Chatterbox: an interactive system of gibberish agents

Ronald Boersen[1], Aaron Liu-Rosenbaum[2], Kivanç Tatar[3], Philippe Pasquier[4]

[1,3,4]School of Interactive Arts & Technology, Simon Fraser University, Surrey, BC, Canada
[2]Faculté de musique, Université Laval, Québec, QC, Canada
[1]rboersen@sfu.ca, [2]aaron.liu-rosenbaum@mus.ulaval.ca, [3]ktatar@sfu.ca, [4]pasquier@sfu.ca

Abstract

We present the interactive multi-agent system Chatterbox, as part of the sound art installation Translanguaging, exploring the notion of translanguaging as a mediation of multilingual and intercultural communication. We discuss the act of languaging as a dual process comprising both semantic language communication, as well as paralanguage that relates to the affective, personal, and cultural aspects related to translanguaging. Through the creation of the Chatterbox agent, generating gibberish vocal streams devoid of semantic content, we aim at highlighting the paralinguistic dimension of languaging. The agent model comprises a kind of *gradient map*, clustering a segmented corpus of vocal sounds in the latent space of a self-organized map, according to its paralinguistic *fingerprint*. We utilize Factor Oracles for the creative generation of novel utterances of paralanguaging gibberish by the agent. Incorporating simple subsumption architecture inspired rules, we further moderate the interaction between the gibberish agents, creating rich and complex multi-agent behavior in "paralanguaging discussion". We outline the artistic and technical considerations in developing our Chatterbox agent throughout the paper. We share several observations made throughout the process of creating the Chatterbox agent, highlighting some of the connections between the notion of (trans)languaging and the implementation of our model.

Keywords

Translanguaging, Paralanguage, Multi-agent systems, Artistic research, Interactive art, Sound art, Generative art.

Introduction

In an attempt to understand the world around us, mankind has explored both the world through an objective perspective of the natural sciences, as well as how we make sense of this world through the subjective perspectives of the experience offered by the humanities and social sciences. The cognitive sciences have explored a plethora of viewpoints and cognitive models, from the computational theory of mind, information processing, and symbolism (e.g., Rescorla 2017), to more embodied forms of cognition, the perception-action relationships, and connectionism (e.g., Varela, Thompson, and Rosch 2017; Noë 2004). The development towards embodied cognitive paradigms has also seen parallel lines of thought arise in the field of music cognition (Leman 2008), and particularly the connection between perception and action is reflected through changes in concepts, such as the development of the term musicking.

The term *musicking* was coined by musician and educator Christopher Small (1998), and challenges the notion of music as a *thing*. Instead it proposes to think of music as an *activity*, where the act of musicking is inclusive of all aspects and actors involved in a musical performance, comprising not only the performance, but also involving the listening, rehearsing, practicing, and composing. In breaking the linguistic barriers and making *music* into a verb, Small has opened up new conceptual possibilities of how we think about music and the meaning that emerges out of the relations between the various acts and actors involved in its creation.

It is in this spirit that the collaborative project Translanguaging developed, comprising two interactive sound installations. The project was inspired by the research on translanguaging in learning by Prof. Angel Lin, Prof. Danièle Moore and Prof. Diane Dagenais as part of an interdisciplinary collaboration at the Faculty of Education,

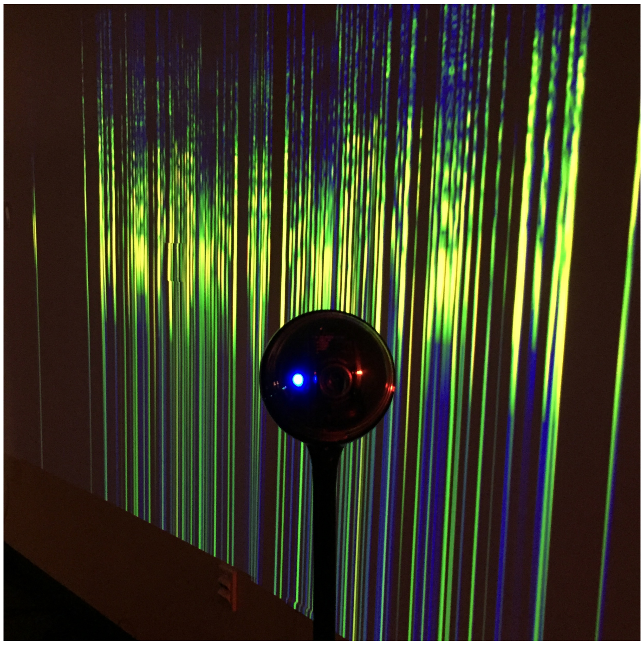


Figure 1. The camera and visualization used for Gesture-wording.

Simon Fraser University, conceived and led by Dr. Aaron Liu-Rosenbaum as Creative-in-Residence. The act of languaging was interpreted in line with the act of musicking, thinking of language not as an object but rather as an activity. The term *translanguaging* here refers to how multilingual speakers employ their multiple languages in the act of communicating, and how language use and language learning for multi-linguals become a negotiation amongst the different languages in one's repertoire, all of which co-exist symbiotically (Lin and He 2017).

