2019 IEEE World Haptics Conference (WHC)
Tokyo, Japan
9-12 July 2019

BarryWhaptics: Towards Countering Social Biases Using Real-Time Haptic Enhancement of Voice

Antoine Weill–Duflos Feras Al Taha Pascal E. Fortin Jeremy R. Cooperstock

***Abstract*— Studies suggest that imbalances in speaking opportunities during meetings often lead to sub-optimal meeting outcomes. These imbalances can be due to a variety of reasons, including people's perception of speakers and their voice. Indeed, speakers with higher pitched voices were shown to be perceived as having lower leadership ability. In an attempt at countering such voice-pitch related biases, this work introduces BarryWhaptics, a real-time speech-to-haptics conversion system that leverages multimodal perception to alter the listener's perception of a speaker. The system operates by augmenting human speech with vibration, applying more intense vibrations to voices that would ordinarily be considered low in dominance. Results from a pilot study assessing the influence of the system in a decision-making task demonstrate that it can meaningfully influence how users choose to follow instructions given by one speaker over another.**

I. Introduction

Our perception of the skills and abilities of others is strongly influenced by non-verbal factors, including proxemics [1], kinesics, vocalics [2], chronemics and haptics [3]. Prior work has shown that we often adopt biased assumptions about social and professional abilities of an interlocutor based on non-verbal vocal properties. For example, a low-pitched voice is likely to be perceived as more dominant than a high-pitched voice when other non-verbal cues remain unchanged. More specifically, studies have shown that fundamental frequency (F_0), loudness, and resonance of voice all affect perceived characteristics of speakers [4], such as physical dominance [5–8], social dominance [4, 7], leadership [9], and attractiveness [5, 6]. Indeed, a louder voice will often be associated with a more physically dominant person [4]. Additionally, a higher pitched voice, while seeming more attractive in a female speaker (up to a F_0 of 280 Hz [6]), will appear less dominant and with less leadership ability in both male and female speakers, independent of the gender of the listener.

These voice-related biases negatively influence the outcomes of social and professional interactions in various respects. People perceived as being more dominant tend to occupy more speaking time, which can reduce the effectiveness of a group meeting and result in a less preferable outcome than a meeting with equal speaking opportunities [10].

Prior work has applied real-time sound processing techniques to alter the speaker's vocal properties in an effort to elicit a different perception among listeners. For example, Aucouturier et al. [11] manipulated perceived vocal emotion by modifying the voice features. Similarly, Rachman et al. [12] created an open-source software platform to manipulate speech and emotional cues by modifying pitch and other voice characteristics. Similar techniques could be applied to transform multiple voices in real-time, equalizing perceived dominance among several speakers. However, a significant drawback of these approaches is that they modify the voice of the speakers, potentially masking nuances in elocution that are important to their vocal identity.

To avoid this situation, we were motivated to investigate the use of haptic augmentation to modify perception without altering the speaker's original speech signal. Consequently, all the non-verbal nuances and identifying features of speech are kept intact. A number of studies have looked into the effect of vibration amplitude, frequency and patterns on participants' pleasure and arousal perception, two of the three most commonly used dimensions of affect. While results vary with the stimulation system, in most cases, an increase in vibration amplitude was associated with higher self-reported arousal, and lower pleasantness ratings [13, 14]. Indeed, an increase in vibration frequency was shown to be correlated with higher reported arousal, although different studies have reached conflicting conclusions regarding the effect on valence [13, 14].

Despite the existing haptic perception literature focusing exclusively on the pleasure and arousal components of the PAD (Pleasure Arousal Dominance) emotion representation space, we hypothesize that vibration amplitude could significantly influence the perceived dominance of a stimulus similar to how a louder voice or brighter light would attract one's attention. If found to be true, these perceptual properties could be used to mitigate the effects of natural biases, as illustrated in Fig. 1.

