Speech Recognition in Java

Breandan Considine
JetBrains, Inc.
Introduction

• How do I build a speech recognition application?
• What techniques do modern ASR systems use?
• Is speech recognition accessible for developers?
• What libraries and frameworks exist for speech?
Why offline?

• Latency – many applications need fast local recognition
• Mobility – users do not always have an internet connection
• Privacy – data is recorded and analyzed completely offline
• Flexibility – configurable API, language, vocabulary, grammar
Why speech?

• Humans are wired for speech (FOXP2)
• Accessibility, mobility, convenience
• Automatic translation for large dictionaries
• Real-time speech recognition is tractable
Why Java?

- CMU Sphinx 4+
- Deeplearning4j (DL4J)
- Java ecosystem
- Cross-platform
- Developer familiarity
Feature Extraction

• Recording in 16kHz, 16-bit depth, mono, single channel
• Reduction to ~40 frames per second
• 40-60 phonemes to model English
• State space reduction,
• Robust signal processing is very important
  • MFCC
  • PLP
Coding Speech
Acoustic Model

- Database of language phones
- Hidden Markov Model (HMM)
- Gaussian Mixture Model (GMM)
- Pretrained models: WSJ, TIDIGITS, RM1, AN4, HUB4
Dictionary

• Mapping for words to phonemes
• Word error rate increases with dictionary size
• Same word, multiple pronunciations (CLOSE – CLOSE)
• Same pronunciation, different word (CELL – SELL)
Dictionary

autonomous  AO  T  AA  N  AH  M  AH  S
autonomously  AO  T  AA  N  OW  M  AH  S  L  IY
autonomy  AO  T  AA  N  AH  M  IY
autonomy(2)  AH  T  AA  N  AH  M  IY
autopacific  AO  T  OW  P  AH  S  IH  F  IH  K
autopart  AO  T  OW  P  AA  R  T
autoparts  AO  T  OW  P  AA  R  T  S
autopilot  AO  T  OW  P  AY  L  AH  T
JSGF Grammar

• Probabilistic context free grammar

<size> = /10/ small | /2/ medium | /1/ large;
<color> = /0.5/ red | /0.1/ navy blue | /0.2/ sea green;
<action> = please (/20/save files | /1/delete all files);
<place> = /20/ <city> | /5/ <country>;}
Language model

<s> generally cloudy today with scattered outbreaks of rain and drizzle persistent and heavy at times </s>
<s> some dry intervals also with hazy sunshine especially in eastern parts in the morning </s>
<s> highest temperatures nine to thirteen Celsius in a light or moderate mainly east south east breeze </s>
<s> cloudy damp and misty today with spells of rain and drizzle in most places much of this rain will be light and patchy but heavier rain may develop in the west later </s>
Language model

- **Unigram**: $P(\text{back})$
- **Bigram**: $P(\text{back}|\text{look})$ and $P(\text{back}|\text{strikes})$
- **Trigram**: $P(\text{back}|\text{don't, look})$, $P(\text{back}|\text{sun, strikes})$, and $P(\text{back}|\text{empire, strikes})$

http://www.cse.sc.edu/~mgv/csce580f12/gradPres/580f12LingareddyNadipally.pptx
Decoder

• Searches grammar for word likelihoods
• Searches dictionary for word pronunciations
• Maps pronunciations to acoustic model HMMs
• FlatLinguist – in-memory, small vocabulary recognition
• LexTreeLinguist – Dynamic search state generation
Search Graph
How to train your own model

• Acoustic model training is very time consuming (months)
• Language model training is easier (~100,000 sentences)
• Some tools:
  • Boilerpipe (HTML text extraction)
  • Logios (model generation)
  • Imtool (CMU Sphinx)
  • IRSLM
  • MITLM
  • SRILM
Recognition accuracy

\[ WER = \frac{\text{deletions} + \text{substitutions} + \text{insertions}}{\text{words}} \]
Resources

• CMUSphinx, http://cmusphinx.sourceforge.net/wiki/
• Portuguese Speech Models (Universidade Federal do Pará), http://www.laps.ufpa.br/falabrasil/downloads.php
• Deep Learning for Java, http://deeplearning4j.org/
• MaryTTS, http://mary.dfki.de/
• FreeTTS 1.2, http://freetts.sourceforge.net/
• JSpeech Grammar Format, http://www.w3.org/TR/jsgf/
• ARPA format for N-gram backoff (Doug Paul) http://www.speech.sri.com/projects/srilm/manpages/ngram-format.5.html
• Language Model Tool http://www.speech.cs.cmu.edu/tools/lmtool.html
Further Research

• The Application of Hidden Markov Models in Speech Recognition, [http://mi.eng.cam.ac.uk/~mjfg/mjfg_NOW.pdf](http://mi.eng.cam.ac.uk/~mjfg/mjfg_NOW.pdf)
• Accurate and Compact Large Vocabulary Speech Recognition on Mobile Devices, [research.google.com/pubs/archive/41176.pdf](http://research.google.com/pubs/archive/41176.pdf)
• Deep Neural Networks for Acoustic Modeling in Speech Recognition, [research.google.com/pubs/archive/38131.pdf](http://research.google.com/pubs/archive/38131.pdf)
Special Thanks

• CMU Sphinx (@cmuspeechgroup)
• Anderson Casimiro (@duodraco)
• Hanneli Tavente (@hannelita)
• Stephen Chin (@steveonjava)