These ideas were explored through sound, image, and gesture, resulting in a two-part interactive sound installation. The first part presented a Gesture-wording interactive installation (see Figure 1). Here visitors could "perform" new instances of words and phrases through the act of gesturing, exploring a virtual latent-space of vocal sounds. Simultaneously projected on the wall were spectral images of the respective frequencies contained in these "new words", visualizing the actions performed. In this way, gesture was de-materialized into sound, and sound was re-materialized into image in a multi-modal experience of languaging.

The second part of the installation, and the focus of this paper, invited visitors to join a gibberish conversation with four loudspeaker-agents (see Figure 2). Each loudspeaker presented a different software agent with its own "languaging" personality that spoke a gibberish language re-constructed from an actual language. Through the sound of their respective voices, the visitors and loudspeaker-agents had to negotiate when it was their turn to speak, being both polite in awaiting their turn as well as sometimes a little rude in anticipation and interrupting.

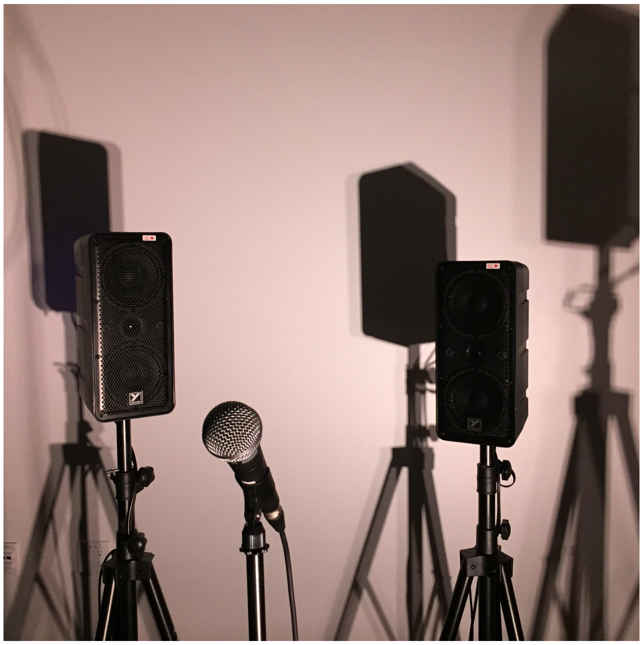


Figure 2. Chatterbox interaction microphone and two of the four speakers.

Conceptualizing a Chatterbox

The project evolved around the creation of a system, named Chatterbox – a gibberish-languaging agent in a multiagent system. The system culminated in an installation, where visitors were invited to join a gibberish conversation with four loudspeaker-agents (see Figure 2). Each loudspeaker was driven by a unique instance of Chatterbox and presented a unique "languaging" agent that spoke gibberish reconstructed from an actual language.

According to Truax (2016), when listening to language, we may speak of a form of dual processing where we perceive two streams of information simultaneously. An audio stream of a voice when speaking affords a listener with both semantic meaning through the organization of vocal sound, as well as paralanguage – those aspects of language that communicate *how* something was said. The paralanguage information stream comprises features such as "pitch inflections, timbre, dynamic changes in loudness, tempo and meter, patterns of stress, and […] the use of silence – exactly those attributes which are used to describe a musical melody" (p. 253).

By creating Chatterbox as a gibberish-languaging agent, removing the semantic information stream of a particular language, we aimed at putting emphasis on this paralinguistic dimension of languaging, and highlighting the affective, personal, and cultural aspects related to translanguaging. The agents construct their gibberish by recombining vocal sound segments, not according to any semantic measures, but by focusing on paralinguistic features, such as inflection and rhythm.

Consequently, we decided on the following base assumptions as guiding the design of the model: *Paralanguage as non-semantic language akin to music, containing both an alphabet and grammar, yet devoid of a semantic representation* (see also Nika, Chemillier, and Assayag 2016). Accordingly, we conceptualized sound as an infinite set that provides an alphabet of vocal sounds for any given language, constituting a subset of sounds relevant to a particular language idiom. The alphabet was structured according to its paralanguage grammar, here considered as musical motives and phrases – analogous to words and sentences, yet intrinsically devoid of semantic meaning. Following the concept of the infinite creativity in a Chompskian generative grammar utterances (Linson and Clarke 2017), we then utilized this paralanguage grammar as the constraint in the creative recombination of the alphabet into new utterances. Such grammar offered us compelling ways to inform a multitude of musical phenomena and generative processes, informed by what may be considered a syntactic mental representation of paralanguage structure.

With a focus on translanguaging, the agents were designed to (re)construct their gibberish based on four unique corpora of speech, each containing about one hour of a single speaker in a specific language. The four corpora comprise a female Chinese speaker; a male French speaker; a male English speaker; and a female speaker of Inuktitut, one of the Inuit languages of Canada.

