Alexander I. Rudnicky

Principal Systems Scientist
School of Computer Science
Department of Computer Science
September 5, 2007, 10:00am

Research Topic
Acoustic models for speech recognition

Research Problem
How to select the most useful data for acoustic training out of large datasets?

Problem Statement
Given a large acoustic dataset, construct extract the most useful training set by eliminating poor training data.

An acoustic dataset is a audio recording of the voice to be recognized by the speech recognition application in condition similar for which it will be used.

Poor training data is that part of the audio recording that is too redundant or of poor quality that no training can be gained.

Problem Description
Dr. Rudnicky is interested in a wide variety of topics related to speech recognition. One problem he is working on is discriminating between useful and not useful data in acoustic modeling. It has traditionally been thought that the primary problem in speech recognition was insufficient datasets. Now that large(>1000 hours) datasets are becoming available, researchers are finding that they have too much data to train acoustic models in a reasonable time period. Dr. Rudnicky was able to select only the most useful data for training. Doing so, he was able to do as well with 150 good hours as he could with the full 840 hours.

Computer Science Perspective
The segregation of data is an extremely interesting problem with applications in a variety of data mining and machine learning contexts in addition to the acoustic modeling paradigm. The common practice is to use all available data to train models. If subsets of the data can produce better models, this could have implications for a wide variety of work.

Disciplines actively involved
Acoustics
Linguistics
Statistics

Description of Disciplines Involved
Acoustic as a discipline is involved because someone has to state the condition under which the sample is taken as well as manipulate the dataset to bring it to it outmost qualify before being processed Dr. Rudbnicky's algorithms.

Linguistic are involved because there is a need to understand the language model to be built as well as understanding how to generalize one language model to others.

Finally a statistician is involved because the language model is based on Markov models and other probability techniques which rely heavily on statistics.

References
Presenter web page:
http://www.cs.cmu.edu/~air/
Speech Group:
http://www.speech.cs.cmu.edu/speech/

By: Jamie Olson

Updated: Guillermo Marinero, September 27, 2007