(a) Without the system (b) With the system

Fig. 1: Explanation of the aim of our system: suppression of voice-related bias in the perception of dominance

The authors are with Shared Reality Lab, Centre for Intelligent Machines, Department of Electrical and Computer Engineering, McGill University, Montreal, Canada
antoine.weill-duflos@mcgill.ca
[feras.altaha,pe.fortin]@mail.mcgill.ca
jer@cim.mcgill.ca

To test our hypothesis, we introduce BarryWhaptics,[1] a system rendering real-time haptic feedback from a voice input. By augmenting higher pitched voices with stronger vibrations, the goal is to balance the perceived dominance across voice pitch in real-time. In theory, such an approach could reduce voice pitch bias, allowing naturally higher-pitched voices to be perceived as equally or more dominant than their lower-pitched counterparts. We provide a detailed description of our design process, the current system implementation, and preliminary results from a pilot test assessing the system's ability to affect speaker perception.

II. The BarryWhaptics architecture

Our system design includes three main components:

1) Voice acquisition
2) Haptic effects synthesis
3) Haptic presentation

For remote usage, additional audio hardware (speaker, headphones, etc.) would also be required. These subsystems are organized in a pipeline, as presented in Fig. 2.

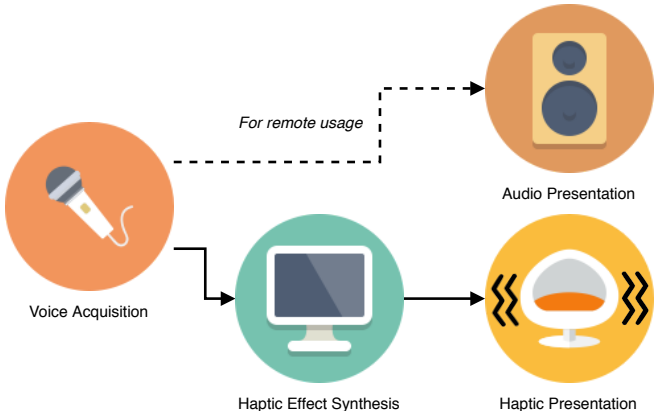

Fig. 2: System description.

A. Voice acquisition

It is essential that the voice acquisition system accurately records voice pitch. A more challenging constraint is the requirement that the voices of multiple speakers in the same room can be separated and processed independently. Finally, the quality of the recording must be sufficient to avoid impacting the performance of the rest of the system. Two solutions are conceivable:

1) An omni-directional noise-canceling microphone sitting in the middle of a meeting table captures speech and sends it to the computing unit as an audio stream. This may benefit from an accompanying speaker diarization algorithm [10] to ensure proper augmentation of distinct voices in overlapping speech scenarios.
2) An individual near-field microphone is assigned to each speaker to ensure that the voices are captured with reasonable isolation, and can be processed as separate audio channels.

For simplicity of hardware in our prototype implementation, we opt for the first solution, and employ a *Zoom H2n Handy Recorder*,[2] chosen for its ability to capture the voices of numerous people in a room. However, informal experimentation suggests that it is possible to replace the Zoom microphone with a wide variety of microphones of varying grades without having a significant impact on system performance, as long as the voices remain clearly audible. The Zoom unit integrates five microphones and a 32 bit processing unit. The signals from the five microphones are combined to offer stereo or surround sound. For this project, we converted the stereo signal to a single input channel. The sampling frequency of the microphone was set to 48 kHz, the same frequency at which the data was processed. A buffer of 1024 samples was used in the real-time audio system.

The next step in the processing pipeline is to use the digitized voices to synthesize the haptic effects.

B. Haptic effect synthesis

The haptic effects are generated on a real-time processing unit. The Python Pyo module[3] is employed for audio signal processing, which makes use of the sound card's DSP hardware, a Realtek ALC892, running on a Linux workstation with an Intel i7 920 CPU and 6 GB of RAM.

