MPhil in Machine Learning and Machine Intelligence

Module MLMI2: Speech Recognition

L1: Introduction

Phil Woodland
pcw@eng.cam.ac.uk

Michaelmas 2021

Cambridge University Engineering Department

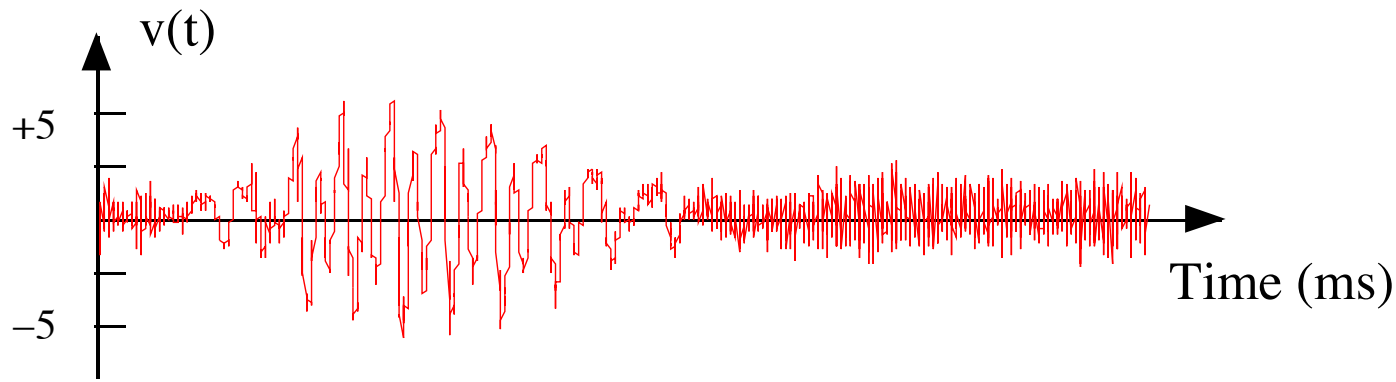
Outline

- Introduction to Speech and Automatic Speech Recognition (ASR)
- Speech Signal Processing
- Hidden Markov Models (HMMs) for Speech Recognition
- Continuous Speech Recognition
- Acoustic Modelling for Large Vocabulary Recognition
- HMMs with Deep Neural Network Acoustic Models: hybrid (DNN-HMMs)
- Language Models & N-Grams
- Front-End Transformations and DNN-based features (tandem models).
- Adaptation
- Large Vocabulary Search and System Design
- Advanced Neural Network Acoustic Models
- Discriminative Sequence Training
- Neural Network Language Models
- End-to-End Trainable ASR Systems

Assessment is via a speech recognition practical on GMM-HMM and DNN-HMM systems. The practicals look at phone recognition with the TIMIT corpus and use the HTK toolkit (see `http://htk.eng.cam.ac.uk`). The write-up includes the practical results with discussion and analysis.

The Speech Waveform

The speech waveform (the electrical signal from a microphone representing the speech pressure waveform) is **non-stationary** and it contains a mix of **pseudo-periodic** and **random** components.



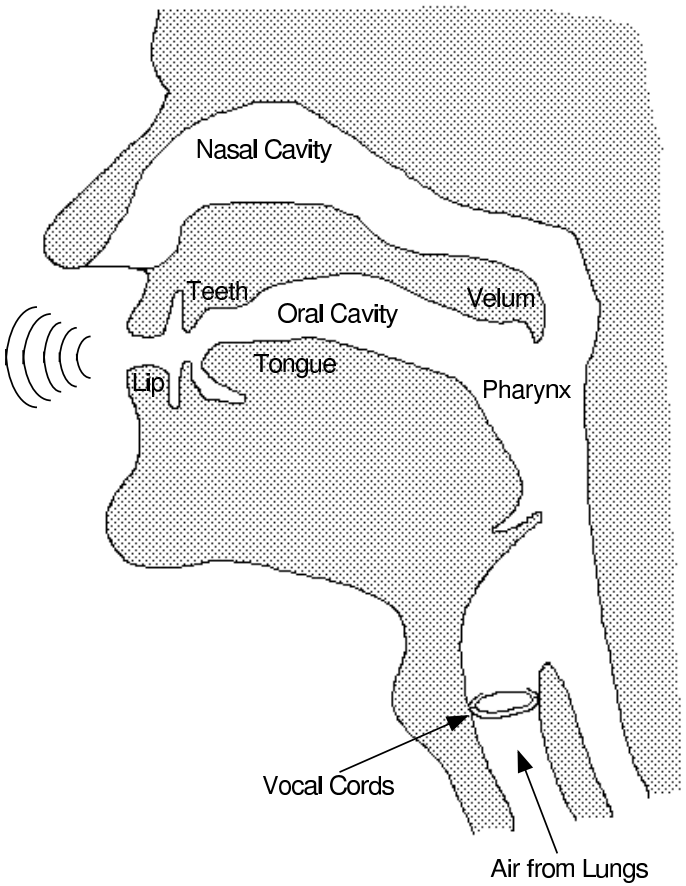
Different classes of speech sounds have properties that depend on how they were produced.

