Stronger Skin: Voice Modulation Through Neurological Feedback

Katarina Richter-Lunn
Graduate School of Design
Harvard University
Cambridge, Massachusetts, USA
krichterlunn@gsd.harvard.edu

ABSTRACT

While currently living in the age of technological overload, and endless hours of video calls, we find ourselves suddenly hyper exposed to our facial expressions, reactions, and unique tendencies. This platform, along with numerous social outlets, has enabled us to gain a stronger understanding of the sentic modulation we are exhibiting to the world. However, our voice remains rarely analyze by our own auditory perception. There has been a significant amount of research in the field of speech emotion recognition, but little application to how technology can account for these vocal cues and alter them to match our desired auditory output. This project proposed a new system of personalized voice modulation which alters prosody in speech by interpreting individuals' neurological feedback while they speak.

Author Keywords
Speech modulation; voice; prosody; timber; anxiety; auditory perception; affective computing; EEG.

CSS Concepts
• **Human-centered computing~Human computer interaction (HCI)**; *Interaction techniques*; Auditory feedback.

INTRODUCTION

Voice might not be the sentic modulation from which we receive feedback on the most often and yet it can be one of the most telling of emotional behavior. Especially when other factors such as expressions, gestures, and posture are not available. There are numerous structural elements of speech such as pitch, rate, volume, and frequency which gives each individual their unique voice, but also informs the emotional undertone of what is being communicated. Among the non-verbal signals translated through speech two of the most relevant are that of prosody and timber. Prosody refers to the pitch, pace, and volume of the sound while timber seeks to identify the resonance by which the ear recognizes sound [1, 2]. Combined - these elements of speech provide the foundation for interpreting auditory perception.

Recent studies have begun to measure the impact of voice modulation during interpersonal conflicts through the lens of self-perception [3]. Additionally to confirming the significant impact emotional regulation has when relying on voice in communication this study highlights the benefits of en vivo voice modulation software in providing feedback to the individual. This near to live translational occurrence allows for participants to be more aware of how they might be vocally interpreted and adjust their tone. This research led me to think of how we might design a system to react proactively to these emotional cues and make slight alterations based on the biofeedback received at the time of communication. Relevant feedback could include physiological signals such as Heart Rate Variability, or Electrodermal Activity, while cognitive signals could extend to electrical activity in the brain (EEG waves). Optimally you are looking to capture feedback with high correlation to changes in vocal prosody, such as elevated heart rate and increased pace of speech. Although it is important to note that all physiological and cognitive signals play a role in altering your emotional response [4].

BACKGROUND AND MOTIVATION

DeepLearning

This project was greatly inspired by the work of Modulate AI and their development of VoiceSkins [5]. This technology uses machine learning to analyze the speech of a person and produce new speech with the exact emotion, inflection, and cadence of the individual. The sophistication of this software and low latency allows for undetectable alterations of the output voice to someone who does not know what the original voice sounds like. Designed primarily for use in video games where someone might not want to use their real voice this software opens a whole new world of voice perception and automated modulation for both preference and security purposes.

Being inspired by the convincing use of neural network-based systems for voice alteration I embarked on a journey into different deepfake models which focus on audio. This led me to two main open-source models:

- SV2TTS - Is three stage deep learning framework that allows the creation of numerical representations of voice and uses it to condition a text-to-speech model trained to generalize new voices [6].

- Mellotron – Is a multispeaker voice synthesis model that can make a voice emote and sing without emotive of singing training data [7].

Each of these models provided some fascinating results in both mimicking of an original voice clip to the transformation of rhythm and pitch to make a voice emote and sing. However, there remains two foundational issues with these models that I was unable to circumvent for the purpose of my project. Those are that of the latency time of the transformation as well as the convincing quality of the output voice modulation. Due to these restrictions in the models I choose to procced with an alternative computational approach which I will review in the methods section of this paper.