The real-time processing unit takes in one or more audio tracks, processes them individually to convert the audio to vibrations, and renders all the vibrations to a single output channel that can be duplicated if there are multiple rendering devices (e.g. for multiple users in a meeting room). The low latency design of our architecture ensures that the haptic effects are well synchronized with the speaker's voice. In the case of remote meetings, the system also outputs the unmodified audio tracks through speakers or headphones as a separate channel.

Two different sound-to-haptic architectures and approaches are explored. Both rely on the idea of associating higher-pitched voices with a stronger haptic effect to counter the correlation of lower pitched voices to higher perceived dominance. In order to minimize the influence of novelty, all voices are augmented with vibrations.

1) Binning method: Initially, the approach taken to convert a voice to vibrations while enhancing higher frequencies was to separate the voice spectrum into several bins using parallel band-pass filters. Only the fundamental frequency components of the voice were captured, in the range 100 Hz to 300 Hz, following the results of Baken and Orlikoff [16]. A more detailed distribution of fundamental frequencies found in male and female voices is shown in Fig. 3. The higher harmonics were suppressed to avoid acoustic, audible effects in the vibrations. The output of each band-pass filter had its amplitude modulated according to the frequency range of the bin.

[1]The name is based on the singer Barry White, known for "his distinctive bass-baritone voice" [15]. Our concept is to recreate the effects of the dominance-enhancing characteristics conveyed by a voice such as Barry White's using haptic effects.

[2]https://www.zoom.co.jp/products/handy-recorder/h2n-handy-recorder

[3]http://ajaxsoundstudio.com/software/pyo/

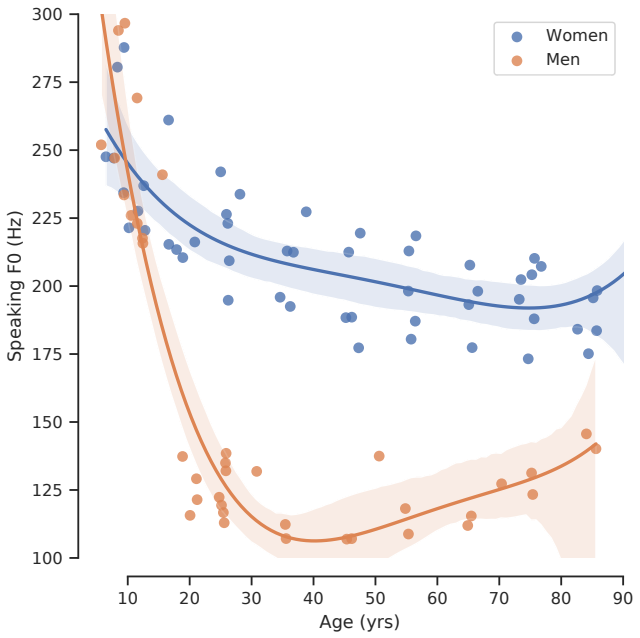

Fig. 3: Distribution of F_0 in male and female by age. Data points were taken from many studies [16].

Linear, logarithmic and exponential amplitude envelopes applied to each frequency bin were considered, with an example of these mappings described in Fig. 4.

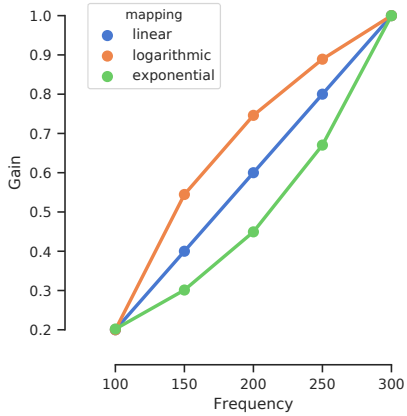

Fig. 4: Amplitude mapping for each band-pass filter

Finally, each filter's output was shifted in the 10 Hz to 300 Hz range to match the frequency range in which vibrations are best perceived [17, 18]. To enhance less dominantly perceived voices in an effort to reduce voice pitch bias, this shift was *inversely proportional* to the bin frequency. This ensured that high-pitch voices resulted in lower frequency vibrations than low-pitch voices. All the bin outputs were then combined into a single output, as summarized in Fig. 5. Fig. 6 illustrates the output spectrum of each filter as well as the total input and output to the system.