There are two main components to the human speech production mechanism: a variably-shaped **acoustic tube** and an **excitation source** for the tube.

Some broad distinctions in speech sound are due to the type of excitation and detailed sounds are due to the shape of the tube.

In this lecture we will mainly look at the **time-domain** features of speech and later we will look at the speech signal in the **frequency domain** and look at some aspects of speech signal processing.

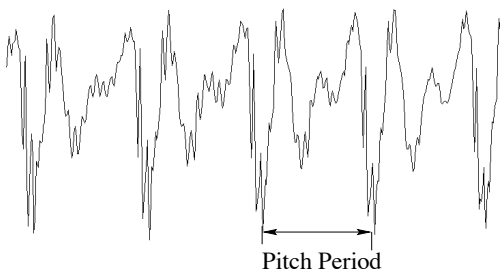
Human Vocal Tract (Cross-section)



- Lungs push air through larynx
- Vocal folds vibrate to release air in puffs, creating a periodic excitation for voiced speech
- Unvoiced sounds from turbulence sound source in vocal tract (fricatives) or build up and release (plosives)
- Articulators (tounge, jaw etc) move to position for particular speech sound: move continually as sounds produced
- Note neighbouring sounds effect realisation of each sound: **co-articulation**.
- Resonances in the vocal tract (formants) alter the signal to shape the sound formed
- Side branch used in nasal sounds
- Sound pressure waves leave mouth and travel through air

Excitation Sources

The acoustic tube has three sources of excitation



1. Vocal cords which vibrate when air from the lungs is forced through them.

This leads to **voiced** sounds as in “feel”, “hit”, “wool”. The sounds are quasi-periodic at the **pitch** frequency.

2. Turbulence caused by forcing air through constrictions formed by raising the tongue to narrow the acoustic tube.

This leads to **fricative** sounds which appear random in the time-domain as in “feel”, “shoe”, etc.

3. Turbulence caused by the release of air following a complete closure of the acoustic tube. This leads to **plosive** sounds as in “take”, “rap”, etc. Note that sounds may also have mixed excitation as for example in “zoo”.

The Sounds of English

It is convenient to view speech as being composed of a sequence of sounds called **phones**.

These sounds are directly associated with basic units of speech, the **phonemes** (use of different phonemes changes meaning in a language)

- Some sounds, particularly vowels, form a continuum.
- Hence various choices of phone set are possible.
- For English about 40 phones used
- For American English, **ARPAbet** is often used.
- The sentence “This is speech” consists of 9 phonemes in sequence (no word boundaries!)

  th ih s ih z s p iy ch

Fricatives		Plosives		Liquids		Nasals	
f	full	p	put	l	like	m	man
v	very	b	but	r	run	n	not
s	some	t	ten	hh	hat	ng	long
z	zeal	d	den	w	went		
sh	ship	k	can	y	yes	Affricates	
zh	measure	g	game			ch	chain
th	thin					jh	judge
dh	then						
Vowels						Diphthongs	
iy	bean	uw	moon	er	burn	ay	buy
ih	pit	uh	good	ax	about	oy	boy
ey	bay	ah	putt	ow	no	aw	now
aa	barn	ao	born	eh	pet	ia	peer
ae	pat	oh	pot			ea	pair
						ua	poor

The ARPAbet American English Phone Set

Consonant Classification

Consonant sounds may be divided into 5 broad classes depending on the type of vocal tract constriction:

- plosives (stops)
- fricatives
- affricates
- liquids (semi-vowels)
- nasals.

Within each class the individual sounds are distinguished by the place at which the constriction occurs and whether or not there is voicing.

The table below shows how each of the consonants in the ARPAbet are classified.