The interaction of dialogue between the loudspeaker-agents, as well as the potential voices of visitors, is driven by the sonic dimension of the interaction alone, and occurs without the use of any visual components or other additional forms of communication. This poses a significant limitation in comparison to human agents that often use a range of cues in communication, such as lip movements, facial movements, and gestures of other speakers. For example, a conference call without video might invite more confusing interaction between speakers than a conference call *with* video. Accordingly, the agents have to negotiate their turn-taking based on the sound of their respective voices only. This behavior was implemented inspired by the subsumption architecture – an architecture originally developed for creating intelligent and complex behavior through the parallel layering of simple rules (Brooks 1991). A description of the implemented rules is elaborated in a later section, as we first turn to explain the design of the agents.

In this paper we describe the development and implementation of the interactive gibberish multi-agent system Chatterbox, as implemented within the interactive sound art installation Translanguaging. We present a number of related works that deal with concepts of language and translanguaging in an artistic context, before providing a report on the design of the agent from both a technical perspective as well as some of the artistic considerations underlying the development process. We conclude the paper with a discussion on a number of observations made in the process that highlight some of the connections between (trans)languaging and the implementation of Chatterbox.

Related Work

The ideas around language, translanguaging, and dual processing have been explored in many artworks in various mediums, a few of which we illustrate here. We begin by observing the work of Xu Bing (2011) and the development of his Square Word Calligraphy (see Figure 3 for an example of his style of calligraphy). At first glance the symbols may seem like traditional Chinese characters, though when asking Chinese viewers to read it, they will not be able to make sense of it. Further inspection reveals how the Chinese looking characters are actually readable by English readers as they are assembled out of English letters, in the case of the example spelling the combination "square word". Tong-King Lee (2015) examined Bing's work through a lens of translanguaging, where the dynamic communications of multi-linguals utilize linguistic and non-linguistic resources across semiotic boundaries when making sense of language. Lee further commented how verbal language may be seen as only one of the multi-modal semiotic resources available to a language user, introducing "an entire range of potentialities for meaning-making, which is not solely dependent upon verbal semantics, but also contingent on the specific configuration of text, mode, and medium in a particular communicative situation" (p. 444).

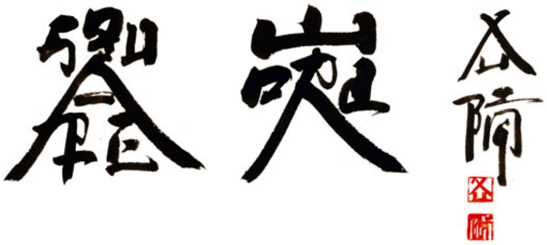


Figure 3. An example of Xu Bing's Square Word Calligraphy, spelling out the words "square word" and the artist's name and signature on the right (Bing, 2011)

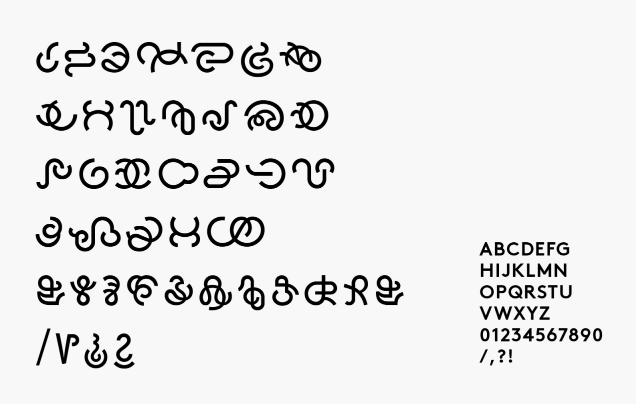


Figure 4. Illustration of the typeface "Backyard" created by Uri Katzenstein (2015)

Another artist, who explored the semiotic disconnect in language representation, is the Israeli multi-media artist Uri Katzenstein. For the exhibition of his work Backyard, he developed a typeface of alternative graphical representations of the letters in the English alphabet, which he referred to as "a hieroglyphic system developed as a common language for man and machine" (Peleg Rotem 2015). While there is much room to interpret the artist's intentions regarding his semiotic endeavors, the typeface would render readable English text incomprehensible. Through obscuring the semantic content of text, the hieroglyphs confront a viewer with the semiotic resources available in languaging when distinguishing abstract symbols as text.

While the previous discussed artists focused on the visual dimension of (trans)languaging, the work of Claude Gauvreau addresses the sonic and poetic dimensions of oral languaging in the creation of his deconstructed, reconstructed, and imaginary language "Explorean". The sonic qualities of Gauvreau's imaginary poetry were further explored in a theatrical performance of "Faisceau d'épingles de verre," where the text was performed utilizing computer-generated speech (Marceau 2005).

While the work of Claude Gauvreau addressed the auditory dimension of languaging through poetry, the aim of our Chatterbox installation is to additionally explore the *interactive* mode of translanguaging in a multi-agent system. We further address the disconnect in language representation, as present in the visual dimension of the work by Xu Bing and Uri Katzenstein, through the dual processing that occurs when languaging. By creating a disconnect from semantic meaning making, through the use of gibberish, we intend to highlight the multi-modal means available to us in the paralanguage dimension of communication.