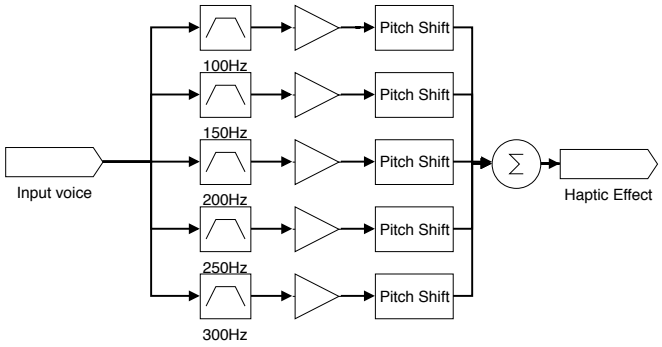

Fig. 5: Block diagram of the binning method

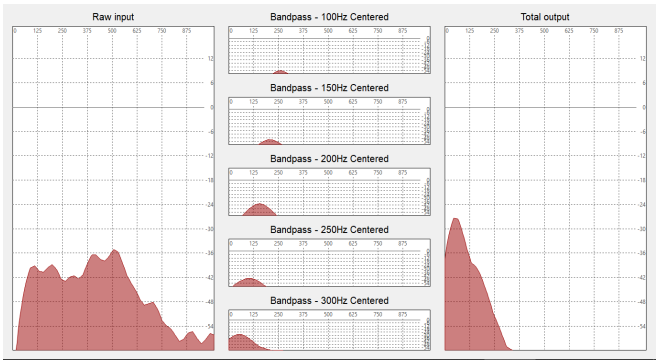

Fig. 6: Input and output spectra with the binning-based method. The displayed spectra of the middle column are the outputs of the band-pass filters *after* frequency shift.

This approach posed a significant challenge because of the numerous parameters that needed to be tuned, i.e., the number of bins, envelopes, shift frequencies, and filter characteristics. Additionally, the entire process was computationally intensive (≥ 5 filters in parallel, all frequency shifted and amplitude modulated in real-time) and ended up creating inconsistent vibrations and time delays. Thus, an alternative vibration synthesis method was considered.

2) Fundamental frequency method: This approach first extracts the fundamental frequency of the voice using the Yin algorithm (de Cheveigne and Kawahara [19]). The signal's amplitude is then amplified or attenuated such that high-pitched voices, i.e., those with a higher fundamental frequency, generate stronger vibrations than low-pitched voices.

The gain is calculated as follows:

$$Gain(F_0) = \frac{F_0 - F_{min}}{F_{max} - F_{min}}$$

where $F_{min} = 100\,Hz$ and $F_{max} = 300\,Hz$.

Next, the signal's frequency spectrum is shifted by an amount dependent on the fundamental frequency:

$$F_{shift}(F_0) = F_0 - 100$$

This ensures that the resulting signal gives rise to perceptible vibrations with a peak at a frequency of $100\,Hz$.

Finally, a low-pass filter with cutoff frequency of $200\,Hz$ is applied to the signal to remove high frequency harmonics and prevent the vibrations from being obviously

audible or distracting. The resulting algorithm is shown in Fig. 7. Adding the low-pass filter at the end, rather than

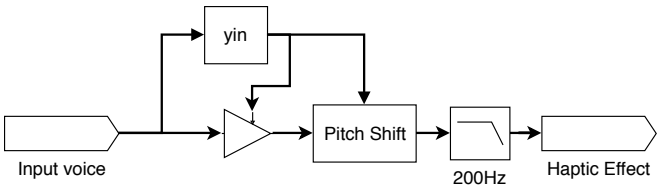

Fig. 7: Fundamental frequency based processing algorithm.

the beginning, preserves all the voice components in the generation of vibrations, thus ensuring a richer output. Since this method reduces the amount of simultaneous filtering and processing, it is significantly less computationally intensive than the former binning approach. It also results in improved quality of vibration output and synchronization with voice, and does not suffer from the issue of overlapping bins that unintentionally enhance overlapped frequencies.