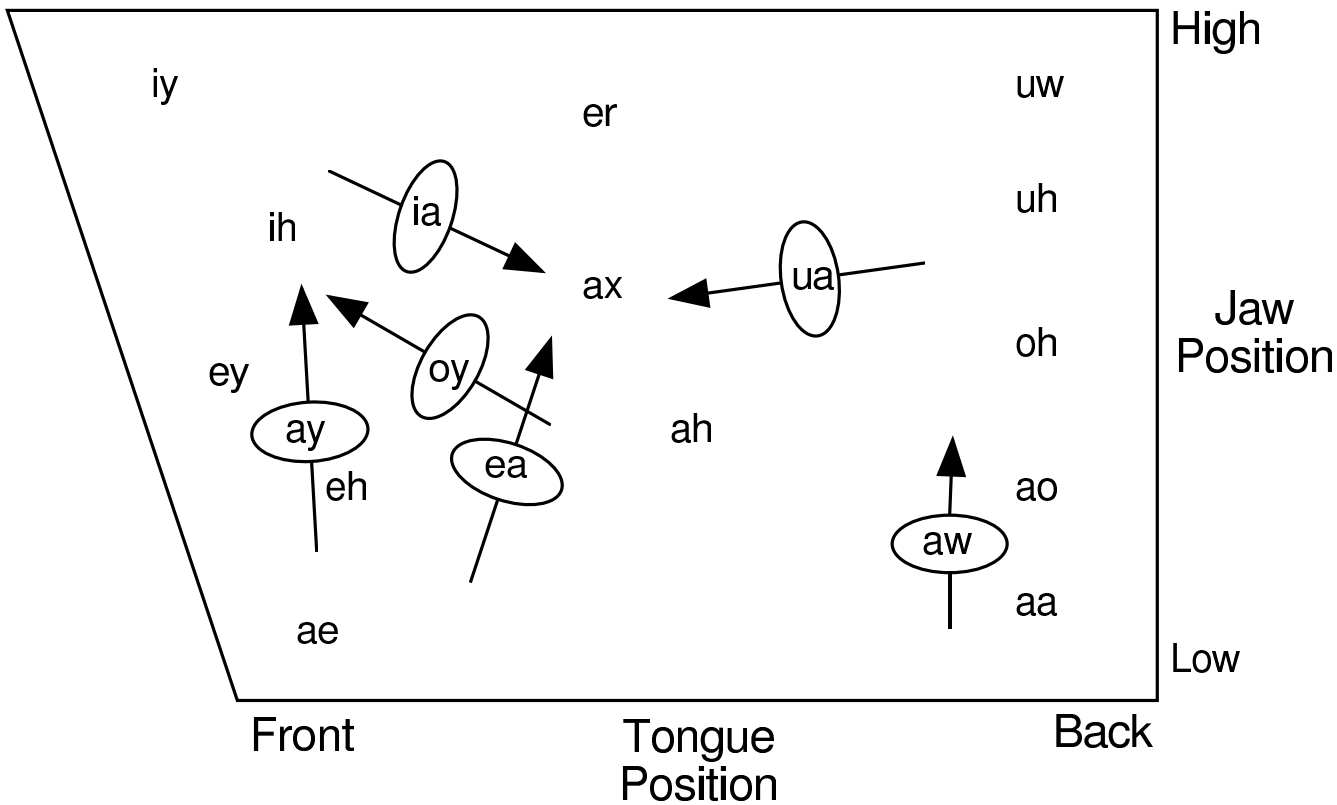
	lip-lip	lip-teeth	teeth-tongue	alveolar-tongue	palate-tongue	velum-tongue
nasal	m			n		ng
stop	p b			t d		k g
fricative		f v	th dh	s z	sh zh	
liquid	w			r l	y	
affricate				ch jh		

