Speech Technology Using in Wechat

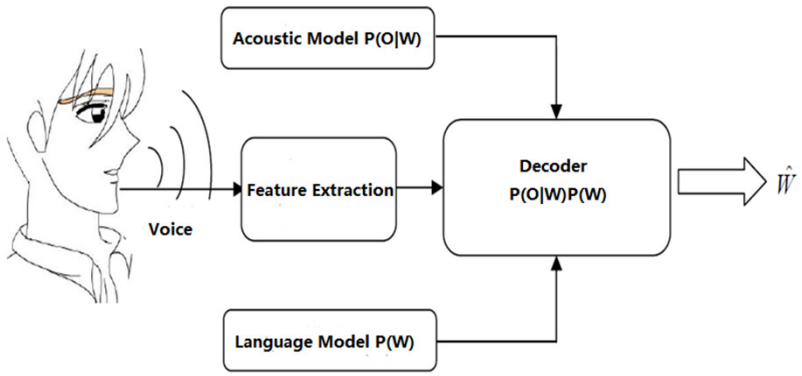
FENG RAO

Powered by WeChat

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

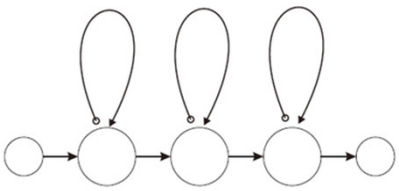
Speech Recognition



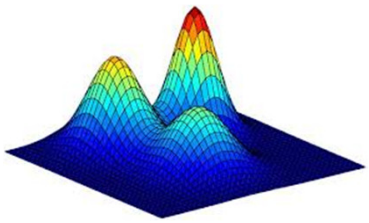
$$\hat{W} = \arg\max_{W \in L} P(W \mid O) \Rightarrow \hat{W} = \arg\max_{W \in L} \frac{P(O \mid W)P(W)}{P(O)} \Rightarrow \hat{W} = \arg\max_{W \in L} P(O \mid W)P(W)$$

Acoustic Model

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



Acoustic Model



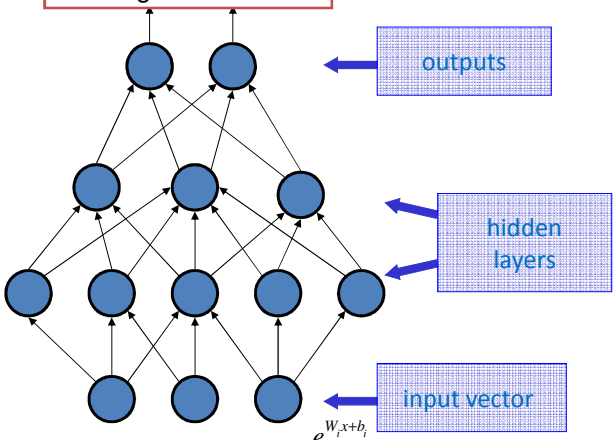
$$P(O \mid S) = \sum_{i=1}^{M} \varpi_i N(O \mid \mu_i, \sum_i)$$

$$P(O \mid W) = \sum_{i=1}^{M} P(O \mid S)$$

Deep Neural Network

Compare outputs with correct answer to get error signal

Back-propagate error signal to get derivatives for learning



$$P(Y = i \mid x, W, b) = soft\max_i(Wx + b) = \frac{e^{W_i x + b_i}}{\sum_j e^{W_j x + b_j}}$$

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?

$$p(S) = p(W_1, W_2, ..., W_k) = p(W_1)\,p(W_2 \mid W_1)...p(W_k \mid W_1, W_2, W_3, ..., W_{k-1})$$

