FONDS FRANCE CANADA POUR LA RECHERCHE
FRANCE CANADA RESEARCH FUND

2022 « New Research Collaboration Program »

Application form

(aussi disponible en français)

Pre-screening
Please respond to the following questions

Eligibility questions:

1/ Have you ever received FCRF funding with your partner or with one of the members of the partner team? **NO**

2/ Have you ever signed a doctoral or postdoctoral contract with the partner team? **NO**

3/ Does your project consist of the organization of a joint symposium? **NO**

If you answered YES to any of these questions, you are not eligible to apply

Secondary questions:

1/ Are doctoral students included in your team or are they associated with the project? **YES**

2/ Are young researchers involved in the project? **YES**

3/ Do you intend to apply to any complementary FCRF financing programs? (see Appendix 1 in the guide) **YES** : **MITACS GRA, L'Oréal-FCRF**

4/ Your project is related to one of the following field(s), please check (several can be checked): **Information Technology/ Security/ Communications, Humanities/ Cultural Diversity/ Francophony/ Society**

Project title: *The Vocal Characteristics of Engaging Second Language Teaching: An Interdisciplinary Study using Artificial Intelligence and Psychoacoustics*

Scientific field:

X Math/Applied mathematics
☐ Physics
☐ Earth & Space sciences
☐ Chemistry
X Engineering sciences
☐ 2022 Focus: Societal impacts of the COVID-19 pandemic
☐ Biology, Health, Medicine
☐ Human sciences
X Social Sciences
X ITSC
☐ Agronomy, Vegetable & Animal Production

1. Partners

	French team	*Canadian team*
Lead researcher		
Surname, given name	AUCOUTURIER, Jean-Julien	LIM, Angelica
Position / grade	Senior CNRS researcher	Assistant Professor
Telephone	+33-06.47.07.20.57	+1.778.980.6568
Fax		
Email	aucouturier@gmail.com	angelica@sfu.ca
Laboratory		
Name	FEMTO-ST Institute	ROSIE Lab, School of Computing Science
Identification No.	UMR 6174	N/A
Director (Surname, given name)	LARGER, Laurent	LIM, Angelica
Home institution	Centre National de la Recherche Scientifique (CNRS)	Simon Fraser University
University of affiliation (if applicable)	Université de Bourgogne Franche-Comté	
Address	24 rue Alain Savary, Besançon, FR	8888 University Dr, Burnaby, BC
Postal code	25000	V5A 1S6
Telephone	+33-03.63.08.24.00	+1.778.782.4277
Fax		
Email	contact@femto-st.fr	cs_assistant@sfu.ca

Names of research unit directors **SIGNATURE**	LARGER, Laurent	

2. Project description

Project focus

How can English speakers more effectively learn French, and how can French speakers more effectively learn English? What words should be stressed, how should one pause -- what does the most effective second language (L2) teaching voice sound like?

Speech, the most fundamental form of communication between humans, has evolved to maintain a listener's attention by modulating characteristics such as pitch, volume, timbre, and variability [1]. Recent works have studied engaging speakers such as Steve Jobs to understand how exactly they use voice to captivate audiences [2,3]. Although we are beginning to understand the general qualities of an engaging speaking voice (e.g., a higher-than-average pitch, greater pitch range, few disfluencies such as “um” or “uh”, and intensified leading consonants and lengthened vowels [2]), we remain much less certain on how to replicate this speech. When exactly do we need to lengthen vowels vs. consonants? When exactly should we pause for emphasis? And when should we vary our pitch range to engage our audience?

This project combines Canadian expertise in artificial intelligence (AI), along with French expertise in electroencephalogram (EEG) and voice psychoacoustics to **build a multilingual model of engaging speaking voice.** The goals of this project are twofold: (1) Develop an **artificial voice model** (for both English and French) that is optimized for maximum retention of message, with **L2 learning as a use case** (2) Understand and **explain the characteristics of engaging voice** towards public impact and social good, e.g. helping teachers improve their speech, improving public dissemination and retention of public health messages.

Potential future outcomes of this project also include interactive voice programs that can effectively support English or French learning using specialized text-to-speech (TTS) for teaching and conversation practice. Furthermore, as we focus on understanding cross-linguistic characteristics of effective teaching voice, this work can potentially be applied to other languages, especially those which have limited language teacher resources such as endangered Indigenous languages.

