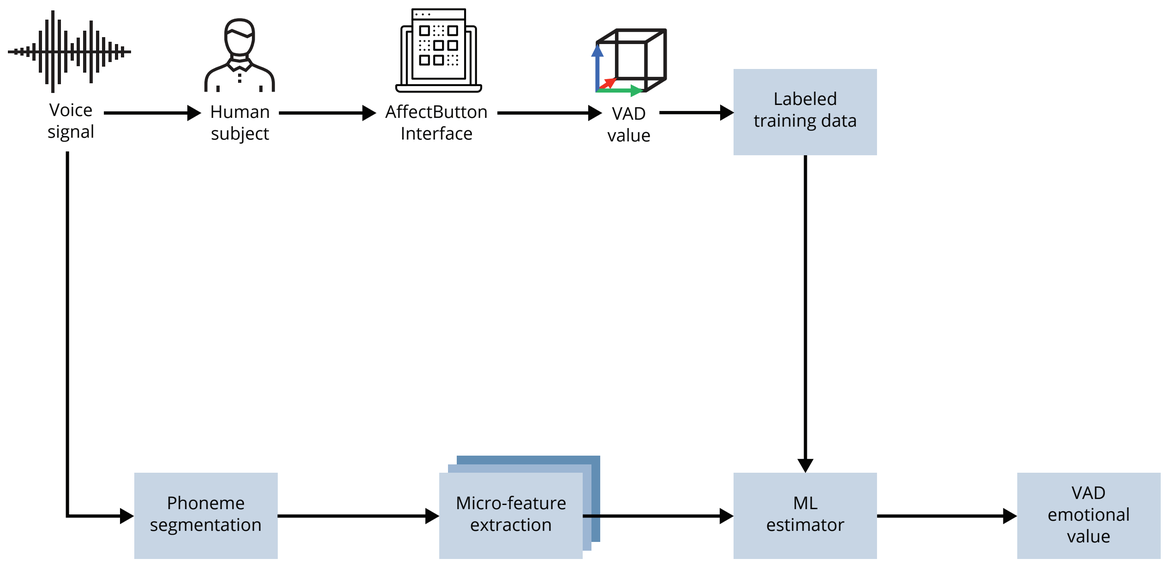
A Novel Approach to Emotion Recognition from Voice

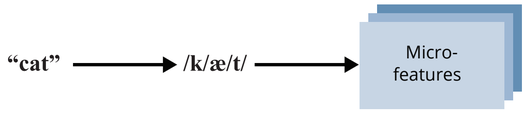
Accurately recognizing emotion from voice is important in defense applications such as speaker profiling and human-machine teaming, but is currently infeasible. We propose a mission-practical prototype using a new, continuous emotional speech database, and a set of micro-articulometry techniques that can capture finer nuances than the current state of the art.



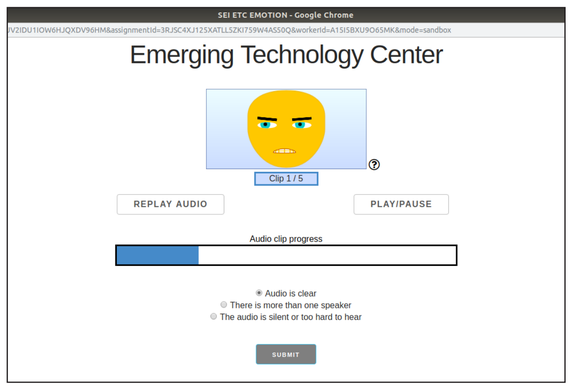
System diagram. Voice data is labeled by crowdsource participants using the AffectButton interface. This labeled data is then used to train classifiers and predict emotion.

The production of the human voice is a complex physical and cognitive process, containing traces of many bio-parameters, including emotional state.

Prior work in speech emotion recognition typically operates at the utterance level—the level of the spoken word or statement. Such approaches are brittle, requiring long, high-quality audio segments. Instead, our approach operates at the phoneme level—the level of the constituent units of speech—using micro-articulometry techniques pioneered by Dr. Rita Singh at CMU's Language Technologies Institute. We will use voice features such as formant position, voicing-onset time, onset of pitch, and phonetic loci as inputs to deep learning classifiers to predict emotional state.



Micro-articulometry: The measurement and modeling of articulatory properties at the phoneme level.



A screenshot of the user interface for data labeling

The AffectButton: An intertactive self-report method for the measurement of human affect. The facial icon changes based on mouse movement [1]

[1] Broekens, J., & Brinkman, W.-P. (2013). AffectButton: A method for reliable and valid affective self-report. International Journal of Human-Computer Studies, 71(6), 641-667.

[2] Mehrabian, A., & Russell, J. A. (1974). An approach to environmental psychology. Cambridge, MA, US: The MIT Press.

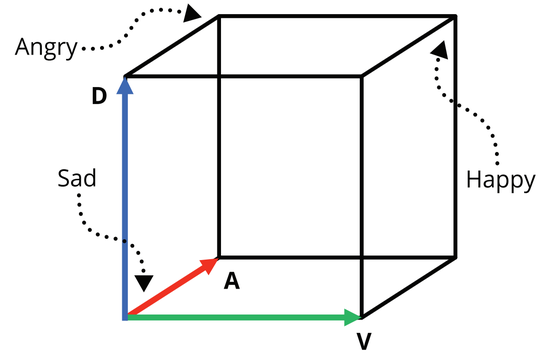
Example Voice Features:

- Diplophonicity
- Flutter
- Formant bandwidth
- Formant dispersion
- Formant position
- Formant Q
- Vocal fry
- Glottalization
- Nasality
- Raspiness
- Resonance
- Shimmer
- Tremor
- Voicing-onset time
- Wobble

Micro-articulatory voice features will be used to train classifiers and predict emotion. Each voice feature requires its own set of signal processing algorithms to extract and measure.

Emotional speech databases today are hand-labeled with discrete categories.

This limitation hinders the mission applicability of any emotion estimator: emotions have varying intensities and overlap with one another, forming a continuum. We are building a new emotional speech database using crowdsourcing techniques to label tens of thousands of speech clips, rather than hundreds, providing the necessary data for deep learning to be effective. Crowdsource participants will label clips using an interface called the AffectButton, based on the VAD emotional state model from psychology, to pick from a continuum of emotions without being biased by explicit, pre-determined labels.



The VAD model: Valence, arousal, and dominance characterize affect in three dimensions [2]

The expected outcome of this project is the creation of the largest ever emotional speech database – which will be open-source and use continuous emotional labels – and an end-to-end emotion recognition prototype built with micro-articulometry algorithms.

Oren Wright | owright@sei.cmu.edu