AVANT, Vol. VIII, Special Issue
ISBN: 978-83-944622-04-6 avant.edu.pl/en
DOI: 10.26913/80s02017.0111.0019

ConversationPiece II: Displaced and Rehacked

Thomas Wennekers[1,3*], **Mathew Emmett**[1,2**], **& Susan L. Denham**[1,4***]

[1] Cognition Institute, Plymouth University, UK
[2] School of Art, Design and Architecture, Plymouth University, UK
[3] School of Computing, Electronics and Mathematics, Plymouth University, UK
[4] School of Psychology, Plymouth University, UK
[*] *twennekers@plymouth.ac.uk*
[**] *mathew.emmett@plymouth.ac.uk*
[***] *sdenham@plymouth.ac.uk*

Received 11 May 2017; accepted 26 September 2017; published 21 November 2017.

Abstract

Conversations are amazing! Although we usually find the experience enjoyable and even relaxing, when one considers the difficulties of simultaneously generating signals that convey an intended message while at the same time trying to understand the messages of another, then the pleasures of conversation may seem rather surprising. We manage to communicate with each other without knowing quite what will happen next. We quickly manufacture precisely timed sounds and gestures on the fly, which we exchange with each other without clashing—even managing to slip in some imitations as we go along! Yet usually meaning is all we really notice. In the ConversationPiece project, we aim to transform conversations into musical sounds using neuro-inspired technology to expose the amazing world of sounds people create when talking with others. Sounds from a microphone are separated into different frequency bands by a computer-simulated "ear" (more precisely "basilar membrane") and analyzed for tone onsets using a lateral-inhibition network, similar to some cortical neural networks. The detected events are used to generate musical notes played on a synthesizer either instantaneously or delayed. The first option allows for exchanging timed sound events between two speakers with a speech-like structure, but without conveying (much) meaning. Delayed feedback further allows self-exploration of one's own speech. We discuss the current setup (ConversationPiece version II), insights from first experiments, and options for future applications.

Keywords: conversation; dialogue; performance; sonification; sound analysis.

Introduction

Sounds offer a very effective means for communicating as they can be rapidly produced and broadcast into the surrounding medium (air, water) for asynchronous information exchange with others. Consequently, most animals have evolved auditory sensorimotor systems. Communication sounds in most land mammals are made by creating broadband (usually harmonic) sounds using vocal cord vibrations, which are spectrally shaped by changing internal (vocal tract, mouth and nasal) cavities to create resonances (Lieberman & Blumstein, 1988), often supplemented by noises produced by rapid tongue or lip movements. The prosody (pitch and amplitude) of the sounds depends on the vocal cord vibrations, while atomic communication sounds (e.g., speech phonemes) typically depend on the dynamic vocal tract resonances in combination with added noises, and in tonal languages on the dynamics of pitch, too. The wide range of different sounds that can be made in this way and concatenated into sequential strings underpins complex information exchange between individuals.

In general, it is not sufficient simply to broadcast messages to others; it is equally important to know whether the intended information is received and understood. Therefore, communication signals must be exchanged and the timing of the signals regulated in order to optimize the flow of information in both directions. This process of social communication is arguably the most important cognitive function of all, providing a basis for social bonding, information exchange and learning from others; possibly reaching a pinnacle in human conversational interactions. Through the exchange of tightly coordinated multi-sensory signals, people in conversation create a shared mental world. The apparent effortlessness of this process and the enjoyment normally derived from conversations raises the question "Why is conversation so easy?" (Garrod & Pickering, 2004). In an effort to answer this question, Garrod and Pickering suggested that conversation should be seen as a joint (bonding) activity in which interlocutors interactively align their thoughts, actions and perceptions at multiple levels, including basic acoustical properties (speaking rate, phonetic characteristics), phonological, lexical, syntactic, and semantic representations and situation models (Menenti, Pickering, & Garrod, 2012); key to successful communication ultimately is the alignment of situation models and the convergence of conceptual spaces. Alignment is achieved through interaction and the percolation of alignment between levels. The resulting tight coupling of sensorimotor systems is evident in phenomena such as chorusing or completing each other's utterances, and even in the entrainment of brain rhythms (Hasson, Ghazanfar, Galantucci, Garrod, & Keysers, 2012). Coupling and alignment are in a sense artifacts of the interaction but are important for its success; consider, for example, the improved comprehension achieved by imitating the (unfamiliar) accent of another (Adank, Hagoort, & Bekkering, 2010). Alignment has also been linked to predictability and

