Project Proposal: Audio Classification for Smart Home Audio Systems

Sean Simon | Springboard Data Science Career Track Capstone 3

Problem Statement

Currently, popular smart home audio systems rely on user input through a controller, either on device or through an app. More advanced products offer voice control through APIs from leading services such as Google or Amazon by using an onboard microphone. But what if a smart speaker can respond to more stimuli beyond voide? The opportunities are plenty. In this vein of thought, can a machine learning method classify sounds into categories a smart loudspeaker can do something with?

Context

Smart home audio systems exist on the market that allow consumers to distribute arrays of loudspeakers throughout their home, office, or otherwise. Such setups can be grouped by room, for example in the home context two loudspeakers in the kitchen, six in the living room, and one in the bedroom. As a service, loudspeakers playing music go a long way to set the mood in an environment with the music they can play. To name a few scenarios, it's not a stretch to imagine your favorite music while cleaning the house, low ambience while working on the computer, or the latest hits to bring a party to the next level. Importantly, in each of those common scenarios the listener's hands are full, and often their minds are more fully on their active task and not on controlling the music. So what if a smart speaker can do the controlling for the listener? Can it recognize footfalls to understand when the listener has moved from the kitchen to the living room, and can the music follow? Can this same feature distinguish pets from humans so as not to turn on erroneously? Can a loudspeaker recognize the click of a keyboard to infer the listener is working, and smartly queue a stimulating but ambient lo-fi playlist? What about finger-snapping or laughter indicating that a party is starting and the volume on the Top 40 should be raised? Within this context sound recognition for home smart audio systems offers myriad promises and unlocks huge potential to offer new consumer experiences.

Criteria for Success

A model or API that can be loaded onto a typical consumer electronic CPU with the ability to classify several sounds common to the home environment. Furthermore, the model must be able to deliver high accuracy (>90%) on certain "on or off" type inputs such as footsteps: if misclassified, it may be jarring to have sound in a room incorrectly turn on or off. For other “mood setting” features, such as laughter or typing, >50% accuracy is reasonable to offer pleasurable consumer experiences for the majority of cases.

Scope of Solution Space

The model must be able to classify sounds common to a household in order to be useful for a home audio environment. However, the model must also be trained on other data so as to strengthen its knowledge of home sounds versus invasive sounds from the street.

Constraints

The model must fit on the available CPU and memory space. The classifications must be well defined enough that customers can be made aware of the experiences they may expect to unlock. The model must work well enough on sound quality produced by the onboard microphone.

Stakeholders

1. Customer experience team
2. Software team
3. Hardware team
4. Audio rendering team

Data Sources

The data for this project will come from the Freesound database, as detailed here: https://www.kaggle.com/c/freesound-audio-tagging

Possible Methods

Split audio from the data source into spectrograms with librosa, and train a deep learning model to classify the labels.

Deliverables

1. Classification model
2. Proposition of at least one platform defining UX experience unlocked by this project