Vowel Classification

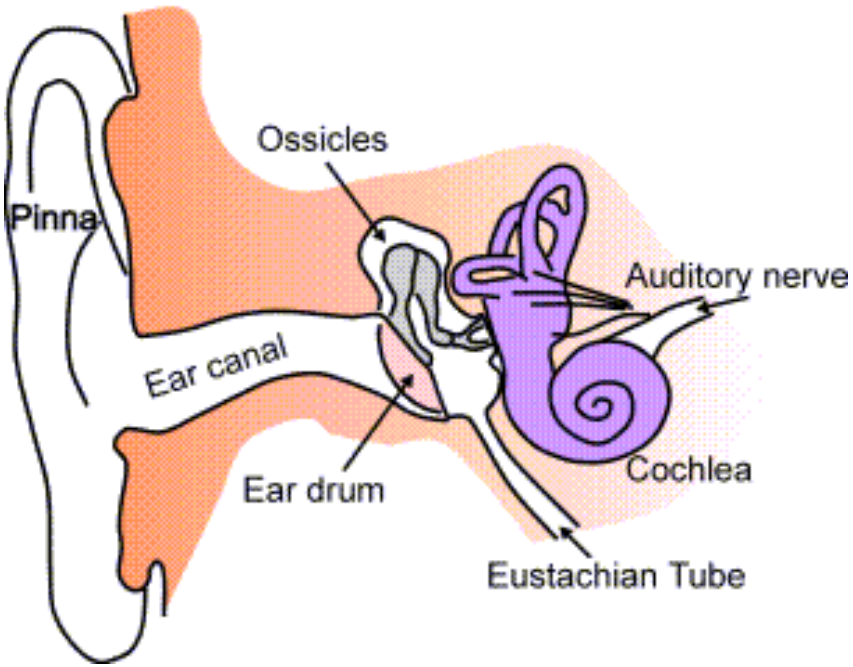
Vowels are mainly classified by the tongue-hump position (front to back) and the jaw position (low to high).

As shown below, these distinctions can be represented by the so-called **vowel quadrilateral**.

Diphthongs can also be shown on this quadrilateral in the form of transitions from one vowel position to another.



The ear

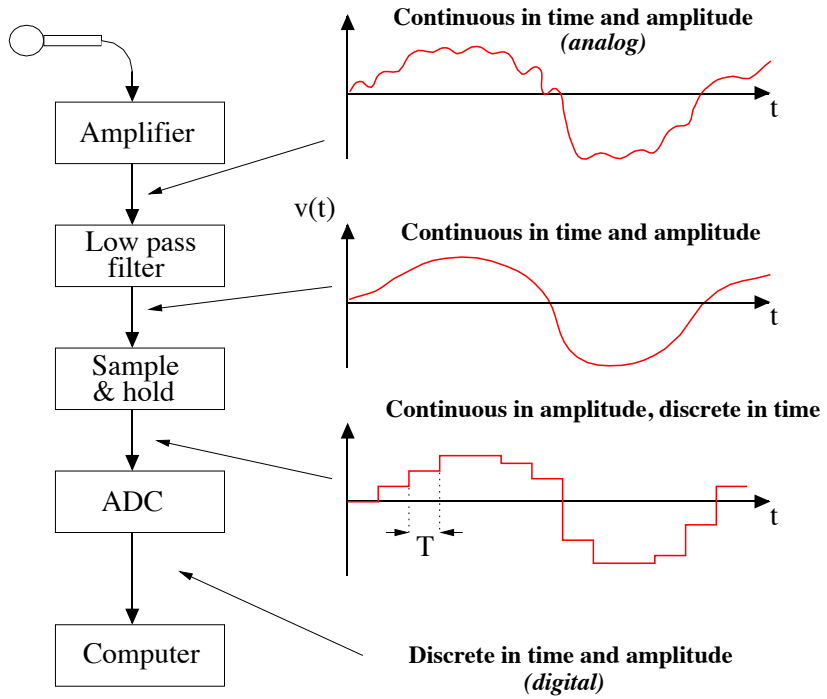


- Outer ear channels sound to middle ear
- Middle ear transmits vibrations to inner ear
- Cochlea is a concentric spiral and narrows along it's length, and contains the organ of Corti
- The organ of Corti is lined with hair cells which respond to certain frequencies
- Respond to high frequencies near the entrance and low frequencies towards the centre of the cochlea
- Nerve response characterises the signal in terms of frequencies and are transmitted to the brain
- Preceptual sensitivity to frequency difference is non-linear (greater resolution at lower frequencies)

Digital Signals

In order to process speech signals in a computer, must convert from **analog** "continuous time" form to a digital signal.

- low-pass filter signal (remove any frequencies greater than half the sampling frequency: **Nyquist's theorem**)
- sampling: measure instantaneous signal value at equally spaced points $v(nT)$: voltage measured at integer multiples of sampling period T: yields a discretisation in time.
- With ideal sampling of a band-limited signal (at greater than Nyquist rate) can convert back to analog signal exactly



Typical sampling frequencies for speech are

- 16kHz for high quality speech with 16 bits of resolution per sample.
- Telephone channels typically use 8kHz sampling & only 8 bits of precision (non-uniform quantisation).