processes whereby interlocutors infer the intended messages of the other by minimizing their mutual prediction errors (Friston & Frith, 2015; Okada, Matchin, & Hickok, 2017). Eliciting predictable responses to one's own communicative actions provides confirmation of the success of communication, and indicators of new (unpredicted) information when unexpected responses occur.

The ease and rapidity with which people are able to dynamically adapt their communication sounds and gestures in order to reflect the demands and goals of the situation make the subtleties of the interaction very difficult to appreciate. In addition, the grouping processes of the auditory system hide the spectrotemporal intricacies of the sounds from perceptual awareness, so it is virtually impossible to hear out the sonic patterns that are created by the dynamics of formant, pitch and amplitude trajectories. In ConversationPiece II, our aim is to create a real-time system that exposes these patterns, transforming conversations into a kind of musical improvisation that enables listeners to appreciate some of the complexities and nuances of the sound world we all create in our everyday lives.

ConversationPiece II incorporates the concept of performance at the intersection between talkers, basically adding a playful element to conversation: speakers become partly detached from their conversation and act as performers of their own speech, which is transformed and sonified as musical sounds. By incorporating performance into our research, we can conceptualize and sonify the linkages between emerging identities, social behaviors and inter-relational human practice. Performance allows us to study the dynamics of conversations; in particular, how people connect using body language and tone of voice, carefully timed to attune themselves to each other. A conversation is a changing sequence of social interactions, following conventions which may be followed or violated, e.g., words imitated or repeated for dramatic effect, timing changed for emphasis. The transient nature of conversations cannot be separated from the performativity of the interaction. ConversationPiece II therefore depends on the level of performative collaboration that exists between the talkers.

Method

In the main interactive setup, incoming sounds produced by two talkers are processed separately. The sound waveform is first filtered by a bank of bandpass filters with center frequencies arranged on a log scale similar to that of the cochlea, here further restricted to the notes of a pentatonic scale. The outputs of the filters are multiplied by weights that counteract the high-frequency fall-off in power typical of human speech. The problem then is to sparsely sample salient points from the output of the filters that capture key structural aspects of the spectrotemporal patterns in the utterances as a sequence of discrete sound events. Sparse sampling both in time

and frequency space is crucial, as dense clusters of individual sound events are perceptually incomprehensible. Interestingly, this very same information is easily processed when the normal auditory grouping processes are at work. To achieve sparse sampling in continuous time, notes are generated at the pitch of the center frequency of a filter when the integrated filter output passes a threshold. Once a sample is generated, to prevent immediate resampling and dense note clusters, the running integration of the filter output and adjacent filter outputs are reset. The time constant of integration and the reset level determine a soft refractory period for each filter that causes temporal sparseness; the suppression of lateral filters further causes sparseness in frequency space. Neural networks in sensory areas of the mammalian neocortex are known to perform similar selection and lateral suppression operations.