  - 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\
ngram 1=4
ngram 2=3
ngram 3=2
\1-grams:-
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\
```

Grammar

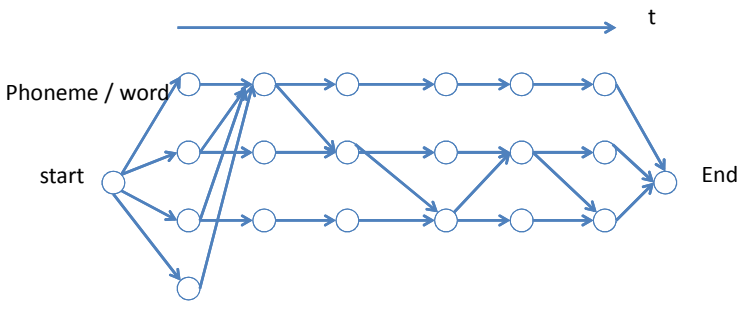
```
public $basicCmd = $digit<1->;
$digit =  (0|1|2|3|4|5|6|7|8);
```

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

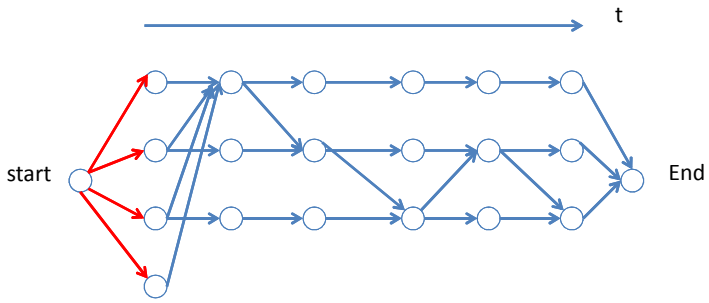
Decoder Network

- Viterbi Decoder Process



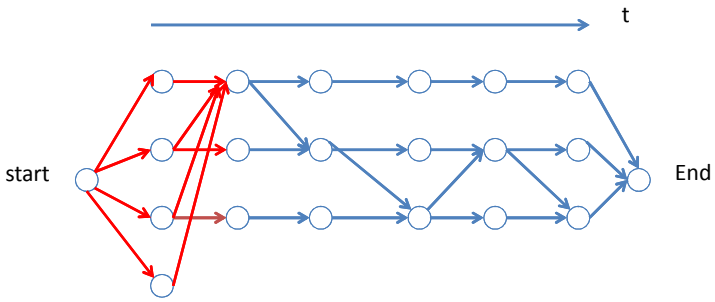
Decoder Network

- Viterbi Decoder Process



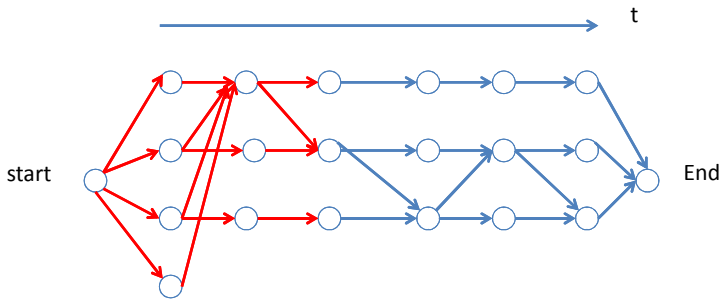
Decoder Network

- Viterbi Decoder Process



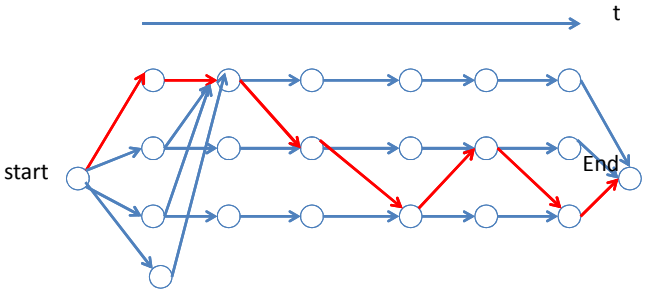
Decoder Network

- Viterbi Decoder Process



Decoder Network

- Viterbi Decoder Process



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

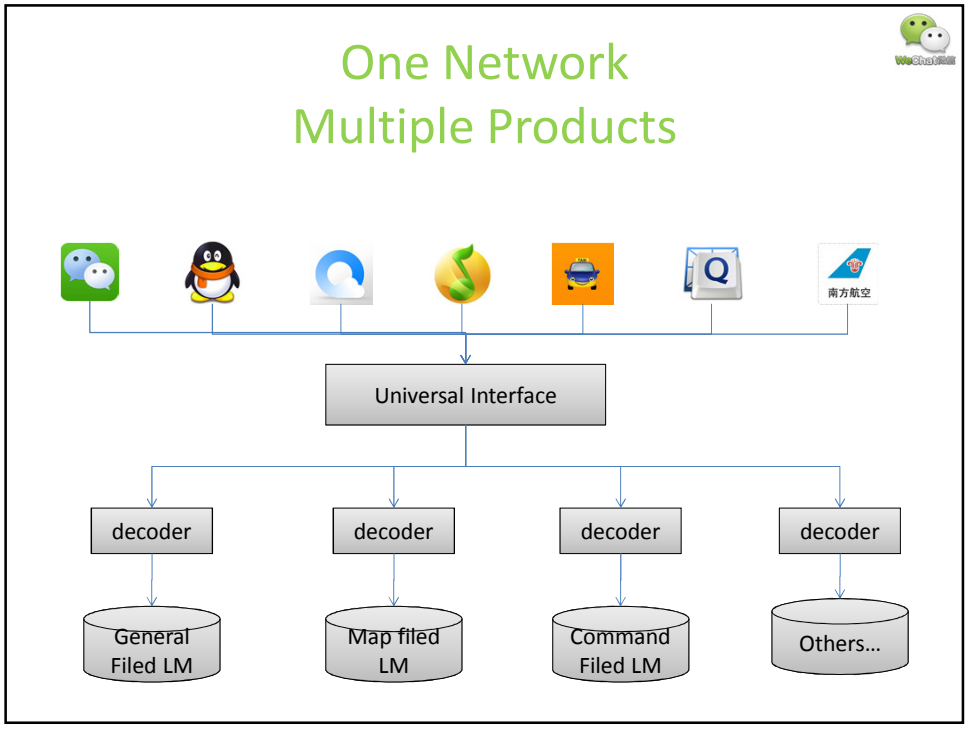
--using in wechat

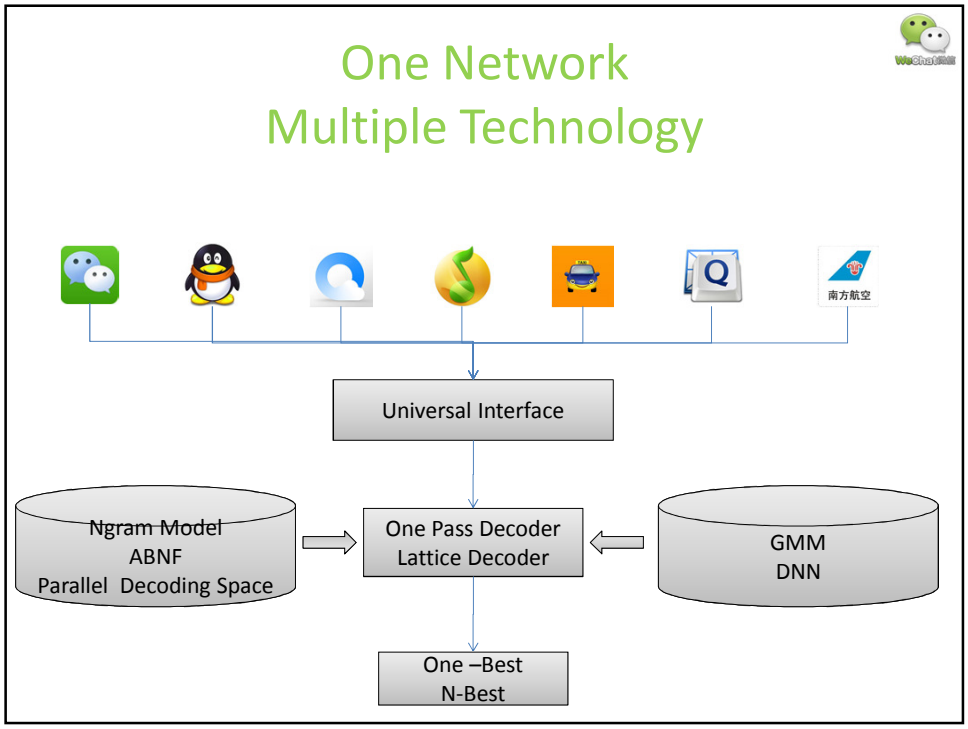
Speech recognition

Speech synthesis

Speaker verification

…


One Network
Multiple Products
Universal Interface
