
Supervised Learning of Human Speech Affect Using Viola-Jones Features

Paul Ruvolo, Ian Fasel, Javier R. Movellan
Machine Perception Laboratory
University of California San Diego

We present a system for automatically categorizing characteristics of human speech affect. Speech is categorized along a series of binary axes regarding emotional content (e.g. unpleasant vs. pleasant, agitated vs. calm). These judgments are made using purely spectral features. Our system learns a rule to discriminate between each axis based on human-labeled training examples.

While most people can identify whether or not speech has a leadership quality, it is not easy for a person to write down a rule by which to make this judgment. We show that using a cascade of Viola-Jones features trained from human-labeled data that we are able to learn a rule to discriminate between speech affect categories.

Our system demonstrates that Viola-Jones features are applicable to the domain of audio. Since Viola-Jones features encode low-level properties of an image patch, it is necessary to project our audio data into pixel space. This projection is done by calculating an audio feature called the Sone from our samples.

Our work shows that Viola-Jones features can be applied to audio in a way that extracts its abstract characteristics (i.e. emotional content). We will present the results of our work applied to a dataset consisting of a monologue performed by actors. Further, we show that the performance of our system on this dataset is comparable to that of the average human labeler.