The system has been implemented in custom C-code (called s2m, short for "speech-to-music") to allow for real-time capabilities. We run it under Linux (kernel version 4.11.12-100.fc24.x86_64 with no additional real-time modules) on a laptop with Intel i7-2820QM CPU at 2.3 GHz; on one core it uses up to 20% CPU time. Two microphones, one for each speaker, are connected to the left and right input of the inbuilt audio card (Intel IDT 92HD90BXX, but any card with a latency smaller than about 10 ms should work) and routed into the s2m-program using the Jack audio connection kit. The s2m software implements the processing scheme described above and outputs MIDI-events using the ALSA sound architecture for Linux for each of the speakers. The MIDI-events are routed using Jack into two independent software synthesizers for sonification. A common choice is Qsynth/timidity, a sound font synthesizer that allows the sound events from each speaker's voice to be mapped to a different instrument from the General MIDI instrument set; any MIDI-controllable software or hardware synthesizer could be used. Generated sounds are finally output through a soundcard and played back to the interlocutors by loudspeakers.

The interactive setup sonifies speaker sounds with an imperceptibly short delay (< 10 ms). A second system, s2m-delayed, has been built that delays detected notes in a speech signal before replaying them. The note detection and sonification follows the same principles as in s2m; however, notes are buffered until the speaker pauses for an adjustable amount of time—typically around 500 ms corresponding with common pauses in speech patterns. This allows speakers to "communicate with themselves"—any utterances are returned as musical instrument sounds reflecting the elements of prosody the system picks up (frequencies, pauses, sound amplitudes). The delayed system allows single speakers to observe and explore their own vocalizations in a transformed sound space.

Performing ConversationPiece II foregrounds conversational interactions as field territory and creates a shared mental world that extends the notion of the individual into a mutual prosthesis or collective entity of networked sensations. This amplifies

the heterotopic quality of communication. Assigning a performative value to the mutually reinforcing relationships created in a conversation makes explicit the form, texture, weight and nuances of the interaction. ConversationPiece II therefore immerses the talkers and listeners within a transient state of reciprocity.

Results

A prototype of ConversationPiece (version I) was exhibited at Off the Lip 2016 with moderate success; children in particular were intrigued by the transformation of their voices. In this version, videos of the interlocutors were simultaneously displayed, and as a result, there were two types of interaction with the system; some people watched and listened to others performing and tried to understand what was happening from the produced sound and video, while others actually took part in conversations, with generated sounds played to them. However, listening to the sounds while talking is very distracting, so those who took part in conversations chose either to play with the system to explore how they could generate sounds using their voices, or to focus on the conversation and largely ignore the resulting soundscape.

What we took away from this prototype exploration was that people were very interested in the concept of the system and the performative aspects it offered, but needed time to explore and play with its possibilities before performing with it. One possibility we intend to explore in the latest version is to optionally mix the voice back in to the instrumental output, thereby making the mapping more explicit. Finally, the simultaneous sonification of gestures would allow the tight embodied coupling also to be demonstrated. For this purpose, a lightweight wearable system has now been built which is able to pick up gestures from body motion using gyrometers and battery-powered micro-processor devices. The detected motion can be wirelessly transmitted to a computer and sonified together with the speech-generated sounds. This enhancement encouraged by ConversationPiece I will allow for a more accentuated form of sonification in the future that takes into account speakers' bodily actions to support their messages.

ConversationPiece II was demonstrated at Off the Lip 2017. The voices of two interlocutors were synthesized in real time using different (usually contrasting) instruments for each speaker. The participants in the conversation could hear both, the words and the sounds produced from their speech, while the audience could only hear the resulting soundscape. The questions and discussions which followed highlighted a number of trade-offs which we had made in developing the system: 1) It is possible to reproduce the speech sounds with quite high fidelity even when using discrete samples; for this, it is simply necessary to sample at about 25 ms (40 Hz) to make the speech fairly comprehensible. 2) If one wants to expose the soundscape or "music" of the speech in terms of rhythms and harmonic sound patterns, however, then a slower sampling closer to a few Hz is required, the trade-off being that the speech is no longer comprehensible even though many prosodic features still persist.