ADCs result in an output number (to represent the voltage) in the range

$$-2^{Q-1} \ldots 2^{Q-1} - 1$$

where Q is the number of bits (e.g. 16). Note: assumes a linear quantiser (unlike standard land-line telephone signals ...)

Speech Analysis & Quasi-Stationarity

Speech is a complex **mixture** of periodic, aperiodic and stochastic signals and thus is non-stationary by nature.

In practice this causes difficulty in analysing and modelling of speech signals. Thus, in practice we normally assume that speech is stationary over a **stationary over a short interval**.

The Interval over which stationarity is assumed is normally about 10-20 ms

- If the interval is too short then there is insufficient time to accurately determine signal properties
- If the interval is too long the speech properties vary significantly so approximation is not valid

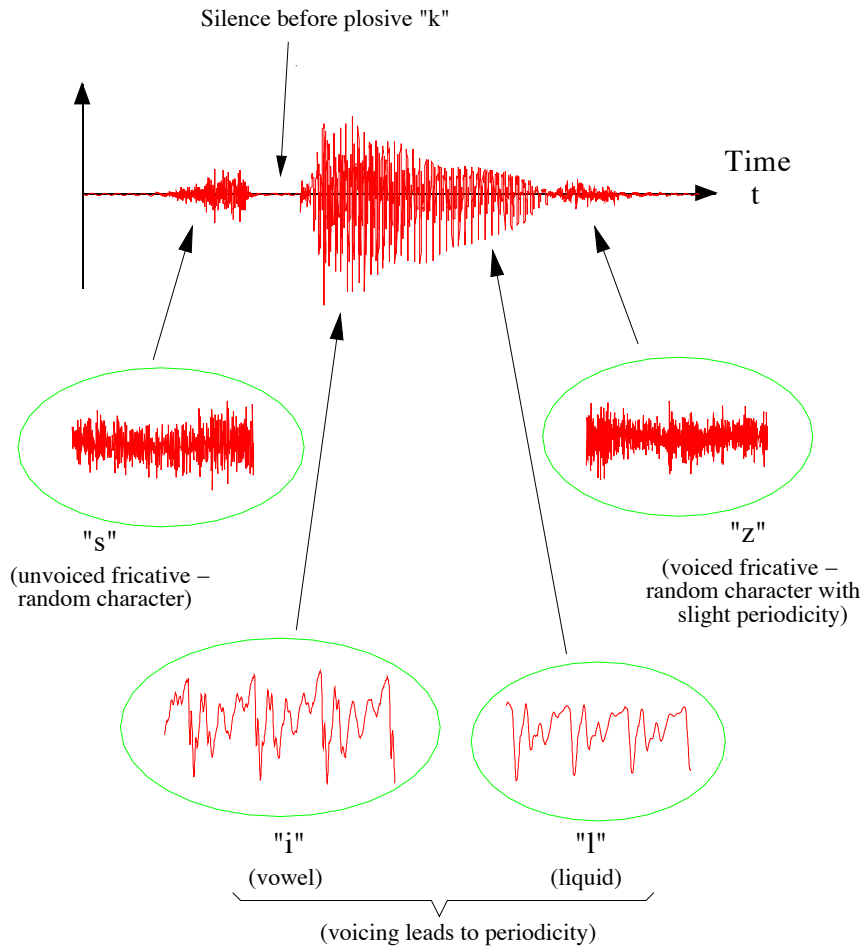
This is the quasi-stationarity assumption used in nearly all speech signal processing

- Despite the fact that this is a considerable approximation.

One important goal of speech analysis can be stated as finding the signal characteristics in each quasi-stationary interval

The word “skills”

- can see different types of speech sounds in time domain
- vowel sounds are often close to periodic (periodic signals repeat exactly with period τ)
- stochastic signals in fricatives
- aperiodic / stochastic signal in stop-consonant



What is Speech Recognition?

Speech is made up changing time-signal, and **Automatic Speech Recognition (ASR)** maps to a corresponding symbol sequence (normally a **string of words)**

Since use machine-learning, need to use a **sequence-to-sequence model**.

Note that this is not **understanding** the input i.e. explicit identification of utterance meaning.