[1] R. I. Dunbar “The social brain hypothesis”. *Evolutionary Anthropology: Issues, News, and Reviews*, vol. 6(5), 178-190, 1998.
[2] O. Niebuhr, J. Voße, and A Brem, “What makes a charismatic speaker? A computer-based acoustic-prosodic analysis of Steve Jobs tone of voice,” Computers in Human Behavior, vol. 64, pp. 366-382, 2016.
[3] O. Niebuhr, J. Michalsky “Computer-Generated Speaker Charisma and Its Effects on Human Actions in a Car-Navigation System Experiment - or How Steve Jobs’ Tone of Voice Can Take You Anywhere,” in Proc Computational Science and Its Applications – ICCSA 2019, vol 11620, pp. 375-390, 2019.

Methodology

The project consists of a 3-year collaborative plan, of which 2 years will seek support from FCRF.
Data preparation. The Canadian team is running qualitative interviews with professional voice coaches who teach teachers how to speak, as well as collecting data from public sources to use as examples of effective teaching voice. The first data source includes professional podcasts, Youtube L2 teaching resources, and language learning recordings. Secondly, we are using a citizen science website to collect vocal samples of people using their best teaching voice, as well as public perceptions about how they best learn. Finally, the voice samples will be analyzed using online participants to rate the voices using declarative measures (e.g. "how effective is this person at teaching?"). The limitations of this data collection is that **collected databases**

are not flexible; i.e. the words that each teacher uses are not the same, and personal vocal characteristics differ; consequently scientific conclusions cannot easily be drawn. Therefore, an artificial voice will be developed and evaluated using a combination of machine-learning and psychoacoustic techniques to understand what characteristics (here, also called "features") are important, and how they should be used.

Task 1. Developing an artificial voice model for maximum retention in language learning.
Most artificial voice generation algorithms trained on large amounts of data produce voices that are relatively human-like and useful for practical tasks like voice navigation (e.g. Siri, Alexa). What is unclear is how exactly the resulting voices impact those who listen to them and how well they can be used for teaching. A recent paper by the **Canadian team** (Lim et al.) proposed a generative algorithm using deep learning that takes text as an input and produces an expressive dynamic gestural sequence corresponding to the text [1]. The resulting dynamic movements were found to be more diverse, appropriate and human-like than baseline models. In this task, the Canadian team will adapt their method to **vocal prosody** by replacing the method's gestural features with audio features. This approach is particularly adapted to our task because (a) it uses a vector-quantized variational autoencoder, which is an **unsupervised machine learning model adaptable to both English and French**, and (b) the trained model also produces an **interpretable** codebook that can be used to understand the features of resulting output, while retaining human-likeness. Using this model, along with the feature-to-speech psychoacoustic package by the **French team** [4snip] we will create a **text-to-prosodic speech system** and apply it to **both English and French**. The model will be trained using a mixture of both normal and engaging voice data sources, and using flow [8] or conditional VAE [9] techniques, we will tailor the latent feature space to allow systematic modulation of features identified in the Data Preparation step, for use in Task 2.

Task 2. Understand and explain the characteristics of the engaging teaching voice.
Secondly, we will use the following methods to examine what exact parameters make the voice engaging and effective for teaching, compared to standard TTS systems, and disentangle the parameters over both English and French. Three studies will be performed.

a) <u>Reverse correlation in France on French-native speakers</u> Psychoacoustic reverse correlation is a data-driven technique developed to uncover what signal features carry information for human sensory processing, by analyzing participant responses to large sets of systematically-varied stimuli [2]. The French team (Aucouturier et al.) has recently explored the application of reverse correlation to reconstruct various prosodies including social judgments of dominance and trustworthiness [3] or the reliability of information [4] . We will similarly use a reverse-correlation approach to help us understand the components of an engaging voice, when they occur, and how exactly they need to fit together relative to one another. The reverse correlation methodology involves having individual listeners rate hundreds or thousands of pairs of sounds, randomly sampled from our generated voices, in the case of this research rating which in a pair is more engaging. Participant responses are then analyzed a posteriori, using techniques such as classification images [2], to automatically discover what signal features drive judgments of whether a given voice is thought engaging. We will finally validate that these signal features impact using real L2 learning tasks, e.g. using a similar word retention task as the French team did previously in [4] (see also Task 2c below for French as an L2).