3) The use of instruments, while more pleasing to the ear and also necessary if the sounds of the two speakers are to be separable, introduces the overtones and some aspects of dynamics of the chosen instruments into the mix. This can make conversation soundscapes musically appealing, as for example, compared to a sonification in terms of mixes of grains of pure tones only, but it also adds semantic references to knowledge about the instruments used, which can distract from the communicative interaction: The analysis process of s2m intentionally strips out the immediate semantics of a conversation ("what it is about"), thereby exhibiting the spectrotemporal communication structure of the message. The sonification, however, can afterwards add artificial features un-intentionally that obscure the intended structure and make it to some degree unrecognizable, something akin to a "grand-piano effect" (but any instrument used for sonification can distract). Despite or because of this range of trade-offs, we believe that the system may offer some possible applications.

For example, we are currently exploring whether an interactive system of the described form might prove engaging to autistic children struggling with social interactions. They could choose a preferred instrument with which to engage. The delayed form of the s2m-system would then allow them to explore the feel of utterances and interactions (with themselves) without the social complexities they find so hard to negotiate.

At present, the system only includes sonification of voices. A next step is to include gestures (as described above) in order to demonstrate the tight coupling between speech and accompanying gestures. Hard- and software for this are in place, but experience with the ConversationPiece has shown that further careful design is required to ensure that the resulting sonifications are not overwhelming; any too unusual features, not explained by actual speech or action, will attract attention and distract from the communicative act – even if "Conversation is Easy," it still seems highly disruptable. At least, what is required is some first explorative or learning phase during which interlocutors become acquainted with the workings, possibilities, and quirks of the s2m-system in order to make efficient use of them afterwards. This may relate to the learning of a musical instrument (or one's voice) in general. To refer back to Pickering and Garrod: in a communication a mental world has to be created, which requires a bonding activity between the interlocutors. For now ConversationPiece forms bonds that are perhaps too strong between speaker and sonification machine; this counteracts its intended purpose as it distracts from the actual interpersonal communication. Ways are currently being explored to make s2m work more transparently to the speakers.

Conclusion

During the prototype performance, interlocutors became more aware of their voices and the collective soundscape that emerged through their outbursts of conversation. These moments led to a discourse on free improvisation whereby the sounds made

by the talkers became an emergent medium that was manipulated to construct a variety of sonic forms. Rather than a singular monologue which fell into an empty space, interlocutors responded to the voice of the other, creating a more enjoyable plural dynamic. ConversationPiece therefore offers a new, somewhat playful, vehicle for performative social interaction.

Acknowledgements

This work was supported by the Marie Curie Initial Training Network FP7-PEOPLE-2013-ITN (CogNovo, grant number 604764). The authors thank Jane Grant, Ilaria Torre, and Frank Loesche for many insightful comments that let to improvements of the manuscript.

References

Adank, P., Hagoort, P., & Bekkering, H. (2010). Imitation improves language comprehension. *Psychological Science, 21*(12), 1903–1909. doi:10.1177/0956797610389192

Friston, K. J., & Frith, C. D. (2015). Active inference, communication and hermeneutics. *Cortex, 68*, 129–143. doi:10.1016/j.cortex.2015.03.025

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences, 8*(1), 8–11. doi:10.1016/j.tics.2003.10.016

Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-brain coupling: A mechanism for creating and sharing a social world. *Trends in Cognitive Sciences, 16*(2), 114–121. doi:10.1016/j.tics.2011.12.007

Lieberman, P., & Blumstein, S. E. (1988). *Speech physiology, speech perception, and acoustic phonetics*. Cambridge, UK: Cambridge University Press.

Menenti, L., Pickering, M. J., & Garrod, S. C. (2012). Toward a neural basis of interactive alignment in conversation. *Frontiers in Human Neuroscience, 6*, 1–9. doi:10.3389/fnhum.2012.00185

Okada, K., Matchin, W., & Hickok, G. (2017). Neural evidence for predictive coding in auditory cortex during speech production. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-017-1284-x