

Talking to Computers and Computers Talking to You

CA107 Topics in Computing Lecture

Nov 8, 2004

John McKenna

John.McKenna@computing.dcu.ie

1

Overview

- What are we dealing with?
 - Sounds and Speech
- Talking to Computers
 - Speech Recognition
- Computers Talking to You
 - Speech Synthesis
- My Research
 - Speaker Characterisation

2

Sounds and Speech

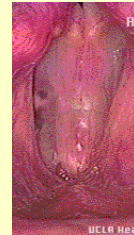
- Words contain sequences of sounds
- Each sound (phone) is produced by sending signals from the brain to the vocal articulators
- The vocal articulators produce variations in air pressure
- These variations are transmitted through the air as complex waves
- These waves are received by the ear and signals are sent to the brain

3

Articulation



Vocal Tract



Vocal Folds

4

Sound Production

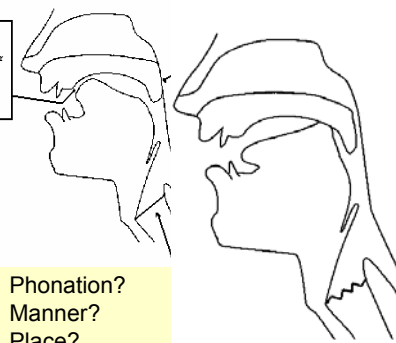


- Vocal folds open and close rapidly
- Their rate of opening/closing determines what we perceive as pitch
- Some consonants are *voiceless*
- Vocal tract configuration determines the sound *quality*

5

How Sounds Vary

Structure of close approximation behind the alveolar ridge causing turbulence (postalveolar (fricative sound))



Phonation?
Manner?
Place?
Nasality?

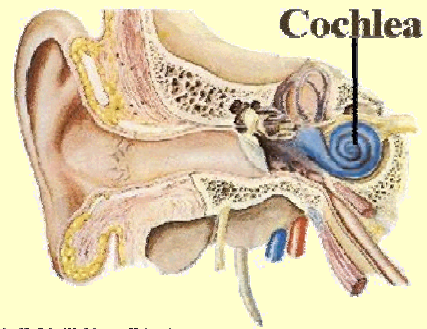
6

Acoustics: Vowels

- All vowels are voiced (except whispered vowels)
- Vocal tract independent of vocal folds
- So we have two things we can vary
 - Rate of vocal folds' opening/closing
 - Vocal tract configuration
- What is it that causes us to perceive differences?
- Let's look at the ear...

7

The Ear



Adv. N. Salt, Washington University

8

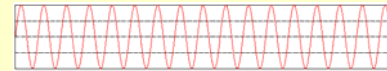
The Ear, Waves & Frequencies

- The cochlea in ear is sensitive to frequency
- What do we mean by frequency?
- We use frequency to describe phenomena that repeat regularly in time
- E.g. a tuning fork vibrates at a certain frequency
- Its oscillations cause air pressure variations

9

Waves and Spectra

Simple wave

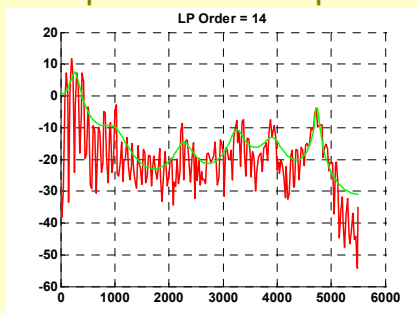


Complex wave



10

Spectral Envelope



Harmonics of F0 vs. Formants (resonances) 11

Computers

- When machines produce sound...
 - Signals are sent from a program to speakers
 - I.e. speakers replace the articulators
- When machines receive sound
 - The microphone replaces the ear
 - Signals are sent from microphone to program
- Sound card: intermediate controller/processor
 - The articulator muscles
 - Cochlea in ear

12

Conclusions

- Speech contains a variety of different sounds called phones. (Pitch not really relevant.)
- If we want to process speech, we analyse/synthesise at the acoustic level
- Acoustically, speech is a series of complex waves which contain oscillations of many frequencies
- The relative strengths of these frequencies characterise sounds
- Knowing/learning these characteristics allows us to process speech

13

Note on Speakers

- Acoustics depend on articulators
- Articulators vary across speakers
- So acoustics vary across speakers
- This can be problematic in speech processing
 - More later...

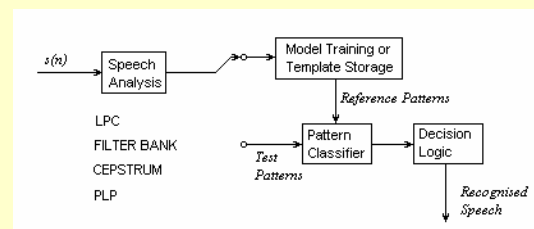
14

Automatic Speech Recognition

- Techniques
 - Template matching
 - Probabilistic modelling
- Examples
- Issues
- Related Tasks
- Demos