decoder
decoder
decoder
decoder
General Filed LM
Map filed LM
Command Filed LM
Others…

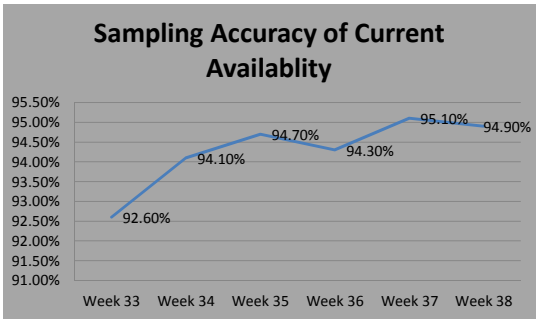

One Network
Multiple Technology
Universal Interface
Ngram Model
ABNF
Parallel Decoding Space
One Pass Decoder
Lattice Decoder
GMM
DNN
One –Best
N-Best

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



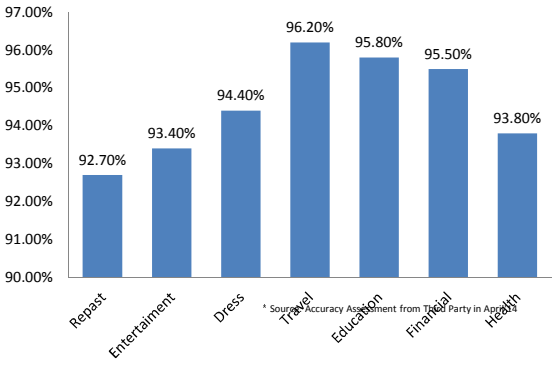
* Source: Accuracy Assessment from Third Party in April'14

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



* Source: Accuracy Assessment from Third Party in April'14

Speech Technology Product

- Speech to text

Wechat Input

QQ Input

Input Tool

21

Speech Technology Product

- Vertical Application

Music Searching

QQ Map

Contact Searching

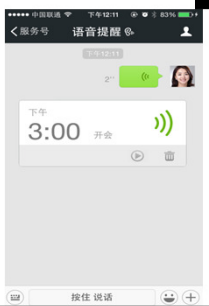
Voice Quality Identify

Voice Awaken To Unlock Mobile Phone

Speech Synthesis

- Features
  - 1. High efficient synthesis.
  - 2. Available SDK for Android and iOS clients.
  - 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