The construction of gibberish

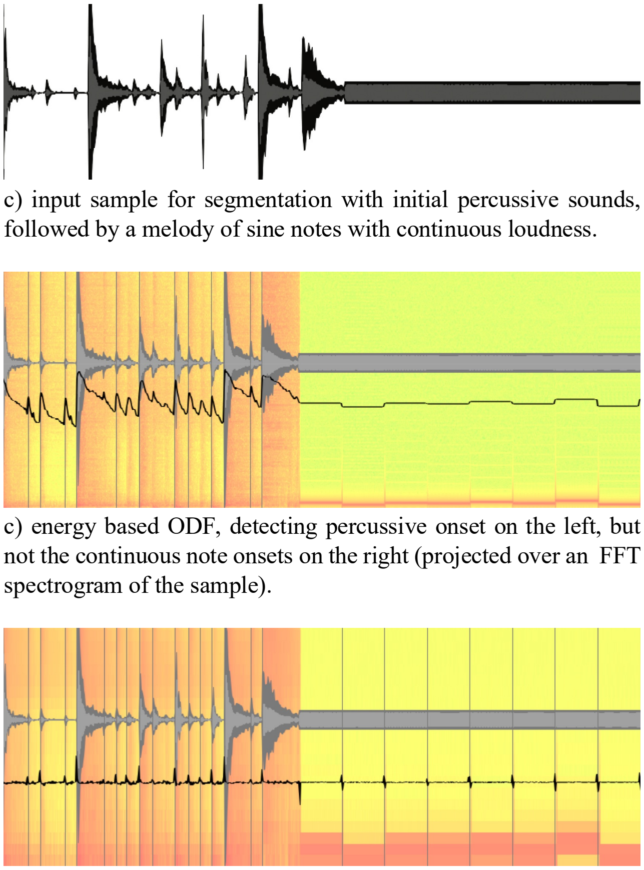
The Chatterbox agent model was developed utilizing the MASOM software agent architecture – a musical agent based on a self-organizing map (Tatar and Pasquier 2017) – and adapting the model to the objectives of this project. The model is implemented in the Max 7 environment and utilizes MuBu (Schnell et al. 2009), PiPo (Schnell et al. 2017), factorOracle (Wilson 2016), and a Self-Organizing Map (SOM) from the ml.* Machine learning toolkit, presented by Smith and Garnett (2012).

This section elaborates on the development of the alphabet, letters, and grammar at the core of the agent. The *alphabet* signifies the identification of a sound "memory"; the *letters* refer to a learning process that assigns letters to this alphabet; and the *grammar* refers to the generative and creative paralanguaging processes involved in turning the memory into novel instances of gibberish.

Segmentation

The memory of the system begins with the construction of its alphabet of sound events X, derived from a single-language audio corpus. The corpus is broken down into short audio segments at the approximate size of syllables, isolated from their semantic context.

The segmentation process identified the boundaries of audio events and placed markers between events within the corpus. This created a corpus of voice segments with an average duration of ±200ms. Traditionally the segmentation process has utilized the *spectral energy* of a given audio excerpt for its Onset Detection Function (ODF), identifying peaks in the signal that signify the onset of sonic events. This approach, however, presents difficulty in detecting event onsets in audio with continuous energy, yet containing sonically dissimilar events (see Figure 5b). Such a phenomenon was significantly present in the audio corpus, as speech contains continuity between syllables in words, and words are often slurred together to form sentences. In order to address this challenge, we opted to implement segmentation based on the temporal evolution of the spectral qualities of the signal. Instead of energy, we use the temporal evolution of the magnitude spectrogram of Mel frequency bands $S_{\text{mel}}(t, b)$, also known as *melFlux* (Böck, Krebs, and Schedl 2012), which shows significant improvement in detecting these onsets as elaborated below (see Figure 5c).



c) input sample for segmentation with initial percussive sounds, followed by a melody of sine notes with continuous loudness.

c) energy based ODF, detecting percussive onset on the left, but not the continuous note onsets on the right (projected over an FFT spectrogram of the sample).

c) melFlux based ODF, detecting both percussive and continuous notes (projected over a 32-band mel spectrogram of the sample).

Figure 5. Illustration of traditional Onset Detection Function in comparison with melFlux ODF

In practical terms, we calculated the magnitude frequency spectrum $S(t, k)$ utilizing an FFT (where t denotes the frame index and k the frequency bin number) with a window size of 23ms (equivalent to 1024 samples at a sample rate of 44.1kHz) and a window hopsize between consecutive frames of 1.45ms (64 samples). The linear magnitude spectrum was subsequently recalculated to the logarithmic frequency representation of the mel scale, a perceptual frequency spectrum more humanly and musically relevant. We achieved this by filtering the linear magnitude spectrum through a 32 bin Mel frequency filter $M(k, b)$ (where b denotes the filter bin number). This process is described by the following formula:

$$S_{mel}(t, b) = S(t, k) * M(k, b)$$

We additionally scaled the output of this filter logarithmically, as this seems to greatly improve performance (further supported by Böck, Krebs, and Schedl 2012).

