Speech recognition

Computer Literacy 1 Lecture 22
10/11/2008

Topics

- Definition of speech recognition
- Brief history
- Technology
- How does speech recognition work
- Speaker recognition
- Problems of speech and speaker recognition

Definition

- Can also be called automatic speech recognition or computer speech recognition
- Definition:

Speech recognition converts the spoken words into machine readable into machine readable input by using binary code!

History - Homer Dudley

- In the 1930s Homer Dudley created the first human voice synthesizer at the Bell Labs
- He started experimenting with electromechanical devices to produce analogues of human speech in the 20s
- His findings let to the patent for “Vocoder” (voice + encoder)

→ a method of reproducing speech through electronic means and allowing it to be transmitted over distances (e.g. telephone lines)

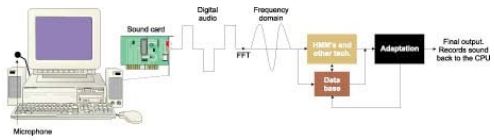
The Vocoder

- Originally developed as a speech decoder for telecommunication
  - Primary use for secure radio communication, where voice has to be encrypted before transmitted
  - Was used in SIGSALY system for high-level communications during WW-II
- Additionally Vocoder's hardware and software has been used as an electronic music instrument (Robert Moog, Kraftwek, Pink Floyd)

Speech recognition - Voice recognition

- What you can already see is that speech and voice recognition can refer to the same technology
- So you can treat these terms as synonyms
- BUT there is also speaker recognition (which falls into the area of speech/voice recognition)

Technology



More Technology

- A speech signal is recoded by a microphone and captured with a sound card
- The speech signal has now to pass through various stages
- Here various mathematical and statistical methods are applied

Inside the computer

- After the voice input is captured on your sound card
- The digital audio output of your card is processed using FFT (Fast Fourier Transform)
- This now already fine-tuned signal is further processed by a HMM (Hidden Markov Model)

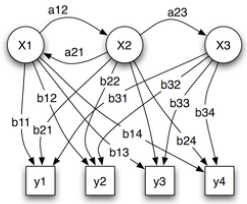
Fast Fourier Transforms (FFT)

- The Fourier Transform is, in mathematics, an operation that transforms one function of a real variable into another
  - It works similar to the way that a chord of music we can hear can be transcribed by notes that are being played
- The FFT is an algorithm to compute the Discrete Fourier Transform (DFT), which is one form of Fourier analysis

Hidden Markov Model (HMM)

- Simply said: **An HMM figures out when speech starts and stops**
  - It is a statistical model
  - An HMM can be considered as the simplest dynamic Bayesian network

HMM



- x = states; y = possible variations; a = state transition probabilities; b = output probabilities

Sound

- Sound itself is analogue that's why we need to translate the signal into a digital signal which is readable by a speech recognising software
- That's what the FFT does, it transforms the incoming signal in a band of frequencies
- When this is done the next step is recognising these bands

How does this work?

- The speech recognition software has a database containing thousands of frequencies → Phonemes
  - A phoneme is the smallest unit of speech in a language or dialect
  - The sound of one phoneme is usually different from another, this can change the meaning of a word
    - E.g. sound 'b' in bat, 'r' in rat
  - The phoneme data base is matching the audio frequency bands that were sampled
  - Each phoneme is tagged with a feature number

How does it figure out the right sound?

- The software has to use complex technique to approximate the sound and figure out what phonemes are used
- One way of identifying relevant phonemes is to train your speech recognition software
- Or you could prune your software for a particular speech

Pruning

- When pruning the software generates several hypothesis on what could have been spoken
- It then generates scores for these hypothesis and decides to go for the one with the highest score
- The ones with the lower scores get pruned out

Train your Speech Recogniser

- When you train your software
- You feed it with many variations of the same phoneme and your software analyses all of these through a statistical methods (e.g. using HMM)
- With the help of this great amount of training phonemes your software gives again feature numbers to specific frequency bands

More training

- So your software applied feature numbers to frequency bands
- Now it uses statistics to figure out the probability of a particular feature number appearing in a phoneme
- The feature number with the highest probability would correspond with the phoneme you've spoken

Speaker recognition

- Speaker recognition = WHO is speaking
  - Speech recognition = WHAT is said
- Identifying characteristics of one voice
- Characteristics of voice are e.g. pitch, melody, hoarse vs soft, frequency

The 2 phases of speaker recognition

- Speaker's voice is recorded and a number of individual features (characteristics) of voice are used to make a voice print
  - In speaker verification this print will be compared to a previous recorded template to verify your voice
  - In speaker identification your voice print is compared to multiple voice prints in order to determine the best match

Possible Problems of Speech and Speaker Recognition

- Speech recognition can't work perfect since people speak in different dialects, use all kind of different pronunciation, HMMs can't always distinguish when speech starts and ends since background noise can be confused with speech, etc…
- Speaker recognition fails as soon as your voice quality is different to your sample, e.g. when you have a cold, aging can have an effect on your voice, etc…

Key points

- The Vocoder, first speech synthesizer
- Speech recognition and it's technology
- Fast Fourier Transformation
- The Hidden Markov Model
- Train and prune your recogniser
- Voice recognition involves verification and identification
- We all speak so differently and our voices are changing through life which makes it very hard to be a good speech recogniser