C. Haptic Delivery

Two different haptic delivery systems are considered to transmit vibrations to the user, either to the wrist, or the entire body, via a chair, as described in Table I.

TABLE I: Haptic rendering device specifications.

	Wristband	Chair
Manufacturer	Tactile Labs	Clark Synthesis
Model	Haptuator TL002-14-A	TST239 Silver
Bandwith	50 *Hz* to 500 *Hz*	15 *Hz* to 17 000 *Hz*
Power handling	1.5 W	100 W
Impedance	6 Ω	4 Ω

The first is a portable version that uses a wristband equipped with a vibrotactile transducer, the Tactile Labs Haptuator TL002-14-A.[4] This option renders vibrations at the wrist level, with a resulting sensation that is subtle and not overly intrusive.

The second option is an office chair augmented with two Clark Synthesis TST239 Tactile Transducers,[5] one on the backrest and a second under the seat, as illustrated in Fig. 8. This option requires the signal to be amplified by a 75 W to 100 W amplifier to drive the actuators.

Although the wearable system is considerably smaller, lighter, and less intrusive than the office chair solution, it is significantly less powerful. We hypothesize that by vibrating the back and seat of the office chair, vibrations delivered in this manner will be perceived as more intimate, and also produce richer effects: the torso, although less sensitive to vibration than the wrist [20], is nevertheless more sensitive to amplitude changes [21]. Therefore, we opted to use the second option for the studies described below.

[4] http://tactilelabs.com/wp-content/uploads/2012/07/TL002-14-A_v1.2.pdf

[5] http://clarksynthesis.com/wp/wp-content/uploads/2015/03/TST-239-Brochure.pdf

III. Experiment

A study was conducted in order to evaluate the influence of the system on voice perception. The study involved a decision-making task in which two neutral instructions are spoken by voices of different pitch. Fig. 8 illustrates the experimental configuration.

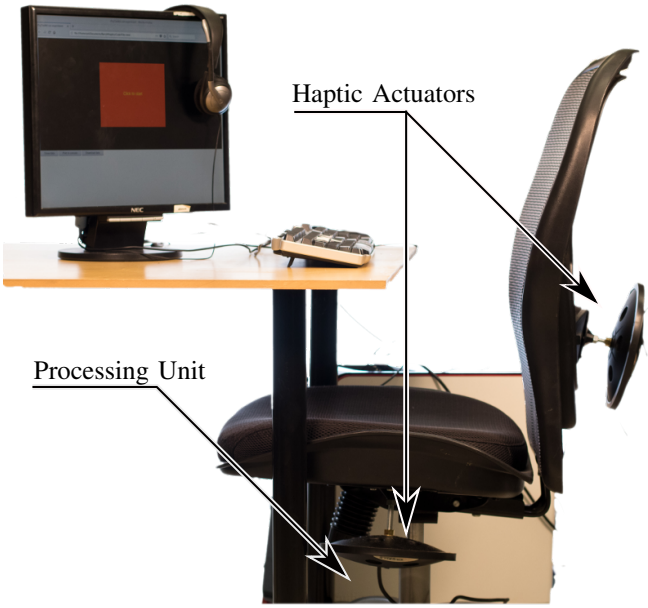

Fig. 8: Picture of the real setup

Seven participants (F:3, M:4) with no involvement or prior knowledge about the project's objectives and methods were recruited from our research group. They were asked to follow one of the two spoken instructions. However, the experimenters did not provide any criteria or instructions as to which voice should be selected.