Auditory Perception

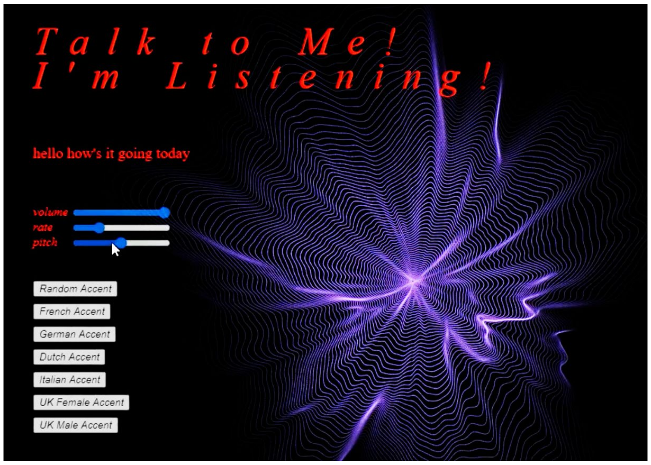
Having already mentioned the current research highlighting the importance of speech prosody in changing voice self-perception I would like to outline the relevance of auditory emotional cues in receptive perception. Voice intonation is among the most apparent sentic modulations when communicating verbally. Among its numerous variances, pitch is one which has been extensively studied [8]. For example, research has shown that low-pitch voices are associated with physical and social dominance [9] while high pitch-voices are collated with weakness or powerlessness [10]. While rate, which represents the pace and volume of speech, tends to be associated with confidence and legitimacy. For instance, decreased speech rate shows to have a deleterious effect on a speakers perceived persuasiveness, fluency, and assertiveness [11]. This all to say that pitch, rate, and volume of speech, are crucial factors in influencing the way in which other perceive and interpret what we are trying to communicate beyond the vocabulary used.

METHODS

Through the process of researching the structural elements of sound I decided to explore these individual signals firsthand through the use Java based code and some additional sound libraries.

Voice Demos

My first voice Demo was written in a java script web based platform called *P5.js* and includes the library *p5.speech* (Figure 1) [12]. This library uses a voice to text converter to record your input sentence and synthesis a response using your choice of accent, rate, pitch, and volume. This first experiment was simply to explore the alteration of each individual characteristic of voice while also testing the fidelity of the voice to text recognition library.



Talk to Me!
I'm Listening!

hello how's it going today

Figure 1. Demo 01. User Interface for recording, altering, and playing the new synthesized voice.

Similarly, to the first demo I continued this exploration by adding in the use of a library called *Minim* which is take audio input and allows you to playback that exact input with alterations to various qualities [13]. This eliminates the necessity for voice-to-text translation as well as the need to synthesize a new voice.

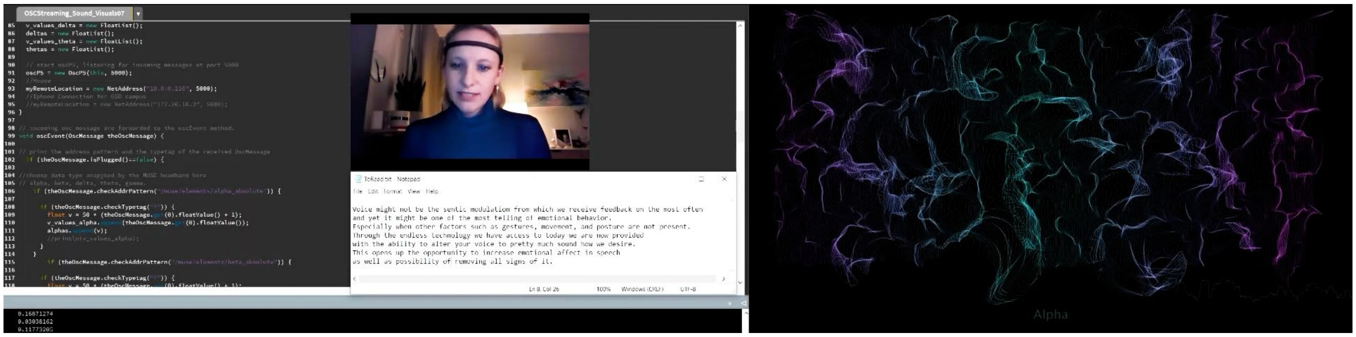
Both initial studies taught me a great deal on the impact of simply adjusting one quality of the sound rather than multiple at the same time and provided me with the tools to proceed with the next step: the introduction of *neurological feedback as the moderator for the output voice alterations.