The melFlux onset detection function $mF(t)$ is subsequently derived by calculating the difference between consecutive frames, with an average calculated over $\delta = 21$ frames (this number was determined through trial and error). The final onset detection function $mF(t)$ is then given by:

$$mF(t) = \sum_{b=1}^{b=32} (S_{mel}^{log}(t, b) - S_{mel}^{log}(t - \delta, b))$$

As the $mF(t)$ contains both spectral, as well as energy information, it allows for a more musically-informed means to find event onsets, including onsets that do not appear

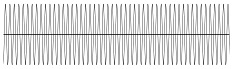
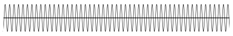
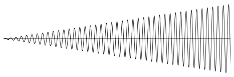
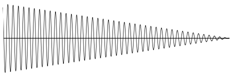
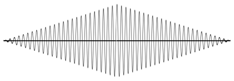
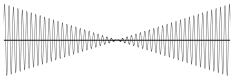
	Absolute loudness	**Loudness slope**
a) Absolute feature more successful		
	$\mu L(x) = 1.0$	$\mu\Delta L(x) = 0$
	$\mu L(x) = 0.5$	$\mu\Delta L(x) = 0$
b) Slope feature more successful		
	$\mu L(x) = 0.5$	$\mu\Delta L(x) = 1.0$
	$\mu L(x) = 0.5$	$\mu\Delta L(x) = -1.0$
c) Neither particularly successful		
	$\mu L(x) = 0.5$	$\mu\Delta L(x) = 0$
	$\mu L(x) = 0.5$	$\mu\Delta L(x) = 0$

Figure 6. Illustration of successes and failures of both absolute and slope feature values, when observing for example loudness.

within the energy contour alone (see Figure 5). The exact threshold for identifying event boundaries was determined through experimentation with the corpus. In practice it provided a robust ODF that outperformed the energy based ODF in determining onsets in the language corpora.

Paralanguage feature labeling

After the segmentation process, the identified segments $x \in X$ were labeled with an n-dimensional feature vector $F(x)$, providing a fingerprint of audio descriptors focused on the paralinguistic features to be used as our final alphabet X. As the paralinguistic features reside in the inflections and changes of the voice, we concentrated on a differential description of the voice features within each segment, including the change of *Loudness* $\left(\Delta L(x)\right)$ from beginning to end, to identify the overall dynamic development; change in *Zero-crossings* $\left(\Delta Z(x)\right)$ as indicator of pitch-inflection and/or noisiness; the change in 13 *Mel Frequency Bands* $(\Delta B_n(x),\ n = 0, \ldots, 12)$, indicating changes in timbre; and *Duration* $\left(D(x)\right)$ as an indicator of rhythm.

Each segment was sliced in 23ms frames with a hopsize of 11.6ms, and the changes in features were calculated between consecutive frames. Rather than analyzing the absolute values of features contained in each segment, we computed the mean and standard deviation of the *slope* of the individual features, i.e. the slope of the respective contours over the segment. This provided a 31-dimensional fingerprint containing $D(x)$ for duration, μ and σ for $\Delta L(x)$, μ and σ for $\Delta Z(x)$, and μ and σ for $\Delta B_n(x)$.

The focus on differential data is further motivated by the fact that the corpus consists of segments from a single speaker. As the loudness, pitch, and timbre of a single person's voice likely remain within a fairly limited range, the mean of the absolute features would likely average out as well. This suggests that the absolute features would offer little information for the vocal characterization of segments. While the differential data would indeed offer little support for identifying absolute pitches and loudness of the voice sounds, it offers a greater emphasis on paralinguistic inflections through *changes* in pitch, loudness, and timbre (see Figure 6).

A latent space of syllables

Having populated our alphabet X with the fingerprints of each segment $F(x)$, we subsequently used our alphabet as a data set for the training of a Self-Organizing Map (SOM).

SOMs are a form of artificial neural networks that utilize unsupervised training (Kohonen 1998). They are used to map a high-dimensional feature spaces onto a 2-dimensional map – organizing, classifying, and clustering the feature space, through the calculation of proximity as similarity of high-dimensional feature sets. In the case of our agent, we utilized a SOM to provide our model with a 2-dimensional representation of the n-dimensional fingerprint of our alphabet. The 2-dimensional map offered an easy-to-navigate "landscape" of sonic material that represents a topological coherence of the original feature space of the segments – a latent space containing the memory of the audio corpus. As the fingerprints $F(x)$ contain differential data, indicating the slope of features within each segment, we regard the latent space of our SOM as a kind of *gradient map* of sonic features ∇F_{SOM}.

We subsequently assigned each segment x to a node coordinate $node_i$ on the SOM according to the best matching unit function $BMU\left(F(x)\right)$ that indicates the nearest node to a particular fingerprint. The nodes of the SOM are then used to cluster various segments in our library according to the respective similarity of their feature array $F(x)$, meaning that multiple segments may be mapped to a single node. As our gradient map ∇F_{SOM} does not represent absolute descriptors of sound, segments that are clustered together on a single node of our SOM might contain entirely different timbres, yet still share similar features of languaging inflections.

