values.*

VOCAL AGENTS	
Individuality (e.g. style of vocalisation)	average ⇔ extreme
Ostention (i.e. stance towards other agents)	connected ⇔ disconnected
Intentionality (i.e. goals wrt other agents)	same ⇔ different
Motivation/effort expended on pursuing others' goals	high ⇔ low
Motivation/effort expended on pursuing own goals	high ⇔ low
VOCALISATIONS	
Intensity (e.g. volume)	low ⇔ high
Clarity (e.g. intelligibility/SNR)	low ⇔ high
Period (i.e. timing)	short ⇔ long
Mark-to-space ratio (i.e. duration)	0% ⇔ 100%
Sentiment (i.e. valence)	-ve ⇔ +ve
Meaning (e.g. category)	named-entity-1 ⇔ named-entity-2
VOCAL INTERACTVITY	
Synchrony (i.e. engagement)	low ⇔ high
Simultaneity (i.e. overlap/interleaving)	0% ⇔ 100%
Dependency (i.e. between vocalisations)	dependent ⇔ independent

In conclusion, this paper has outlined some of the key issues facing contemporary voice-user interfaces, with a special focus on emergent properties of collective vocal behaviour, especially ostensive interaction and timing. The focus has been on closed-loop negative-feedback control as a regulatory mechanism, which implements a coincidence detection scheme that is compatible with known neural mechanisms (Matell and Meck, 2000). The simulation of real-time interacting vocal agents has already provided a number of insights into such behaviour, and more are expected as the full parameter space is investigated. In particular, it should be possible to show (i) how dialogue emerges as a compensatory response to the automatic regulation of intelligibility, not as a trivial action-reaction behaviour (Benichov et al., 2016), (ii) how cooperative *vs.* competitive interaction conditions vocalisations, and (iii) how communicative behaviour emerges from vocal interaction (Rosenthal et al., 2015).

REFERENCES

Assaneo, M. F., Ripollés, P., Orpella, J., Lin, W. M., de Diego-Balaguer, R., and Poeppel, D. (2019). Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning. *Nature Neuroscience*, 22:627–632.

Benichov, J. I., Benezra, S. E., Vallentin, D., Globerson, E., Long, M. A., and Tchernichovski, O. (2016). The forebrain song system mediates predictive call timing in female and male Zebra finches. *Current Biology*, 26(3):309–18.

Bo, L., Tian-Guang, C., Long, W., and Zhan-Feng, W. (2005). Swarm dynamics of a group of mobile autonomous agents. *Chinese Physics Letters*, 22(1).

Couzin, I., Krause, J., James, R., Ruxton, G., and Franks, N. (2002). Collective memory and spatial sorting in animal groups. *Journal of Theoretical Biology*, 218:1–11.

Cummins, F. (2009). Rhythm as an affordance for the entrainment of movement. *Phonetica*, 66(1-2):15–28.

Cummins, F. (2011). Periodic and aperiodic synchronization in skilled action. *Frontiers in Human Neuroscience*, 5(170):1–9.

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

De Looze, C., Scherer, S., Vaughan, B., and Campbell, N. (2014). Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction. *Speech Communication*, 58:11–34.

DiStefano III, J. J., Stubberud, A. R., and Williams, I. J. (1990). *Feedback and Control Systems*. McGraw-Hill, New York, 2nd edition.

Edlund, J., Heldner, M., and Hirschberg, J. (2009). Pause and gap length in face-to-face interaction. In *INTERSPEECH*, Brighton, UK.

Ermentrout, B. (1991). An adaptive model for synchrony in the firefly Pteroptyx malaccae. *Journal of Mathematical Biology*, 29(6):571–585.

Farnell, A. (2008). *Designing Sound*. Applied Scientific Press Limited, London.

Fitch, W. T. (2013). Rhythmic cognition in humans and animals: distinguishing meter and pulse perception. *Frontiers in systems neuroscience*, 7:68.

Fusaroli, R., Rączaszek-Leonardi, J., and Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas in Psychology*, 32:147–157.

Handegard, N. O., Boswell, K. M., Ioannou, C. C., Leblanc, S. P., Tjøstheim, D. B., Couzin, I. D., Walczak, A., Parisi, G., Procaccini, A., Viale, M., and Al., E. (2012). The dynamics of coordinated group hunting and collective information transfer among schooling prey. *Current Biology*, 22(13):1213–7.

Heldner, M. and Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

Katz, Y., Tunstrøm, K., Ioannou, C. C., Huepe, C., and Couzin, I. D. (2011). Inferring the structure and dynamics of interactions in schooling fish. *Proceedings of the National Academy of Sciences of the United States of America*, 108(46):18720–5.

