CS221 Project Proposal

Modeling Infant Statistical Learning in Lexical Acquisition through Machine Learning

Jihee Hwang (jiheeh), Krishman Kumar(krishank), Eric Ehizokhale(eokhale)

1. Overview/Intro

Infant language acquisition is an efficient process that has been studied by linguistics and psychologists for decades. While we still don't fully understand yet how an infant can learn a language simply by listening to the environment around them, the most prominent theory proposed to explain the phenomenon so far is the statistical learning model.

The statistical learning model for infant language acquisition theorizes that children obtain lexical, grammatical and even phonological knowledge about a certain language by extracting statistical regularities from any input stream of words. The most well-observed one out of those is the learning in lexical acquisition; while spoken language does not have clear boundaries between words, infants can statistically find patterns of certain phonemes that repeatedly occur across different utterances to figure out the individual words making up the lexicon. For example, consider the given two grammatically acceptable sentences: *This is a dog* and *This was a lemon*. While the two sets of three morphemes *is-a-dog, was-a-lemon* would be pronounced without any breaks in the middle, the child can *statistically* realize that *a* is the repeatedly occurring component, and thus conclude that it is a single word . For this project, we aim to replicate such lexical statistical learning given a set of speech audio inputs by incorporating AI principles learned in class.

2. Input-output examples

We will be primarily collecting spoken language audio files for our dataset. We will test the performance of our system on a variety of sentence types: self-recorded set of sentences given a lexicon that we already have decided on; elementary level speech involving repeated, simple, monosyllabic words most ideally taken from storybook readings; and lastly, on complicated spoken English taken from popular podcasts (RadioLab, This American Life, etc). Each of these data sets would be a milestone to accomplish.

An output with these data would be the lexicon, or words that have been used in the data sets. The 'word' would be in audio form, rather than text form. We expect both our 'audio' input and output to be in a vectorized form. Ideally, if our model works successfully on English audio inputs, then it should be able to work as efficiently in other languages as well.

3. Baseline-Oracle

- Baseline: Look for 'breaks' in speech to detect words. The break is defined as local minima in volume. However, we expect low accuracy with this method, because spoken language does not contain noticeable breaks between words.
- Oracle: Words recognized by a human listener who is a native speaker of the language used in the data set.

The gap between the baseline and oracle performance would be very large. The oracle would have nearly perfect accuracy, only subject to mishearing unclear speech. The baseline has would have a very low accuracy on the other hand, as we don't pause too much between individual words while speaking naturally. Closing the gap between the two objectives would be our main challenge.

4. Challenges

One big challenge we will face with this project is finding a way to generalize a variety of different voices (timbres), accents, and unique colloquial speech patterns that will exist within our dataset. Studies show that infants might perceive external audio as if it was processed through a low-passed filter -- this could potentially help generalize our audio data set.

The high level topics we can use to address these challenges include machine learning - potentially unsupervised learning. Self-designed features could also prove to be important in handling the characteristic of spoken language. Furthermore, we're looking into using k-means during the process of clustering different 'word' audio data.

5. Existing related work

- http://science.sciencemag.org/content/274/5294/1926.full.pdf+html

In 1996, Saffran, Aslin, and Newport conducted a study with 8-month old infants using an artificial grammar and found that they could detect word boundaries based only on transitional probabilities. The study sample audio was spoken in a monotone, constant volume. But in real spoken language, prosodic and phonotactic information also helps distinguish between words.

- http://www.aclweb.org/anthology/E03-3001

Piroska Lendvai used machine learning algorithms to detect fragmented, disfluent words (such as self-corrections, hesitations, etc.) in spoken Dutch. Algorithms included a memory-based classification system and a rule induction classifier. The machine learning optimizations used for spoken language will be relevant to our project.