(Re)constructing non-semantic meaning

Having assigned a node coordinate to each segment, we created a node sequence that followed the order in which the segments appeared in the original audio. We then used this order to encode what is called a Factor Oracle (FO).

FOs are a tool originally developed for data compression (Allauzen, Crochemore, and Raffinot 2002) that offer great potential in creative applications. By encoding the representation of structure of a string, FOs allow for the efficient querying of a string for existing substrings (also known as factors) through the inclusion of forward-links; additionally, they enable the identification of repeating patterns through the encoding of suffix-links. Additionally, the FO representation provides a powerful tool for the generative exploration of novelty, utilizing the encoded paralanguaging grammar built from the provided corpus. This strategy presents us with a promising form of style imitation that explores the creative space presented by the original corpus.

By assigning each segment in our corpus with a coordinate in our latent space, we represented the corpus as a string of node coordinates, encoding the "route" a particular sequence of segments represents on the gradient map. For example, if the first segment was mapped to $node_a$, the second to $node_c$, the third to $node_b$, and the forth to $node_c$, we get the sequence:

$$s = node_a, node_c, node_b, node_c$$

Consequently, the clustering of segments on the SOM nodes according to their similarity, offered the possibility to identify patterns within the corpus. It is these repeating patterns of consecutive nodes that provided the motives and phrases for the construction of our paralanguaging grammar.

To conclude, by building an FO out of the string of nodes as encountered in the original audio, we encoded the paralanguaging grammar of our corpus. Subsequently, in the generative phase of our agent, we used the FO to explore and predict the order of segments as encoded. Utilizing the generative capabilities of the FO, the agent was able to speak new utterances devoid of semantic content, by playing node sequences as generated by the FO – thus creating a paralanguaging gibberish agent.

The emergence of dialogue

Having established the paralanguaging ability of the gibberish agent Chatterbox, we are now faced with the task of getting the agents to enter into a dialogue. In order to create a dynamic and interesting behavior of the agent's turn taking, we drew inspiration from the subsumption architecture in layering simple rules for the creation of complex behavior (Brooks 1991). We subsequently defined a number of heuristics that govern the turn taking behavior of the model.

At the base of the model's subsumption architecture we defined the agent's listening capabilities to detect silence and start speaking, as well as to stop speaking at the end of an utterance. This constitutes the first layer of the model's heuristics:

- **Layer 1: "be chatty"**
  - *Start speaking when detecting silence*
  - *Stop speaking at the end of an utterance*

Through experimentation we found that the node with the minimum sum of all features in the latent space of the SOM, offers a reasonable indication of an end-of-utterance. Accordingly, we defined the agent to stop speaking when the node being played crossed a certain threshold, defined as a Euclidian distance from the minimum on the map.

While this rule would successfully govern the turn-taking between two agents, the addition of a third agent would cause two agents to respond and continue talking simultaneously. We therefore added another layer to check whether other agents are talking:

- **Layer 2: "politely test the waters"**
  - *If after n-segments there is silence, continue speaking*
  - *Otherwise stop speaking*

By defining the checkpoint after n-segments (default n=3), small variations in segment length between the various uttering agents, determine which agent will continue speaking and which agent will stop.

In order to also include the anticipatory nature of how people interact through interrupting each other in a dialogue, we allowed the agents to do the same when expecting another agent to stop speaking:

- **Layer 3: "also be a bit rude"**
  - *Start speaking when predicting another agent's end of an utterance*

The ability to anticipate the end of a utterance was implemented through the use of an FO conditioned by the same sequence that also governs the speaking of the agent. By analyzing the input audio-stream according to the same 31 paralanguage features on which the agent was trained, the agent can match the input-speaker to a current node on its own SOM. Subsequently, the agent can predict a possible continuation of the external speaker with the use of its FO, rendered from the subjective perspective of how it would possibly continue the utterance itself. When the agent anticipates the external speaker to enter into the minimum of the latent space, suggesting a potential end-of-utterance, it may start speaking and somewhat rudely interrupt.

To promote dialogue to occur between two distinct agents, we added another layer to encourage the agent to respond when testing the waters in layer 2:

- **Layer 4: "engage in dialogue"**
  - *Gradually increase n while speaking, encouraging engagement in discussion.*

As a final layer, we moderated the discussions by encouraging agents to stop uttering when other are speaking:

- **Layer 5: “moderate discussion”**
  - *Gradually increase proximity threshold to end-of-utterance when speaking simultaneously.*

By enlarging the threshold that governs layer 1 to stop speaking, the agent is gradually encouraged to stop speaking when other agents are speaking concurrently.

It should be emphasized that each agent is trained on a different corpus with a unique voice and a different language. As such, the inclusion of the subjective dimension of anticipation is what promotes the translanguaging behavior of the multi-agent system. In its essence, the subjective perspective of the agent in perceiving and responding to its environment, encourages the agent to interpret, and quite possibly misinterpret, the other agents. However, the agent (mis)interprets and foresees its environment according to how it would expect to behave itself.

