Erdös Institute Spring 23 – Speech Recognition

EXECUTIVE SUMMARY BY JACOB MASHBURN, BENJAMIN WARREN, & SURAJ SINGH

Overview

We use an artificial neural network to recognize when a potentially noisy recording contains human speech, for voice-activated consumer electronics, machine-learning-assisted hearing aids, and more.

The Problem

In recent years, speech recognition technology has been integrated into many consumer electronics, such as cell phones, vehicles, and smart homes, and will continue to be in demand for the foreseeable future. Speech-in-Noise (SIN) perception is the ability to recognize relevant voice within a noisy background and is crucial for devices operating in environments where such a background is unavoidable, such as the inside of a moving vehicle. The first step in speech recognition for such devices, and the objective of our neural network model, is to determine whether a given raw audio sample includes someone speaking or not.

Proposed Solution

Our solution is to train an artificial neural network (ANN) on Google's AudioSet, which consists of thousands of 10-second audio excerpts from randomly selected YouTube videos. The raw audio set is not available to the public, but rather, 128-dimensional feature vectors, each representing 960 millisecond pieces of the excerpts, are free to download. These features were extracted using a convolutional neural network (CNN), quantized, and subjected to principal component analysis (PCA). Moreover, for a small subset of this dataset, Google provided a temporally detailed label set, describing for each 960 ms piece which sound events (out of 527 possibilities) are present.

We train our ANN on this subset; our objective is binary classification in the sense that the model should determine whether speech is present or not in a 960 ms audio sample just from the 128-dimensional feature vector representing it. Google provided 527 sound event labels, and several of these easily can be considered "speech", and so every feature vector including one or more such events is to be assigned a "true" label. Because the subset is imbalanced, with roughly 18% of excerpts containing speech by our definition, we minimize a weighted binary cross entropy loss function.

Our Model and Results

Our key results are 82.2% accuracy, 66.0% recall, and 64.4% precision, all of which were measured over the evaluation set and averaged across 30 trainings. The latter two are metrics that try to compensate for datasets with imbalanced classes such as ours. Overall, this is still a good result which shows promise for our model's use in speech recognition applications. However, in some testing and examination of the ontology, we note that some human noises, such as laughter, singing, or babbling, tend to be a source of confusion for our model, given their similarities to speech. We classify them as non-speech because we aimed to distinguish deliberate speech from other human-made sounds. More work will need to be done in the future to strengthen the distinction.

Conclusion

In the coming years, machine-learning-aided speech-in-noise detection will become more widespread in use as the technology improves. These results show promise for the use of artificial neural networks to recognize speech, even amid other noise. As more uses for such technology are discovered, such as better hearing diagnostic tests, machine-learning-based approaches such as this will prove useful in meeting the increasing demand for quality detection products.