15

Architecture



16

Techniques in ASR

- Template Matching
 - Used in voice dialling on mobiles
 - Calculate distances between test utterance and each stored template
 - Choose template with minimum distance
- Probabilistic Modelling
 - Train models with multiple utterances
 - Calculate the likelihood that the test utterance was produced by each model
 - Choose model with highest probability
- Examples

17

Issues in ASR

- Speaker dependent/independent
- Vocabulary size
- Isolated word vs. Continuous speech
- Language modelling constraints
 - Level of ambiguity in vocabulary
- Environment e.g. noise considerations

18

Related Tasks

- Speaker Recognition
 - Speaker Identification
 - Speaker Verification
- Speaker Adaptation
- Speaker Normalisation

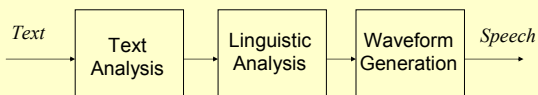
19

Speech Synthesis

- Text-To-Speech (TTS)
 - Typical architecture
 - Festival
- Demos
 - MBROLA

20

TTS Architecture



21

Text & Linguistic Processing

- Language modelling generates phone sequence and prosody of the target utterance.
 - Tokenisation
 - Parsing
 - POS tagging
 - Word pronunciation and letter to sound rules
 - Prosodic modelling

22

Tokenisation

- Easier for English than Chinese
- Wewenttotheseasideadayago

23

Parsing & Tagging

- Parsing helps phrasing
 - Phrasing helps decide natural pauses
 - He went to the drive in to see the movie
- POS tagging aids disambiguation
 - I live in Reading
 - I can fish
 - Homographs: e.g. 1996

24

Dictionary/Lexicon

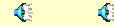
- Phonemic info
- Prosodic info
- E.g.:

```
( "present" v (((pre) 0) ((z@nt) 1)) )  
( "monument" n (((mo) 1) ((nyu) 0) ((m@nt) 0)) )  
( "lives" n (((lai v z) 1)) )  
( "lives" v (((liv z) 1)) )
```
- What if no entry?
 - E.g. proper nouns
 - Letter-to-sound rules

25

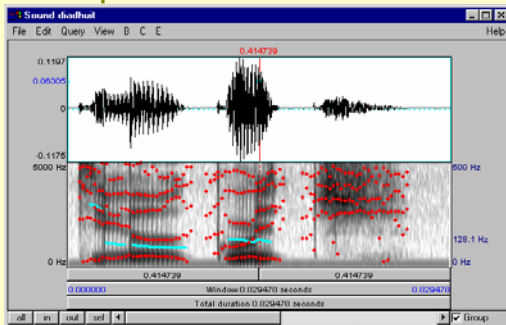
Speech Signal Generation

- Concatenative
 - What units?
 - Words, syllables, phones, diphones
 - Unit selection
 - Post-processing
- Other types
 - Formant
 - Articulatory



26

Diphone concatenation



diəgwɪf

27

Signal Postprocessing

- Manipulate duration
- Manipulate pitch
- Smooth joins



28

Evaluation

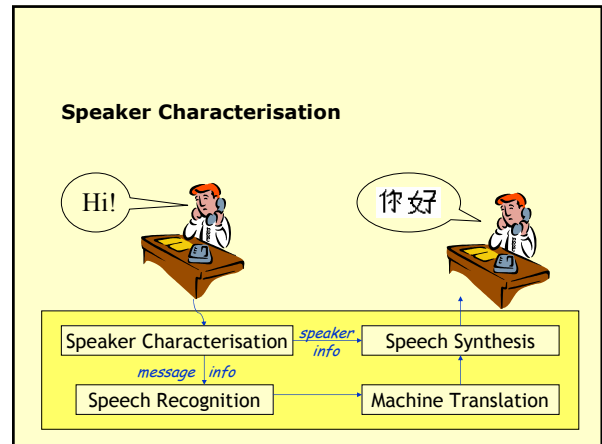
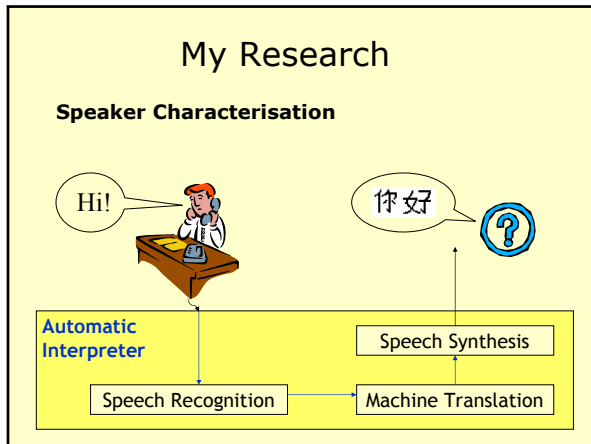
- Intelligibility
- Naturalness
- Perceptual tests
- Psychoacoustics

29

Future Trends

- Best synthesis units?
- Speech signal modification
- Voice conversion
- Variability: style, mood, ...
- Better models

30



- ## Note
- All the software used in the demos today is either available in the CA labs and/or is downloadable for free.
 - In the CA labs, choose
 - > Programs
 - > Computational Linguistics
 - > *Package*
 - Feel free to experiment
- 33