Discussion

We have presented the interactive multi-agent system Chatterbox, as part of the sound art installation Translanguaging, exploring the notion of translanguaging as a mediation of multilingual and intercultural communication. We have discussed the act of languaging as a dual process comprising both semantic language communication, as well as paralanguage that indicates the affective, personal, and cultural aspects related to translanguaging. Through the creation of the Chatterbox agent, generating gibberish vocal streams devoid of semantic content, we aimed to highlight the paralinguistic dimension of languaging.

Additionally, we provide a technical description of the Chatterbox agent model, and outlines the artistic motivations for the technical considerations of the implementation. The agent model comprises a *gradient map* that clusters a segmented audio corpus in the latent space of a SOM. The audio segments are assigned a node in the latent space of the SOM according to their paralinguistic *fingerprint*. Describing the audio corpus as a sequence of nodes on the SOM, we construct an FO for the encoding of the paralanguage *grammar*, and use the FO for the creative generation of novel utterances by the agent according to the encoded grammar. Incorporating simple subsumption architecture inspired rules, we further moderate the interaction between the gibberish agents, creating rich and complex multi-agent behavior in “paralanguaging discussion”.

While the above describes the final implementation of the gibberish agent model, we would like to share several observations we made throughout the process of creating the Chatterbox agent, highlighting some of the connections between the notion of (trans)languaging and the implementation of our model.

Firstly, after implementing the first three layers of the turn-taking heuristics, the agents were beginning to display rich and complex behavior when interacting with one another. We tested the system and its interaction with a human agent through a microphone in a studio environment using a simple small speaker setup. We worked on fine-tuning the model in this environment, before working with the eventual 4 speaker setup, until the agents behaved satisfactory and the interaction was engaging. However, when eventually scaling the setup and having the speaker-agents spread out comparable to how human agents would converse around a table, a curious phenomenon emerged: the human agent interacting through a microphone would respond to a particular speaker agent, and expect the same agent to respond. In actuality however, often another agent would respond causing some cognitive confusion for the human agent, needing to redirect the attention and physically rotate in order to face the newly engaged speaker agent. Surprisingly, this conflict was entirely missed when observing the agents from a localized perspective of the single speaker setup, and only became apparent when human embodiment became part of the interaction, having to rotate one’s body to align with our perceptual focus. This observation led to the addition of the fourth heuristics for the interaction, encouraging the agents to engage in local two-agent dialogues within the larger multi agent discussion.

Secondly, we observed how the agents would often start speaking simultaneously as they were implemented with the same heuristics. This behavior created a sense of commotion in an often dense gibberish soundscape, and made the interaction seem somewhat hectic. The addition of the fifth layer in the turn-talking heuristics was intended to moderate the discussion accordingly and encourage agents to let the others speak, limiting the undesired behavior (this in addition to the second layer already stating to “test the waters” before continuing to speak). While this indeed improved the interaction, we noticed it did not eliminate the recurrent commotion altogether. However, rather than regarding this as a failure of the model, this behavior may actually be indicative of the multimodal dimension of human communication. Akin to a conference call without video, the agents were limited in medium to negotiate their turn and ended up “accidentally” speaking together. We recalled how this issue also often emerged during the pre-video conference calling era. Any improvement for this behavior would likely come from incorporating a multimodal dimension to the interaction, such as providing the agents with visual feedback, as possible future directions for developing the Chatterbox agents.

As a final observation, we noticed that when the agents would speak with one another, there seemed to be a lacking connection in how one agent would answer the other. Admittedly, the agents were merely generating sentences according to *their own* FO, exploring the latent space of the SOM; however, the challenge lay in how to create the connection in between responses when there is no semantic content. The approach for addressing the challenge was motivated from the same subjective perspective explored in the third heuristic, where the agent anticipates another agent to finish speaking by mapping the input to its own latent space. By fingerprinting the input audio-stream of the other agents, the agent can map and interpret the input as a node sequence from the subjective perspective of how it would paralanguage itself. Subsequently, when initializing speaking, the agent can prime the FO governing its paralanguaging gibberish, offering a subjective interpretation of continuation of the external speaker. This solution follows a similar line of thought as the Continuator (Pachet 2003) in providing a stylistically consistent continuation between human and machine, both creating and anticipating according to a concept of style-imitation.

To conclude with a quote from Prof. Angel Lin, discussing her experience of the Chatterbox installation in a panel of language education researchers:

> "I think you can't escape that kind of feeling of […] otherness, when you hear something that you don't understand, chopped into bits and pieces. […]. It kind of represents what you feel when you're encountering someone from a different culture, from a different background, who speaks with a different accent, or different tone, or just a different style, or just a different rhythm. That kind of alienness or otherness is so well represented there […], it is the beauty of what it is."

Acknowledgements

This research was supported by the Social Sciences and Humanities Research Council of Canada. The project was commissioned and supported by the Faculty of Education at the Simon Fraser University under the guidance of Dean and Prof. Susan O'Neil. The work was inspired by the collective research of Prof. Diane Dagenais, Prof. Angel Lin, and Prof. Danièle Moore, as part of their research in Languages and Societies at the Faculty of Education.

