IIS for Speech Processing

Michael J. Watts

http://mike.watts.net.nz

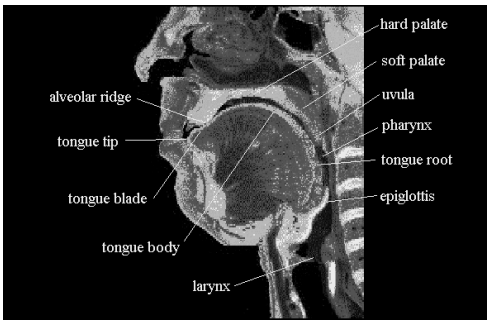
Lecture Outline

- Speech Production
- Speech Segments
- Speech Data Capture
- Representing Speech Data
- Speech Processing
- Speech Recognition

Speech Production

- Speech is a sound wave
  - pressure wave
- Created by air passing over the organs of speech
- Sounds modified by the tongue, teeth and lips

Speech Production



- Http://www.umanitoba.ca/faculties/arts/linguistics/russell/138/sec1/anatomy.htm

Speech Segments

- There are many levels of speech that linguists deal with
  - Sentence
  - Clause
  - Phrase
  - Word
  - Phoneme
- Not a strict hierarchy

Speech Segments

- A sentence consists of one or more clauses
- A clause has at least a predicate, a subject and expresses a proposition
  - predicate expresses something about the subject
  - proposition is the part of the clause that is constant

Speech Segments

- A phrase is more than one word, but less than a clause
  - less organised
- A word is a part of a phrase
  - minimal possible meaningful unit
  - consists of one or more phonemes
  - there are 250,000+ words in the English language

Speech Segments

- A phoneme is the smallest unique segment in a spoken language
  - there are 44 phonemes in standard New Zealand English
  - standard (English) English has 43
- A phoneme consists of one or more phones

Speech Data Capture

- With a microphone
  - well, Duh!
- Digitising sound
  - two parameters
  - sample rate
  - Resolution

Speech Data Capture

- A sound wave is a continuous wave
- Digital capture (sampling) is discrete
- How often a sample is taken is the sample rate
  - 44kHz = 44,000 samples / second
- The more samples (higher sample rate) the better the quality of the recording

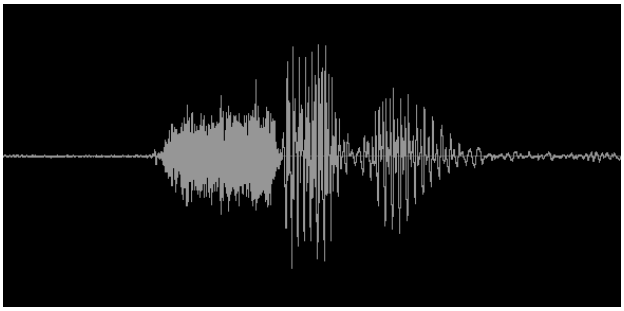
Speech Data Capture

- Resolution is the number of bits used to "describe" each sample
- Higher sample resolution means a greater number of sounds can be represented
- Most CDs are 16-bit, 44.1kHz

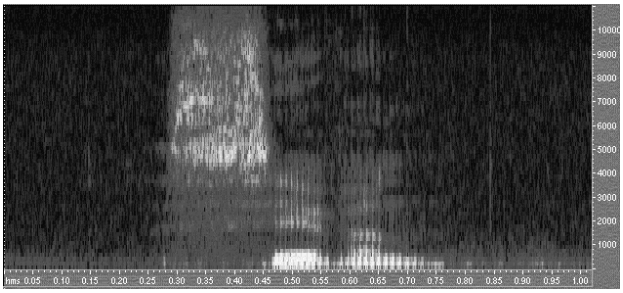
Representing Speech Data

- Two main methods of visualisation
- Waveform
  - plot of signal amplitude against time
- Spectrogram
  - energy-frequency-time plots

Representing Speech Data



Representing Speech Data



Representing Speech Data

- Speech signal has a lot of variation
- Hard to recognise speech from the raw signal
- Need to reduce the signal somehow
- Various transforms are available
- Many based on FFT

Representing Speech Data

- Common transform is the mel-scale
- The mel-scale is a log scale
- Models human perception
- Divides the signal into frequency bands
- Returns the log-energy for each frequency band

Speech Processing

- Speech encoding
  - Compression or de-noising of a speech signal
  - ANN are good at compression
- Speaker identification
  - Who is speaking at any one time?
- Language identification
  - what language is being spoken?
- Speech recognition

Speech Recognition

- Two general types
- Continuous
  - recognises words without pauses between them
  - recognise individual words, but uses surrounding words for increased accuracy
- Discrete
  - recognises single words

Speech Recognition

- Systems can be word or phoneme based
- Word based systems are easier to build
- Require a restricted vocabulary
  - require a recogniser for each word
    - how many words in English?
- The most common used, however

Speech Recognition

- Phoneme based system require fewer recognisers
  - ~40 phonemes in most English accents
- Require a means of assembling the phonemes into words
  - harder than it sounds

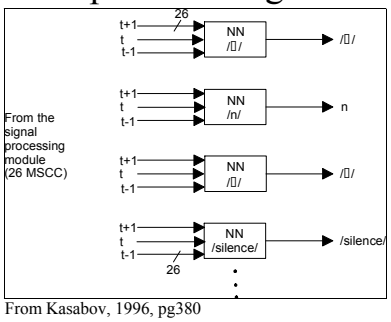
Speech Recognition

- ANN can be trained for both word and phoneme based recognition
- Common models used are MLP, SOM and recurrent networks
- Monolithic vs. Multi-modular
  - speed vs. simplicity

Speech Recognition

- Phoneme based speech recognition with MLP
- Create a MLP for each phoneme
- Train the MLP to activate for the target phoneme and reject all others
- Need to combine the outputs in a meaningful way

Speech Recognition



From Kasabov, 1996, pg380

Speech Recognition

- Phoneme base speech recognition with SOM
- Train a SOM on the phonemes
  - phonemes will cluster together
- Node that is active identifies the current phoneme
  - one of the earliest applications of SOMs

Speech Recognition

- Word based systems work in a similar manner
- Obviates the need for assembling phonemes into words
- Large number of words restricts the application areas

Speech Recognition

- Fuzzy systems can also be used
- Fuzzy rules assemble the phoneme streams into words
- Hybrid systems

Summary

- Speech processing is a complex problem
- Large variability in speech signals
- Main application is speech recognition
- Many types of speech recognition methods
- Intelligent systems can be useful for this problem