The main questions investigated by this experiment were the ability of the system 1) to modify the subconscious choice, and 2) to control the direction of a change in choice.

A. Voice acquisition in pilot experiment

For the pilot experiment, prerecorded speech files were used instead of a real-time voice acquisition system to ensure consistency between the trials. However, the haptic effects were generated in real-time from the prerecorded audio, demonstrating the capabilities of the processing algorithm. Speech recordings were generated using Google Cloud's Text-to-Speech service[6] with the "en-US-Wavenet-C" voice, which imitates a female voice. The pitch of the generated voices can be up-shifted or down-shifted by increments of semitones (adding one semitone is equivalent to multiplying the frequency by $\sqrt[12]{2}$) to produce both a high-pitch ($+5$ *semitones*) and low-pitch version (-5 *semitones*). Subjectively, speech samples were of high quality, hardly distinguishable from recordings of human voices.

[6] https://cloud.google.com/text-to-speech/

B. Procedure

The pairs of text-to-speech generated sentences, instructing participants to chose one of two options, were presented with and without the proposed haptic reinforcement in two experiment blocks. Instructions included a choice between (1) square and triangle, (2) the numeric values 343 and 434, (3) left and right, (4) up and down, and (5) two invented words: “oxilmani” and “busallue”. These were selected to be as unbiased as possible:

- shapes that are similar; the circle was avoided because of potential special interpretations, for example, driven by cultural references [22]
- the selected numbers avoid possible birthday dates or common “lucky numbers”

Paper labels representing the choices were placed on adjacent keycaps of a regular keyboard. For each choice, participants were asked to select one of the two options by pressing the key corresponding to the chosen instruction.

Each of the five pairs of sentences was presented in counter-balanced order, with both high- and low-pitch voice versions of each instruction, four times, with and without haptic augmentation, for a total of 160 trials.

C. Results

Without haptics, instructions provided in the higher-pitch voice were followed, on average, $51\,\%$ of trials.

For the purpose of analysis, we divide the results into those for whom the haptic augmentation had a positive impact (i.e. those who were influenced in choosing higher pitched voices), vs. those for whom it had the opposite effect. In this respect, we note that the two participants who fit the latter category also expressed, at the end of the experiment, that they avoided the strongest vibration as it felt uncomfortable. Fig. 9 clearly demonstrate the effect of this vibration aversion behavior.

For the remaining participants, the results indicate a strong positive effect of the haptic voice augmentation. As shown in Fig. 9, the fraction of trials for which participants chose to respond to the instruction with the higher frequency voice increased from a baseline of $54.7\,\%$ to $63.4\,\%$ with haptic augmentation. A Friedman’s test, chosen because of the ordinal nature of the value measured and the repetition of the measures, indicated that this increase was statistically significant ($p < 0.05$, $\chi = 5$). It is confirmed by a Conover posthoc test with $p < 0.01$.

Based on these results, we can confirm that the haptic augmentation does succeed in modifying the subconscious choice, but not in a consistent direction across participants. We discuss this further in the following section.

IV. Discussion and Future Work

The varied reactions of participants exposed to the vibrations constitutes a significant limitation of the system. A larger study is needed to determine more precisely the reactions to haptic augmentation, and whether it is possible to relate the different reactions to social criteria.