b) <u>Electroencephalogram (EEG) entrainment study</u> An electroencephalogram (EEG) entrainment study will be used as a method to further verify our results and clarify the physiological bases of how the vocal features identified in Task 2 (a) drive psychological engagement. Psychology and neuroscientific studies using both EEG and functional magnetic resonance imaging (fMRI) have shown that brain-to-brain synchrony between both students and speakers (teachers) and within student groups is positively associated with learning outcomes [5, 6]. Furthermore, student engagement could be predicted by group synchrony, suggesting that increased engagement can result in increased synchrony and in turn increased cognitive gains [7]. Using a similar method to research by Dikker [5] done in a classroom with human teachers we will study brain-to-brain synchrony amongst students and between the students and teaching voice during language lessons provided using our generated voice. Pre and post test methodologies testing students' retention of the lesson, as well as surveys to assess how engaging and natural the students find the voice will also be employed.

c) Reverse correlation in Canada on English-native speakers The same methodology as (a) but with English speakers in Canada learning French will be conducted.

Our aim is to have Task 2 (a) and (b) funded by FCRF for mobility of PhD candidate Paige Tuttosi to France.

[1] P.J. Yazdian, M. Chen, A. Lim. "Gesture2Vec: Clustering Gestures using Representation Learning Methods for Co-speech Gesture Generation". Submitted to *The Tenth International Conference on Learning Representations*, 2022.

[2] R. Adolphs, L. Nummenmaa, A. Todorov, JV. Haxby. "Data-driven approaches in the investigation of social perception". *Phil Trans R Soc B*., vol. 371(1693):20150367, 2015.

[3] E. Ponsot, J.J. Burred, P. Belin, and J.J. Aucouturier. "Cracking the social code of speech prosody using reverse correlation," *Proc. National Academy of Sciences*, vol. 115, no. 15, pp. 3972-3977, Apr 2018.

[4] L. Goupil, E. Ponsot, D. Richardson, G. Reyes, JJ. Aucouturier. "Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature." *Nature communications* 12.1 (2021): 1-17.

[5] G.J. Stephens, L.J. Silbert, and U. Hasson. "Speaker-listener neural coupling underlies successful communication," Proc. National Academy of Sciences, vol. 107, no. 32, pp. 14425-14430, Aug 2010.

[6] I. Davidesco et al. "Brain-to-brain synchrony between students and teachers predicts learning outcomes" bioRxiv, Preprint.

[7] S. Dikker et al. "Brain-to-Brain Synchrony Tracks Real-World Dynamic Group Interactions in the Classroom," Current Biology, vol. 27, no. 9, pp. 1375-1380, 2017.

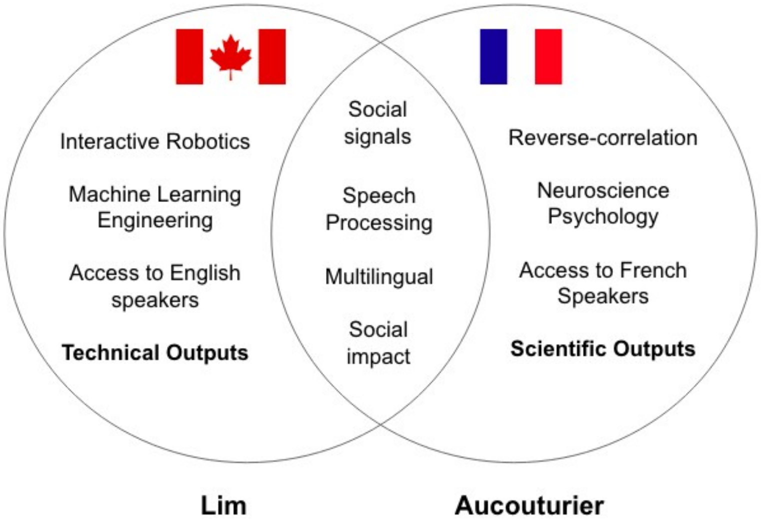
[8] D.P. Kingma, P. Dhariwal. "Glow: Generative Flow with Invertible 1x1 Convolutions", Neurips, 2018