References

Allauzen, Cyril, Maxime Crochemore, and Mathieu Raffinot. 2002. "Factor Oracle: a New Structure for Pattern Matching." In *SOFSEM'99: Theory and Practice of Informatics, Lecture Notes in Computer Science*, edited by Pavelka J, Tel G, and Bartošek M, 1725:295–310. Berlin, Heidelberg: Springer Berlin Heidelberg.

Bing, Xu. 2011. *Square Word Calligraphy: Square Word*. http://www.aaa-a.org/events/xu-bing-square-word-calligraphy-classroom/.

Böck, Sebastian, Florian Krebs, and Markus Schedl. 2012. "Evaluating the Online Capabilities of Onset Detection Methods." In, 1–6.

Brooks, Rodney A. 1991. "Intelligence Without Representation." *Artificial Intelligence* 47 (1-3): 139–59.

Kohonen, Teuvo. 1998. "The Self-Organizing Map." *Neurocomputing* 21 (1-3): 1–6.

Lee, Tong-King. 2015. "Translanguaging and Visuality: Translingual Practices in Literary Art." *Applied Linguistics Review* 6 (4): 75–26.

Leman, Marc. 2008. "Paradigms of Music Research." In *Embodied Music Cognition and Mediation Technology*, 27–49. Cambridge, MA: MIT Press.

Lin, Angel M Y, and Peichang He. 2017. "Translanguaging as Dynamic Activity Flows in CLIL Classrooms." *Journal of Language, Identity & Education* 16 (4). Routledge: 228–44.

Linson, Adam, and Eric F Clarke. 2017. "Distributed Cognition, Ecological Theory and Group Improvisation." In *Distributed Creativity: Collaboration and Improvisation in Contemporary Music*, edited by Eric F Clarke and Mark Doffman, 1:52–69. Distributed Cognition, Ecological Theory and Group Improvisation. Oxford University Press.

Marceau, Andre. 2005. "Review of [Faisceau D'épingles De Verre D'après « L'objet Dramatique » De Claude Gauvreau : Du Mélange Dans Les Genres... Une Petite Étude De Cas]." *Inter*, no. 91: 42–44. https://www.erudit.org/en/journals/inter/2005-n91-inter1120238/45787ac/abstract/.

Nika, J, Marc Chemillier, and Gérard Assayag. 2016. "ImproteK: Introducing Scenarios into Human-Computer Music Improvisation." *Computers in Entertainment* 14 (2): 1–27.

Noë, Alva. 2004. *Action in Perception*. Cambridge, MA: The MIT Press.

Pachet, François. 2003. "The Continuator: Musical Interaction with Style." *Journal of New Music Research* 32 (3): 333–41.

Peleg Rotem, Hagit. 2015. ""Hatzer Aharonit": Te'aruchah Chadashah Shel Uri Katzenstein beMuzeon Tel Aviv (Translation: "Backyard": a New Exhibition by Uri Katzenstein at the Tel Aviv Museum)." https://www.globes.co.il/news/article.aspx?did=1001062736.

Rescorla, Michael. 2017. "The Computational Theory of Mind." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N Zalta, 2017 ed. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2017/entries/computational-mind/.

Schnell, Norbert, Axel Röbel, Diemo Schwarz, Geoffroy Peeters, and Riccardo Borghesi. 2009. "MuBu and Friends - Assembling Tools for Content Based Real-Time Interactive Audio Processing in Max/MSP." In. Montreal.

Schnell, Norbert, Diemo Schwarz, Joseph Larralde, and Riccardo Borghesi. 2017. "PiPo, a Plugin Interface for Afferent Data Stream Processing Modules." In *International Symposium on Music Information Retrieval*. Suzhou, China.

Small, Christopher. 1998. *Musicking: the Meanings of Performing and Listening*. Middletown, CT: Wesleyan University Press.

Smith, Benjamin D, and Guy E Garnett. 2012. "Unsupervised Play: Machine Learning Toolkit for Max." In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 1–4.

Tatar, Kıvanç, and Philippe Pasquier. 2017. "MASOM: a Musical Agent Architecture Based on Self Organizing Maps, Affective Computing, and Variable Markov Models." In *Proceedings of the International Workshop on Musical Metacreation*.

Truax, Barry. 2016. "Acoustic Space, Community, and Virtual Soundscapes." In *The Routledge Companion to Sounding Art*, edited by Marcel Cobussen, Vincent Meelberg, and Barry Truax, 253–63. New York: Routledge.

Varela, Francisco J, Evan Thompson, and Eleanor Rosch. 1991/2017. *The Embodied Mind: Cognitive Science and Human Experience*. 1st ed. Cambridge, MA: The MIT Press.

Wilson, A. J. 2016. factorOracle: an Extensible Max External for Investigating Applications of the Factor Oracle Automaton in Real-Time Music Improvisation. In *Proceedings of the International Workshop on Musical Metacreation*, Paris, France.