System Structure

In this project I used the *Muse Headband* to communicate the EEG frequencies of an individual to the voice modulation software. Communication between these devices is hosted via *Processing* and occurs through the OSC streaming library *oscP5*, and the app *Mind Monitor* [14, 15, 16, 17]. All audio influence is made possible by the *Processing* sound library and *Minim* library. These applications and libraries allow for the data to remain continuous and current so that correlation between EEG waves and voice are synchronous.

Specifically, this means that as the microphone is recording the input voice, the rate, pitch and volume are instantaneously being adjusted according to the frequency level of your EEG waves. For example, if a person is resonating at a high frequency wavelength, namely Beta or Gamma waves, which tend to signal heightened cognitive processing and excitement, the pace and pitch of the output speech would be moderated to decrease and slowdown [18, 19]. This action would then reverse when wave frequencies are low, such as Delta or Theta waves, and remain constant to the original input when in Alpha waves (Figure 2).

*Due to limited accessibility to resources at this time I choose to use Electroencephalogram waves as the moderating feedback for voice modulation. Despite there being a correlation between EEG patterns and sentic modulation I believe that this project would be best supported through the use of physiological feedback such as HRV or EDA.



```
  v_values_delta = new FloatList();
  deltas = new FloatList();
  v_values_theta = new FloatList();
  thetas = new FloatList();

  // start oscP5, listening for incoming messages at port 5000
  oscP5 = new OscP5(this, 5000);
  //House
  myRemoteLocation = new NetAddress("10.0.0.216", 5000);
  //Iphone Connection for GSD campus
  //myRemoteLocation = new NetAddress("172.20.10.3", 5000);
}

// incoming osc message are forwarded to the oscEvent method.
void oscEvent(OscMessage theOscMessage) {

// print the address pattern and the typetag of the received OscMessage
  if (theOscMessage.isPlugged()==false) {

//choose data type anapysed by the MUSE headband here
// alpha, beta, delta, theta, gamma.
    if (theOscMessage.checkAddrPattern("/muse/elements/alpha_absolute")) {

      if (theOscMessage.checkTypetag("f")) {
        float v = 50 * (theOscMessage.get(0).floatValue() + 1);
        v_values_alpha.append(theOscMessage.get(0).floatValue());
        alphas.append(v);
        //println(v_values_alpha);
      }
    }
      if (theOscMessage.checkAddrPattern("/muse/elements/beta_absolute")) {

      if (theOscMessage.checkTypetag("f")) {
        float v = 50 * (theOscMessage.get(0).floatValue() + 1);
```

```
0.16871274
0.03038162
0.11773205
```

```
Voice might not be the sentic modulation from which we receive feedback on the most often
and yet it might be one of the most telling of emotional behavior.
Especially when other factors such as gestures, movement, and posture are not present.
Through the endless technology we have access to today we are now provided
with the ability to alter your voice to pretty much sound how we desire.
This opens up the opportunity to increase emotional affect in speech
as well as possibility of removing all signs of it.
```

Figure 2. Project demonstration using *Muse Headband* and *Processing* code to generate output speech modulation.

RESULTS AND DISCUSSION

The results of the project show great promise in both the use of the software for fun experimentations of one's own voice manipulation as well as the potential for very subtle alterations in prosody to alleviate the often-unintentional fluctuations in speech. I personally found it quite enlightening to work with the software to recognize first the change in tone that occurs so quickly by simply adjustments to certain speech characteristics, as well as the correlations you can start to recognize between your neurological signals and output speech. For example, as I was working on the code for this project, I would test it over the course of the day, and realized that in the morning, when my cognitive processing and attention are high, the software would output slower, and quieter renditions of my voice in contrast to late at night which would produce faster pace and higher pitch output. This might not necessarily always be what we are looking for but nonetheless was an interesting observation of my performance throughout the day. The beauty of this software is its malleability to being tweaked to produce whatever output you might desire at the time. Perhaps there are times you might want these alterations to be more subtle and others when you would like to exaggerate them. Additionally, as I will discuss in my conclusion, I see there being great potential for this to be associated with other types of processing signals.