[9] M. Marmpena, F. Garcia, A. Lim. "Generating Robotic Emotional Body Language of Targeted Valence and Arousal with Conditional Variational Autoencoders." ACM/IEEE International Conference on Human Robot Interaction,pp. 2020

3. Team overview (provide CVs in the Appendix: maximum of 4 CVs)

Composition of teams involved in the project *(name, position, grade, % of involvement in the project)*
France:
Researchers AUCOUTURIER, Jean Julien, senior CNRS researcher *(Directeur de recherche CNRS),* involved at 17.5% fulltime (40% during 6m visit, 10% remaining 18m), i.e. 50% of total PI involvement
Canada:
Researchers LIM, Angelica, Assistant Professor, involved at 17.5% fulltime (20% during 18m Canada supervision, 10% during 6m travel), i.e. 50% of total PI involvement
Students *(jointly-supervised doctorates, potential joint PhDs)* TUTTOSI, Paige, PhD Candidate, Year 1

4. Synergy between the French and Canadian research teams



The **Canadian** and **French** teams form a strong **interdisciplinary** team with **complementary** skills. Lim (computer science) and Aucouturier (psychology, neuroscience) are both experts in **speech and audio processing**, giving them a **common technical language and understanding of vocal features**. At the same time, they approach speech problems from different perspectives: Lim has focused on the **technical development** of **artificial intelligence** algorithms to make **machines more socially intelligent** [1], and Aucouturier has focused on **understanding the social intelligence of humans** using the **scientific method** [2].

On one hand, psychologists face challenges to create stimuli to properly study vocal features over long periods of time, and require a method to generate flexible, realistic voices for experimental studies that are also human-like (e.g. produces stresses on probabilistically common words). On the other hand, computer scientists, while able to develop such technologies, are not equipped with methods such as reverse correlation to understand how their generated models impact those who use them.

In this way, the proposed project creates virtuous synergies: the Canadian team will **generate parameterizable artificial voices** for which the French team will **validate experimentally and give feedback** towards maximum economic and social impact. **<u>Both English and French will be used as target languages</u>**, which is not common for prosody research (usually only one language is used), and this multi-country, multi-language perspective will result in more widely generalizable outcomes.

All team members, including the PhD trainee, speak both **English and French fluently**, and are deeply interested in **culture** (both have post-graduate experience in Japan and abroad, and published on the topic [3][4][5]). The French team has **psychology** and **neuroscience** expertise, **reverse-correlation** know-how, and resources including **EEG equipment** and access to native French speakers learning English. The Canadian team has **machine learning** expertise for **social signal generation**, backgrounds in French linguistics and French as an L2, past L2 learning projects for Canadian languages [6], and access to English speakers learning French. PhD trainee Paige Tuttosi has a **cross-disciplinary background** in Statistics, Computer Engineering, Anthropology and French Linguistics which make her an ideal candidate. Together they aim for a broadly impactful output with deliverables in (a) **technical innovation** for economic development and social good, as well as (b) **scientific understanding** for use in education and public policy, with publications in high-impact technical and scientific venues.

[1] A. Lim, H.G. Okuno. "The MEI Robot: Towards using Motherese to Develop Multimodal Emotional Intelligence." IEEE Transactions on Autonomous Mental Development, no 6.2, pp. 126-138, 2014

[2] P. Arias, L, Rachman, M. Liuni, M., J.J. Aucouturier. "Beyond Correlation: Acoustic Transformation Methods for the Experimental Study of Emotional Voice and Speech". Emotion Review, 1-13. 2020

[3] E. Hughson, R. Javadi, J. Thompson, A. Lim. "Investigating the Role of Culture on Negative Emotion Expressions in the Wild." Frontiers in Integrative Neuroscience, 2021

[4] A. Lim and A. Matsufuji. "How a Robot Should Speak Depends on Social, Environmental, Cognitive, Emotional, and

Cultural Contexts." Sound in HRI Workshop, ACM/IEEE International Conference on Human-Robot Interaction, 2021
[5] T. Nakai, L. Rachman, P. Arias, K. Okanoya, JJ. Aucouturier (2020) “A language-familiarity effect on the recognition of computer-transformed vocal emotional cues”. bioRxiv, 521641
[6] http://blackfoot-revitalization.cs.sfu.ca/

