A Surprising Density of Illusionable Natural Speech

Melody Y. Guan
Stanford University, Stanford, California, United States

Gregory Valiant
Stanford University, Stanford, California, United States

Abstract

Recent work on adversarial examples demonstrates a brittleness of many state-of-the-art machine learning systems. We investigate one human analog, asking: What fraction of natural speech can be turned into illusions which alter humans perception or result in different people having significantly different perceptions? Using generated videos, we first empirically estimate that 17% of words occurring in natural speech have some susceptibility to the McGurk effect–the phenomenon by which adding a carefully chosen video clip to the audio channel affects the viewers perception of the message. We develop a bag-of-phonemes prediction model for word-level illusionability that we extend with natural language modeling to build a sentence-level framework. We train an instantiation using Amazon Mechanical Turk evaluations on sentence-level illusions. Finally we generate several new instances of the Yanny/Laurel illusion, demonstrating that it is not an isolated occurrence. The surprising density of illusionable instances warrants further investigation from cognitive and security perspectives.