Speech recognition is difficult because of

- differences between speakers **inter-speaker** variability;
  - physiological, accent/dialect etc
- differences in how a speaker utters the same word/sentence **intra-speaker** variability;
  - different styles of speech (prepared, spontaneous, casual, whispered etc).
  - a person's speech changes over time (ageing etc) and transitory effects (e.g. a cold)
- acoustic channel & noise - microphone differences, background noise;
- need to model language in general and of particular domain;

From the word strings other "downstream" applications may be used e.g.

1. machine-human dialogue systems
2. machine translation
3. information retrieval and information extraction

Example Applications of ASR

Numerous applications. Currently include

- **Dictation**: **desktop:** “voice typewriter” also **mobile** messaging
- **Voice dialling**: situations where hands/eyes busy elsewhere.
- **Telephony & Information Access systems**:
  - Banking, account access and management, share quotes etc.
  - Voice access to web pages/services
  - Mobile voice search
- **Transcription Systems**
  - Often specialised domains: legal, medical, voicemail etc
  - Transcription for information retrieval/extraction from audio
  - Transcription of video calls/lectures
- **Command and control**:
  - Navigation around computer windows system etc.
  - Control of many home (Amazon Echo, Google Home)/office/military items
  - Voice dialling/control of mobile phones (hands/eyes busy elsewhere)
  - Car navigation systems
  - Interaction with virtual assistants (e.g. Siri etc).

Applications have been rapidly growing as technology/compute power/access to “the Cloud” improves.

Task Complexity

Tasks have been typically constrained to limit the variability by e.g.

- isolated word format
- limited vocabulary
- constrained syntax
- low noise conditions

Hence, historically, there have been several styles of speech recognition

	input mode	vocab size	basic unit	grammar
isolated	discrete	small	word	none
continuous	continuous	medium	phone	FS / N-gram
discrete large vocabulary	discrete	large	phone	N-gram
continuous large vocabulary	continuous	large	phone	N-gram

- **input mode**: discrete (gaps between words) vs. continuous speech
- **vocabulary size**: small ($<$ 200 words) to very large ($>$ 60k words)
- **basic unit**: *acoustic model* units used (discussed later)
- **grammar**: assigns a probability to possible word sequences
  - finite state networks only allow paths through an explicit word network;
  - N-gram grammars (or other models) allows all word sequences with non-zero probability,

Current speech recognition systems are nearly all continuous speech and normally (fairly) large vocabulary. Aim to **reduce task constraints**.

Also operating condition affects performance and task complexity.

- **Speaker mode**: speaker dependent (SD) vs speaker independent (SI) vs speaker adaptive (SA)
- **Microphone**: close vs far-field; fixed vs variable; bandwidth; number of microphones
- **Background noise**: clean vs noisy (& noise type)
- **Channel**: high quality/std telephone/cellular
- **Speaking style**: prepared/read/careful vs spontaneous/casual

Research tends to focus on making recognition systems **more general** ("all purpose"): Large vocabulary speaker independent continuous speech recognition systems trained on data from a variety of different sources.

There are also a number of other factors:

- **Languages** Historically most work in English. Systems exist in most European languages, Mandarin, Japanese etc., multi-lingual,
- **System Input** Might not just be speech but multi-modal interfaces of various types, touch screens, lip-reading etc.
- **System output** Multiple hypotheses, confidence scores (estimated probability of correctness), rich transcripts, human readable ...

All modern ASR systems are based on **machine learning** and **statistical pattern recognition**.

System Aspects: What makes a "good" ASR system ?

- **Low error rate** Performance of speech recognisers can be measured by comparison of the output string with a manually transcribed version (reference transcript).
- **User satisfaction** Recognisers form part of a larger system (e.g. a text input system or an enquiry system). Users (customers!) are interested in overall system performance (e.g overall transaction time, **latency**) rather than the raw error rate.

For many applications it is necessary that the recogniser must be able to support these activities.

- alternative recognition hypothesis strings
- rejection of utterance / confidence scoring (how likely is output correct)
- vocabulary extension / spelling
- etc.

Speech recognition systems have a history of reporting worse performance in field with real users than in lab (different noise conditions, different user behaviour etc.).

Recogniser robustness is very important.

Hence it is very important to collect realistic databases for system development/test & perhaps **train on field data**.

Measuring Isolated Word Recognition Performance

Recognition results give recognition word/phone accuracy/error rates on **independent test sets**.