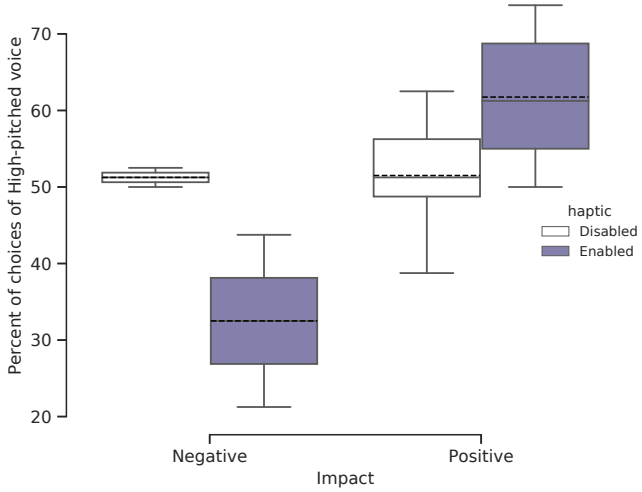

Fig. 9: Overall results, sorted by positive and negative impact

The discrepancy of results with the documented perceived dominance of lower-pitched voices over higher one was unexpected. This may be the result of small imperfections or unnatural artifacts appearing in the voices when the pitch is strongly shifted. The second hypothesis is a side effect of not imposing a specific criterion for the choice of one voice over the other. Since female voices can be perceived as more attractive when higher pitched [6], it is entirely possible that this factor outweighed perceived dominance in the participant’s choice of which instruction to follow. Unfortunately, explicitly determining how participants should choose the instruction to follow would draw explicit attention to dominance factors, which was preferable to avoid in order to assess how participants subconsciously perceived speakers.

A further challenge is that if two people with different vocal ranges speak concurrently, the synthesis of haptic effects will not produce the desired output. One solution considered is the use of another identification algorithm or the use of several audio tracks each with their own gain, ensuring that each voice is processed separately.

In addition, the system does not attempt to adjust vibration intensity as a function of the speaker’s loudness. On the contrary, a louder speaker would give rise to stronger vibrations, regardless of differences in voice pitch. This is obviously at odds with our design objectives. A possible solution would be to normalize voice loudness during voice acquisition, but this raises numerous questions of microphone placement and time window over which normalization should be applied.

Other desirable future enhancements include the addition of more actuators to the system, which could vibrate differently depending on the location (side of the room) of the speaker, thus directing listeners’ attention appropriately. The system could also selectively enhance voices based on other factors, for example, speaking time, such that less talkative individuals are “promoted” during the discussion.

V. Conclusion

We presented a possible solution to augment voice with haptic effects as a prototype effort to address voice-pitch-related biases in perception of dominance. The system relies on identification of the fundamental frequency of a speaker voice. As the literature suggests that higher-pitched voices are perceived as less dominant, the system associates high frequencies with stronger vibrations.

Results of our pilot study indicate the promising effect of the system on decision-making of listeners, and more generally, on their perception of speakers and of voice-related biases. However, as noted above, our experiment did not control for the *decision-making criterion* by which participants favored one instruction over another, and thus, we cannot determine what factor led to this result. Moreover, the variability of reactions to the system is worth investigating as a topic for future study. If the system's effect can be properly controlled, such a system could be used in workplaces to equalize people's perceived dominance. This would result in more equal contribution opportunities by neutralizing voice-related biases, potentially improving the outcomes of workplace interactions and meetings.