5. Resources available for project execution: *(ex: field/analytic/bibliographic resources, databases)*

In France:

Equipment: Human electrophysiology laboratory, fully fitted for psychoacoustics and EEG research
Participants: Access to French-native speakers

In Canada:

Equipment : Compute Canada (SFU hosts Cedar, one of the most powerful academic supercomputers in Canada)
Study Participants: Access to English-native speakers

6. Viability of and prospects for collaboration *(ex: research-based training, academic collaboration, publications, communication, organization of symposia, economic/social/industrial optimization)*

Research-based training: The PhD student, Paige, will visit Besançon (FR) and gain experience and training working with reverse correlation and EEGs that she can apply to Canadian participants upon her return.
Academic collaboration: Aucouturier and Lim will jointly explore the generative models and requirements for scientific output and innovation in Task 1 and 2, and provide complementary know-how and resources
Publications: We will publish the research output in both technical and scientific venues for publication including those more specific to speech and audio processing, e.g. ICASSP, Interspeech. Our overall goal will be to produce a broadly impactful, **open-access** cross-disciplinary publication, e.g. Nature Communications or PLOS One.
Public outreach: Both Lim and Aucouturier have extensive public dissemination experience, including TEDx and World Economic Forum online talks (with ~60,000 views) and interviewing with the BBC, New Scientist, Vox, CTV News, MIT Technology review, and others. Lim, who lived in France for 5 years, hosted and narrated "Ma Vie Avec un Robot", a TV documentary for CANAL+ distributed in 3 continents, finalist at the *ParisScience* International Film Festival. Similarly, press releases and online talks discussing outcomes will be planned for maximal worldwide reach. Demonstrations at local science centres using the ROSIE Lab's robots or interactive voice software (e.g. Science World in Vancouver or Cité des Sciences in Paris pending further funding through complementary sources) are also envisioned.
Communication: Regular online group meetings will take place at least once per month on Zoom, and asynchronous communication will be carried out via Slack and email exchanges.
Organization of symposia: The PIs and PhD trainee will jointly propose and organize an interdisciplinary workshop on speech prosody at the Society for Affective Sciences (SAS) annual congress.
Economic/social/industrial: Understanding how to both generate and how humans can adapt their voices to be engaging and increase retention of relayed information have important impacts in education, specifically applications for social good such as teaching of underrepresented languages. For example, outcomes may feed back into a SSHRC-funded project between SFU Indigenous Studies founding director Eldon Yellowhorn and Lim in developing a text-to-speech system for the endangered indigenous language Blackfoot and/or France-Canada commercialization opportunities of the AltaVoce voice-transformation startup founded by the French team (https://alta-voce.tech).

7. Projected project timeline

May 2022 - September 2023

Task 1**:** Developing an artificial voice model for maximum retention in language learning

September 2023 - May 2024

Task 2**:** Understand and explain the characteristics of the engaging teaching voice

September - November 2023

(a) Reverse correlation in France on French-native speakers

December 2023 - February 2024

(b) EEG entrainment study

February 2024 - May 2024

(c) Reverse correlation in Canada on English-native speakers

8. Budgetary requests

Type of eligible expenses	Details of estimated cost (Number of persons involved, number of trips)	Cost	%
Air transport	1 person, 1 x $2300 per round trip = $2300	2300	15
Travel expenses in France	1 person, 1 x Return train from Paris to Besançon = $240 1 person, 6 x $65 monthly transit pass = $390	630	4
Accommodations (France)	1 person, 6 x $1500 monthly = $9,000	9000	60
Other (**not including salaries**). 10% maximum.	1 person, food stipend, per diem $17, $510 x 6 months = $3060	3060	21
Total		**14990**	**100**

Budget justification

All expenses for SFU Ph.D. candidate Paige Tuttosi to undertake research in Besançon with Aucouturier.

Air transport: Return flight Vancouver to Paris, includes relevant expenses such as PCR testing if applicable

Travel expenses: Return train from Paris to Besançon ~$240; Monthly transit pass in Besançon, Passe Sésame 43,50 euros ~$65

Accommodations and Food Stipend: Airbnb studio example, daily meal costs at $17

Appendix – detailed CV for each team member

(Lead researcher for each team and other collaborators—maximum of 4 CVs)