For isolated word recognition simply count the number of words correctly classified and express as a percentage of the total number tested.

$$\%\text{Correct} = 100 \times \frac{N\,(\text{words correctly classified})}{N\,(\text{words in the test set})}$$

Consequently the **word error rate (WER)** is measured as

$$\%\text{Word Error Rate} = 100 - \%\text{Correct}$$

In the following example the error rate is 60%:

REFERENCE	This	is	the	reference	sentence
RECOGNISER OUTPUT	This	recogniser	produced	a	sentence

For the isolated word recognition we assume that the number of words in a sentence are known, i.e. automatically detectable by inter-word silence.

Measuring WER for Continuous Speech Recognition

In the general continuous speech case, there may also be deletion and insertion errors.

The number of words in the reference will in general be different to the number of words in the recogniser output $\Rightarrow$ **word alignment** is required.

REFERENCE	One	day	we	can	talk	to		all	machines
OUTPUT	One	way	we		walk	to	the	wall	
	CORR	SUB	CORR	DEL	SUB	CORR	INS	SUB	DEL

The errors types are substitutions (SUB), insertions (INS), and deletions (DEL).

$$\%\text{Word error rate} = 100 \times \frac{N(\text{INS}) + N(\text{DEL}) + N(\text{SUB})}{N(\text{words in the reference})} = 100 - \%\text{Accuracy}$$

Theoretically the word error rate can be higher than 100% (e.g. one word in the reference could yield one substitution + one insertion). This corresponds to a negative %Accuracy.

Sometimes the percentage of correct words is used. Note that this disregards insertions.

$$\%\text{Correct} = 100 \times \frac{N(\text{CORR})}{N(\text{words in the reference})}$$

Statistical Speech Recognition

In ASR need to find the most likely word sequence $\hat{W}$ from an utterance i.e. $\max_W P(\mathbf{W}|O)$ and will need some type of **sequence-to-sequence model**.

Could try and model the speech signal and directly output words (and these **end-to-end trainable systems** are a recent technology direction discussed later).

Still many deployed systems use a **generative model approach** and find

$$\hat{W} = \arg\max_W \{p(\mathbf{O}|W)P(W)\}$$

Note that $\mathbf{O}$ is a (stream of) feature vectors describing the utterance. Hence, the essential parts of an ASR system are

- **acoustic model** giving $p(\mathbf{O}|W)$
- **language model** yielding $P(W)$

Need to design and estimate model parameters **training**, and use models to find most likely word sequence **recognition**.

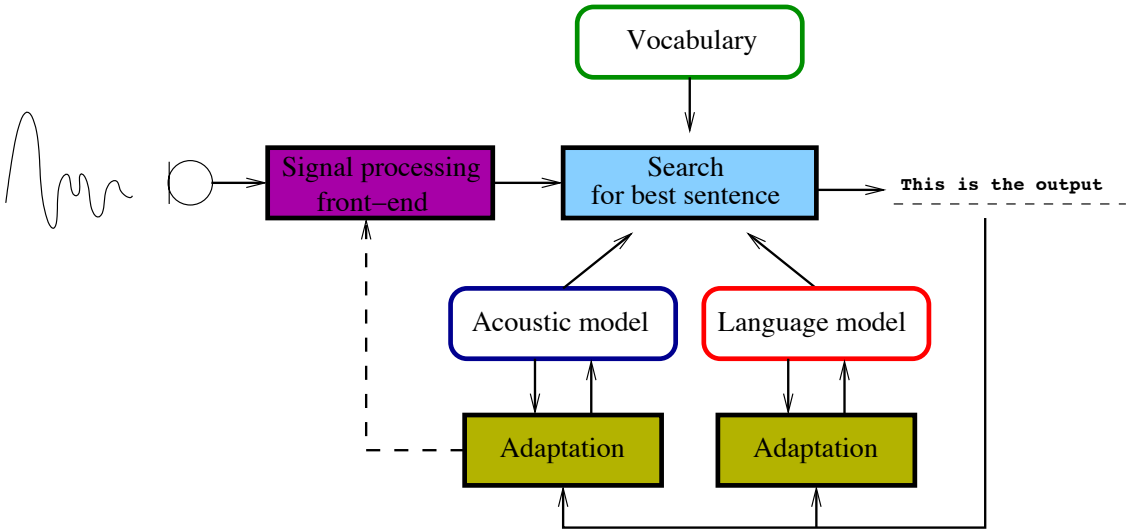
The most common form for the acoustic model is a **Hidden Markov Model (HMM)**.