Kim, Y. (2019). *Usage of Speech Technology Systems*. Final-year project dissertation, Dept. Computer Science, University of Sheffield.

Kuramoto, Y. (1975). Self-entrainment of a population of coupled non-linear oscillators. In Araki, H., editor, *International Symposium on Mathematical Problems in Theoretical Physics*, pages 420–422.

Matell, M. S. and Meck, W. H. (2000). Neuropsychological mechanisms of interval timing behavior. *Bioessays*, 22(1):94–103.

Matheny, M. H., Emenheiser, J., Fon, W., Chapman, A., Salova, A., Rohden, M., Li, J., de Badyn, M. H., Pósfai, M., Duenas-Osorio, L., Mesbahi, M., Crutchfield, J. P., Cross, M. C., D'Souza, R. M., and Roukes, M. L. (2019). Exotic states in a simple network of nanoelectromechanical oscillators. *Science*, 363(1057).

Mirollo, R. E. and Strogatz, S. H. (1990). Synchronization of pulse-coupled biological oscillators. *SIAM Journal on Applied Mathematics*, 50(6):1645–1662.

Moore, R. K. (2012). Finding rhythm in speech: a response to Cummins. *Empirical Musicology Review*, 7(1-2):36–44.

Moore, R. K. (2015). From talking and listening robots to intelligent communicative machines. In Markowitz, J., editor, *Robots That Talk and Listen*, chapter 12, pages 317–335. De Gruyter, Boston, MA.

Moore, R. K. (2016). Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. In Jokinen, K. and Wilcock, G., editors, *Dialogues with Social Robots – Enablements, Analyses, and Evaluation*, pages 281–291. Springer Lecture Notes in Electrical Engineering (LNEE).

Moore, R. K., Li, H., and Liao, S.-H. (2016a). Progress and prospects for spoken language technology: what ordinary people think. In *INTERSPEECH*, pages 3007–3011, San Francisco, CA. ISCA.

Moore, R. K., Marxer, R., and Thill, S. (2016b). Vocal interactivity in-and-between humans, animals and robots. *Frontiers in Robotics and AI*, 3(61).

Moore, R. K. and Nicolao, M. (2017). Towards a Needs-Based Architecture for 'Intelligent' Communicative Agents: Speaking with Intention. *Frontiers in Robotics and AI*, 4(66).

Néda, Z., Ravasz, E., Brechet, Y., Vicsek, T., and Barabási, A.-L. (2000). Self-organizing processes: The sound of many hands clapping. *Nature*, 403:849–850.

Pecora, L. M., Sorrentino, F., Hagerstrom, A. M., Murphy, T. E., and Roy, R. (2014). Cluster synchronization and isolated desynchronization in complex networks with symmetries. *Nature communications*, 5(4079).

Powers, W. T. (1973). *Behavior: The Control of Perception*. Hawthorne, NY: Aldine.

Ravignani, A., Bowling, D. L., and Fitch, W. T. (2014). Chorusing, synchrony, and the evolutionary functions of rhythm. *Frontiers in psychology*, 5:1118.

Reynolds, C. (1987). Flocks, herds and schools: A distributed behavioral model. In *SIGGRAPH '87: Proc. of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, pages 25–34, Anaheim, USA.

Rosenthal, S. B., Twomey, C. R., Hartnett, A. T., Wu, H. S., and Couzin, I. D. (2015). Revealing the hidden networks of interaction in mobile animal groups allows prediction of complex behavioral contagion. *Proceedings of the National Academy of Sciences of the United States of America*, 112(15):4690–4695.

Scott-Phillips, T. (2015). *Speaking Our Minds: Why human communication is different, and how language evolved to make it special*. Palgrave MacMillan.

Still, G. K. (2000). *Crowd Dynamics*. Phd, University of Warwick.

Strogatz, S. H. (2012). *Sync: How Order Emerges from Chaos In the Universe, Nature, and Daily Life*. Hatchette Book Group.

Strogatz, S. H. and Stewart, I. (1993). Coupled oscillators and biological synchronization. *Scientific American*, 269(6):68–75.

Takahashi, D. Y., Narayanan, D. Z., and Ghazanfar, A. A. (2013). Coupled oscillator dynamics of vocal turn-taking in monkeys. *Current biology : CB*, 23(21):2162–8.

von Zimmermann, J. and Richardson, D. C. (2016). Verbal synchrony and action dynamics in large groups. *Frontiers in Psychology*, 7:2034.

Wiener, N. (1965). *Cybernetics: or Control and Communication in the Animal and the Machine*. The MIT Press, Cambridge, Mass., 2nd edition.