References

1. Hall, E. T. The Hidden Dimension (1966).
2. Burgoon, J. K., Dunbar, N. E. & Segrin, C. en. in *The Persuasion Handbook: Developments in Theory and Practice* 446–474 (SAGE Publications, Inc., 2002). ISBN: 978-0-7619-2006-9. doi:10.4135/9781412976046.n23.
3. Knapp, M. L. & Daly, J. A. *Handbook of Interpersonal Communication* en. ISBN: 978-0-7619-2160-8 (SAGE, Oct. 2002).
4. Ko, S. J., Sadler, M. S. & Galinsky, A. D. The Sound of Power: Conveying and Detecting Hierarchical Rank Through Voice. *Psychological Science* **26,** 3–14. ISSN: 0956-7976 (Jan. 2015).
5. Jones, B. C., Feinberg, D. R., DeBruine, L. M., Little, A. C. & Vukovic, J. A Domain-Specific Opposite-Sex Bias in Human Preferences for Manipulated Voice Pitch. *Animal Behaviour* **79,** 57–62. ISSN: 0003-3472 (Jan. 2010).
6. Borkowska, B. & Pawlowski, B. Female Voice Frequency in the Context of Dominance and Attractiveness Perception. *Animal Behaviour* **82,** 55–59. ISSN: 0003-3472 (July 2011).
7. Puts, D. A., Hodges, C. R., Cárdenas, R. A. & Gaulin, S. J. C. Men's Voices as Dominance Signals: Vocal Fundamental and Formant Frequencies Influence Dominance Attributions among Men. English. *Evolution and Human Behavior* **28,** 340–344. ISSN: 1090-5138, 1879-0607 (Sept. 2007).
8. Tusing, K. J. & Dillard, J. P. The Sunds of Dominance Vocal Precursors of Perceived Dominance During Interpersonal Influence. en. *Human Communication Research* **26,** 148–172. ISSN: 0360-3989 (Jan. 2000).
9. Klofstad, C. A., Anderson, R. C. & Peters, S. Sounds like a Winner: Voice Pitch Influences Perception of Leadership Capacity in Both Men and Women. *Proceedings of the Royal Society of London B: Biological Sciences.* ISSN: 0962-8452, 1471-2954. doi:10.1098/rspb.2012.0311 (Mar. 2012).
10. Hung, H., Huang, Y., Friedland, G. & Gatica-Perez, D. Estimating Dominance in Multi-Party Meetings Using Speaker Diarization. *IEEE Transactions on Audio, Speech, and Language Processing* **19,** 847–860. ISSN: 1558-7916 (2011).
11. Aucouturier, J.-J. *et al.* Covert Digital Manipulation of Vocal Emotion Alter Speakers' Emotional States in a Congruent Direction. *Proceedings of the National Academy of Sciences of the United States of America* **113,** 948–953. ISSN: 0027-8424 (Jan. 2016).
12. Rachman, L. *et al.* DAVID: An Open-Source Platform for Real-Time Transformation of Infra-Segmental Emotional Cues in Running Speech. *Behavior Research Methods* **50,** 323–343. ISSN: 1554-3528 (Feb. 2018).
13. Yoo, Y., Yoo, T., Kong, J. & Choi, S. *Emotional Responses of Tactile Icons: Effects of Amplitude, Frequency, Duration, and Envelope* in *World Haptics Conference* (IEEE, 2015), 235–240.
14. Wilson, G. & Brewster, S. A. *Multi-Moji: Combining Thermal, Vibrotactile & Visual Stimuli to Expand the Affective Range of Feedback* in *Proc. Human Factors in Computing Systems* (ACM, New York, NY, USA, 2017), 1743–1755. ISBN: 978-1-4503-4655-9. doi:10.1145/3025453.3025614.
15. Barry White. en. *Wikipedia.* Page Version ID: 858991676 (Sept. 2018).
16. Baken, R. J. & Orlikoff, R. F. in *Clinical Measurement of Speech and Voice* 145–223 (Cengage Learning, 2000). ISBN: 1-56593-869-0.
17. Hayward, V. & Astley, O. R. in *Robotics Research* 195–206 (Springer, 1996).
18. Reynier, F. & Hayward, V. Summary of the Kinesthetic and Tactile Function of the Human Upper Extremities. *McGill Research Center for Intelligent Machines, McGill University, Montreal, Canada* (1993).
19. De Cheveigné, A. & Kawahara, H. YIN, a Fundamental Frequency Estimator for Speech and Music. *The Journal of the Acoustical Society of America* **111,** 1917–1930 (2002).
20. Karuei, I. *et al. Detecting Vibrations across the Body in Mobile Contexts* in *Proc. Human Factors in Computing Systems* (ACM, 2011), 3267–3276.
21. Jones, L. A. & Sarter, N. B. Tactile Displays: Guidance for Their Design and Application. *Human factors* **50,** 90–111 (2008).
22. Abbott, E. A. *Flatland* fr. ISBN: 978-963-522-672-6 (Booklassic, June 2015).