We will first discuss HMMs that model acoustic vectors both using Gaussian Mixture Models (GMM-HMM) and using Deep Neural Networks (DNN-HMM) (referring to both feed-forward networks and other deep-learning architectures).

The lectures on ASR will explain HMM-based speech recognition for both isolated word recognition, extend these concepts to **large vocabulary continuous speech recognition**.

Classical Generic Recognition Architecture

- Search for most likely word or sentence given acoustic and language models
- Normally known as recognition or decoding.
- A finite set of words is defined in the vocabulary of the ASR system.



- The acoustic models usually make use of phonetic representation of words (although models individual words or letters can also be used).
- Also need to supply the pronunciations for each word
- Words not in the vocabulary (Out-Of-Vocabulary) cause errors.
- Normally an N-gram language model is used (or more advanced language model)
- ASR system output may be used to adapt (adjust) the models.

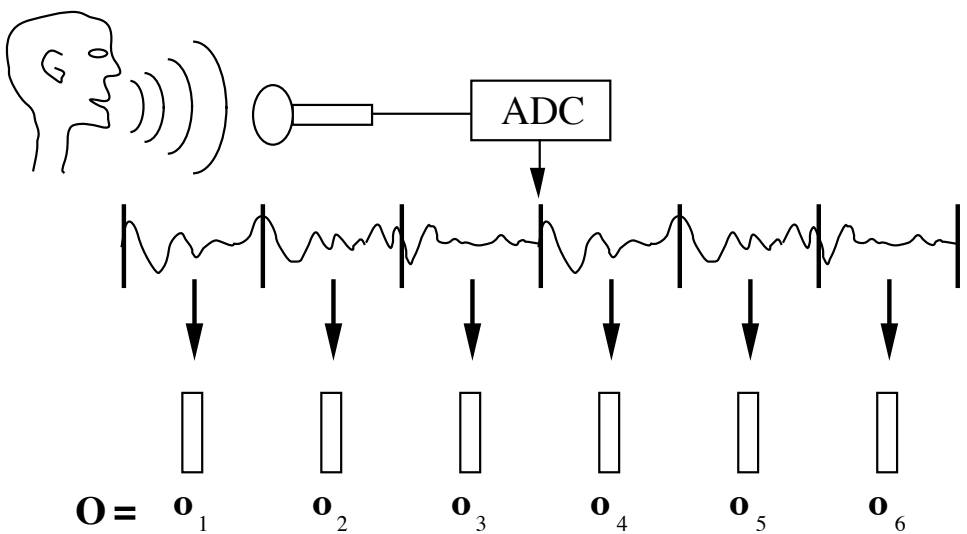
Isolated word recognition simplifies search by pre-segmentation of speech into words, but segmentation is not straightforward in anything but a low-noise audio environment.

Feature Extraction

The aim of the feature extraction is to

- transform raw-data to a form suitable for the subsequent models (if needed ... some models can take raw waveform as input)
- reduce the data rate;
- keep only information that discriminates between classes

Since speech signal is not fixed length, use **sequence** of fixed duration feature vectors.



- Normally in ASR a fixed duration of speech signal (a **frame**) is used for each feature vector
- Feature vector describes the short-term spectrum
- Spectral information can be encoded as a **cepstral representation** such as **MFCCs** (discussed in the next lecture).
- Typically have a new frame every 10-20 ms.

Summary

- Speech is produced by the human vocal apparatus: excitation source and acoustic tube
- Speech is a continuous waveform and articulators move continuously
- Types of excitation sources: voiced and unvoiced speech
- Classification of speech sounds
- The human ear (& most speech recognisers) perform frequency analysis

- Speech signals converted to digital form before processing (sampling and quantisation)
- Speech analysis assumes signal is stationary over 10-25 ms
- Time domain signal exhibits clear features of different types of excitation

- Automatic speech recognition converts speech into text
- Most modern systems aim at large vocabulary continuous speech recognition
- Speech recognition uses sequence-to-sequence models
- ”Classical” systems use generative approach: separate acoustic & language models
- Need to search given all models for best word sequence
- Generative model approach normally based on hidden Markov acoustic models
- Use a feature extraction stage that represents speech by short-term spectral information

- Many machine-learning techniques used for sequences were first developed in the context of speech recognition