CONCLUSION

In this project I investigated the potential for one to correlate biofeedback with speech characteristics to modulate one's voice output. By leveraging research behind the sentic modulation perceived through vocal signals I created an automated system which adjusts parameters of prosody to compliment one's neurological frequencies. With results ranging from playful ways in which one can become more self-aware of their voice qualities to other being more subtle and useful for presentation or security purposes. I see great potential for this project to grow in the sophistication of the modulation software, such as the use of VoiceSkins by Modulate AI, but also am interested in the exploration of other sensor feedback which might be more in tune with your perceived vocal fluctuations. I imagine this initial investigation to be just the onset of personalized voice modulation and see that in the near future we might be able to design and activate different desired voice alterations based on personal sensor feedback or simply preference. However, just as face swapping and deepfakes has demonstrated, with good intentions also comes misuse and abuse of such technology. This is not to say that the pursuit of these system should not be explored but simply that the ethical considerations should always be at the forefront of each innovation.

ACKNOWLEDGMENTS

I would like to thank the incredible teaching staff of MAS.S60 – *Experiments in Deepfakes: Creativity, Computation and Criticism* - Roy Shilkrot, Pat Pataranutaporn, Joanne Leong, and Prof. Pattie Maes. Thank you!

REFERENCES

[1] Anne Cutler, Delphine Dahan, and Wilma Van Donselaar. 1997. Prosody in the comprehension of spoken language: A literature review. *Language and speech* 40, 2 (1997), 141 201.

[2] Mariëlle Stel, Eric van Dijk, Pamela K Smith, Wilco W van Dijk, and Farah M Djalal. 2012. Lowering the pitch of your voice makes you feel more powerful and think more abstractly. *Social Psychological and Personality Science* 3, 4 (2012), 497–502.

[3] Jean Costa, Malte F. Jung, Mary Czerwinski, François Guimbretière, Trinh Le, and Tanzeem Choudhury. 2018. "Regulating feelings during interpersonal conflicts by changing voice self-perception." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1-13.

[4] Rosalind W. Picard. 2000. *Affective computing*. MIT press.

[5] "VoiceWear". 2020. Modulate.ai. https://www.modulate.ai/voice-wear

[6] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. 2018. "Transfer learning from speaker verification to multispeaker text-to-speech synthesis." *Advances in neural information processing systems* 31: 4480-4490.

[7] Rafael Valle, Jason Li, Ryan Prenger, and Bryan Catanzaro. 2020. "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens." In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6189-6193. IEEE.

[8] Kyle James Tusing and James Price Dillard. 2000. The sounds of dominance. *Human Communication Research* 26, 1, 148–171.

[9] David Andrew Puts, Steven JC Gaulin, and Katherine Verdolini. 2006. Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior* 27, 4, 283–296.

[10] Renée Van Bezooijen. 1995. Sociocultural aspects of pitch differences between Japanese and Dutch women. *Language and speech* 38, 3, 253–265.

[11] William Apple, Lynn A. Streeter, and Robert M. Krauss. 1979. "Effects of pitch and speech rate on personal attributions." *Journal of personality and social psychology* 37, no. 5: 715.

[12] Lauren Lee McCarthy. *P5.js*. 2020. https://p5js.org/

[13] Damien Quartz. *Minim*. Processing 2.0. 2020. http://code.compartmental.net/tools/minim/

[14] InteraXon Inc. *Muse*. 2020. https://choosemuse.com/.

[15] Ben Fry, and Casey Reas. *Processing*. 2001. https://processing.org/.

[16] Andreas Schlegel. *oscP5*. 2011. https://www.cnmat.berkeley.edu/OpenSoundControl/.

[17] James Clutterbuck. *Mind Monitor*. 2020. https://mind-monitor.com/.

[18] Michael Thompson, and Lydia Thompson. 2015. *The Neurofeedback-Book: an Introduction to Basic Concepts in Applied Psychophysiology* . Toronto: Association for Applied Psychophysiology and Biofeedback.

[19] Michal Teplan. 2002. "Fundamentals of EEG measurement." *Measurement science review* 2, no. 2: 1-11.