Speech Technology Using in Wechat

FENG RAO

Outline

• Introduce Algorithm of Speech Recognition
  – Acoustic Model
  – Language Model
  – Decoder

• Speech Technology Open Platform
  – Framework of Speech Recognition
  – Products of Speech Recognition
  – Speech Synthesis
  – Speaker Verification
Speech Recognition

\[ \hat{W} = \arg\max_{W \in L} P(W \mid O) \]

\[ \hat{W} = \arg\max_{W \in L} \frac{P(O \mid W)P(W)}{P(O)} \]

\[ \hat{W} = \arg\max_{W \in L} P(O \mid W)P(W) \]

Acoustic Model

- Spoken words: “I think there are”
- Phonemes: ‘ay-th-in-nk-dh-eh-r-aa-r’
- Each tri-phone corresponds to a HMM
- H.M.M: 5-state representation
- Each state corresponds to a mixture Gaussian model
Acoustic Model

\[ P(O|S) = \sum_{i=1}^{M} \sigma N(O|\mu_i, \sum_j) \]

\[ P(O|W) = \sum_{i=1}^{M} P(O|S) \]

Deep Neural Network

\[ P(Y = i|x,W,b) = \text{softmax}(Wx + b) = \frac{e^{Wx+b_i}}{\sum_j e^{Wx+b_j}} \]
Language Model

- **N-Gram Model**
  - Build the LM by calculating n-gram probabilities from text training corpus: how likely is one word to follow another? To follow the two previous words?

\[
p(S) = p(W_1, W_2, \ldots, W_i) = p(W_i) p(W_{i-1} | W_i) \ldots p(W_1 | W_2, W_3, \ldots, W_{i-1})
\]
  - Smooth methods
    - KN, GT, Stupid Backoff

- **Grammar**
  - ABNF, is to describe a formal system of a language to be used as bidirectional communication protocol.
  - Quick, Small

---

N-Gram

- \data\n- ngram 1=4
- ngram 2=3
- ngram 3=2
- \1-grams:\n  - 0.60206 hello -0.39794
  - -0.60206 world -0.3979
  - -0.60206 </s> -0.39794
  - -0.60206 <s> -0.39794
- \2-grams:0
  - 0 hello world -0.39794
  - 0 world </s> -0.39794
  - 0 </s> hello -0.39794
- \3-grams:
  - 0 hello world </s>
  - 0 </s> hello world\end\n
Grammar

```perl
public $basicCmd = $digit<1->;
$digit = {0|1|2|3|4|5|6|7|8};
```
Decoder

- Find the best hypothesis $P(O|W) P(W)$ given
  - A sequence of acoustic feature vectors ($O$)
  - A trained HMM ($AM$)
  - Lexicon ($PM$)
  - Probabilities of word sequences ($LM$)
- For $O$
  - Weighted finite state transducer
  - Build network composed with HMM trip-hone and words in Am and Lm.
  - Calculate most likely state sequence in HMM given transition and observation probs.
  - Trace back through state sequence to get the word sequence.
  - Viterbi decoder
  - N best vs. 1 best vs. lattice output
- Limiting search
  - Lattice minimization and determination
  - Pruning: beam search

Decoder Network

- Viterbi Decoder Process
Decoder Network

• Viterbi Decoder Process

Decoder Network

• Viterbi Decoder Process
Decoder Network

• Viterbi Decoder Process

start → t → End
Challenge Under Internet

• Big Training Data
  – Txt corpus is TB level and thousand hours of speech data as training data
  – Speed Optimized methods
• Large Mount of Users
  – Real time response
  – More machines, Robust service
• Quick Update
  – Content in Internet in changing every day.
  – Update model especially on language model

Speech Open Platform

  --using in wechat

  Speech recognition
  Speech synthesis
  Speaker verification
  ...

One Network
Multiple Products

Universal Interface

decoder

General Filed LM

decoder

Map filed LM

decoder

Command Filed LM

decoder

Others...

One Network
Multiple Technology

Universal Interface

Ngram Model
ABNF
Parallel Decoding Space

One Pass Decoder
Lattice Decoder

GMM
DNN

One –Best
N-Best
Recognition rate

Non-Finite Field
The core performance of Speech Recognition is optimized and developing

- Accuracy rate: 94% (Audio sampling at 16kHz)
- Usage amount: 18 million per day

![Sampling Accuracy of Current Availability](chart)

Vertical Fields

Multi Verticals
Unify entrance with Parallel decoding of space technology

- Parallel recognition supports 11 classifications of verticals
- 30% better in performance than speech input in Verticals
- Recognition rate: 96%, more accurate than common

![Vertical Fields Chart](chart)
Speech Technology Product

• Speech to text
  
  Wechat Input  QQ Input  Input Tool

Speech Technology Product

• Vertical Application
  
  Music Searching  QQ Map
Contact Searching

Voice Quality Identify

Voice Awaken To Unlock Mobile Phone

Speech Synthesis

• Features
  – 1. High efficient synthesis.
  – 3. Offline and Online TTS

• Applications
  – 1. WeChat Official Account.
  – 2. WeCall.
  .
Speaker verification

- Application of scene
  - User login verification
  - Bank transfer, payment verification
  - Forgot password

- Advantage:
  - Convenient, fast
  - Safety
  - Good user experience

How To Get Speech Technology

- http://pr.weixin.qq.com/voice/intro
Thanks