Engineering Part IIB: Module 4F11
Speech and Language Processing
Lecture 1: Overview / Introduction

Bill Byrne

Lent 2014

Cambridge University Engineering Department

http://mi.eng.cam.ac.uk/∼pcw/local/4F11/index.html

Course Contents

16 Jan	1	Overview / Introduction	WJB
17 Jan		**No Lecture**	
23/24 Jan	2,3	Acoustic Analysis	PCW
30/31 Jan	4,5	ASR Introduction and Isolated Word Recognition	PCW
6 Feb	6	Sub-word Acoustic Models	PCW
7 Feb	7	Language Models	PCW
13 Feb	8	ASR Search Issues	PCW
14 Feb	-	Examples Class	PCW
20/21 Feb	9,10	Weighted Finite State Transducers for Speech and Language Processing	WJB
27 Feb	11	Introduction to Statistical Machine Translation	WJB
28 Feb	12	SMT - Alignment	WJB
6 Mar	13	SMT - Translation	WJB
7 Mar	14	Text-to-Speech Synthesis	WJB
TBD	-	Examples Class	WJB

No lecture on Friday 17 January 2014!

The course will be illustrated with speech and language processing demonstrations and examples.

Why Speech Processing?

Speech Processing aims to model and manipulate the speech signal to be able to transmit (code) speech efficiently; to be able to produce natural speech synthesis and to be able to recognise the spoken word.

Since speech is the natural form of communication between humans it reflects a lot of the variability and complexity of humans! This makes modelling speech an interesting and difficult task.

The speech signal contains information from many levels and encodes information about the speaker and the acoustic channel; the words and their pronunciation; the language syntax and semantics etc.

Speech technology is becoming increasingly well established with quite sophisticated technology now incorporated into many widely deployed applications and speech technologists are much in demand!

Why Speech and Language Processing?

(1) Speech technology - recognition, synthesis, coding - is now often only a single component within a complex information processing system.

- Engineers now study speech processing applications in real applications
- The R&D effort aims for optimum integration

(2) Modeling techniques developed for speech processing - speech recognition in particular - are applied to other language processing tasks. Statistical Machine Translation is a prime example. There is a surprising need for mathematical sophistication.

(3) Consumers have become very sophisticated in their demands for fast and easy-to-use interfaces for devices such as the iPhone. But natural interaction lags behind other aspects of interface design.

(4) Language is fun !

Some Speech and Language Processing Applications

Human-Machine Communication

desktop dictation, telephony services, found-speech transcription/indexing

Siri: `http://www.apple.com/ios/siri/`

Google for iPhone: `http://www.google.com/mobile/iphone/`

Google Voice: `http://googleblog.blogspot.co.uk/2009/03/here-comes-google-voice.html`

Machine-Human Communication

output from information systems

Toshiba / Cambridge Talking Head:

`http://www.bbc.co.uk/news/technology-21827924`

`http://youtu.be/kOil2HSDq0E`

Human-Human Communication

speech coding (reduction in bit-rate/storage);

speech enhancement (removal of noise);

voice transformation / voice morphing

personalized speech translation

aids for disabled

Video Search

Most video search is currently based on metadata. The contents of the video is *not* indexed. Searching relies on text accompanying the video when it is posted on the web, or text that appears on the web page on from the video is linked.

Ideally it should be possible to search for videos based on automatically generated transcripts of the speech in the video. This remains a research problem, although many projects focus on it.

Automatic captions in YouTube

`http://www.youtube.com/watch?v=qzXLM-Kx3kI`

(turn on automatic captions)

Voice Morphing

Voice Morphing, or voice transformation or voice conversion, is a technique to modify a source speaker's speech utterance to sound as if it was spoken by a target speaker.

– Speech is transformed from one person's voice to another person's voice
– Example: Female to Male conversion

Source: `http://mi.eng.cam.ac.uk/~wjb31/local/4F11/src01.wav`

Target: `http://mi.eng.cam.ac.uk/~wjb31/local/4F11/tgt01.wav`

Morphed: `http://mi.eng.cam.ac.uk/~wjb31/local/4F11/vm01.wav`

Adaptive Model-Based Synthesis

Statistical models which can generate speech in a variety of modes:

- speaker neutral
- speaker dependent / speaker adapted

These systems can 'read aloud' directly from text.

Google Map - UEDIN

`http://homepages.inf.ed.ac.uk/jyamagis/Demo-html/map-new.html`

Synthesizing speech in a particular voice makes it possible to consider mapping voices from language-to-language as well from speaker-to-speaker

EMIME: Personalized Speech-to-Speech Translation

English Speech Recognition $\rightarrow$ English-to-Japanese Translation $\rightarrow$ Personalized Japanese Speech Synthesis

Statistical Machine Translation

Techniques and approaches have been borrowed from speech recognition and applied to Statistical Machine Translation (SMT). Important aspects of the SMT problem draw heavily on engineering techniques. Approaches are very computational and rely on algorithms based on statistical models of how translation maps sentences from a source language to a target language.

Modern SMT is (nearly) entirely statistical and relies on large amounts of translations and monolingual text.

Interactive, multi-engine translation:
`http://labs.reverso.net/default.aspx?lang=EN#`

FAUST project: `http://faust-fp7.eu/faust/`

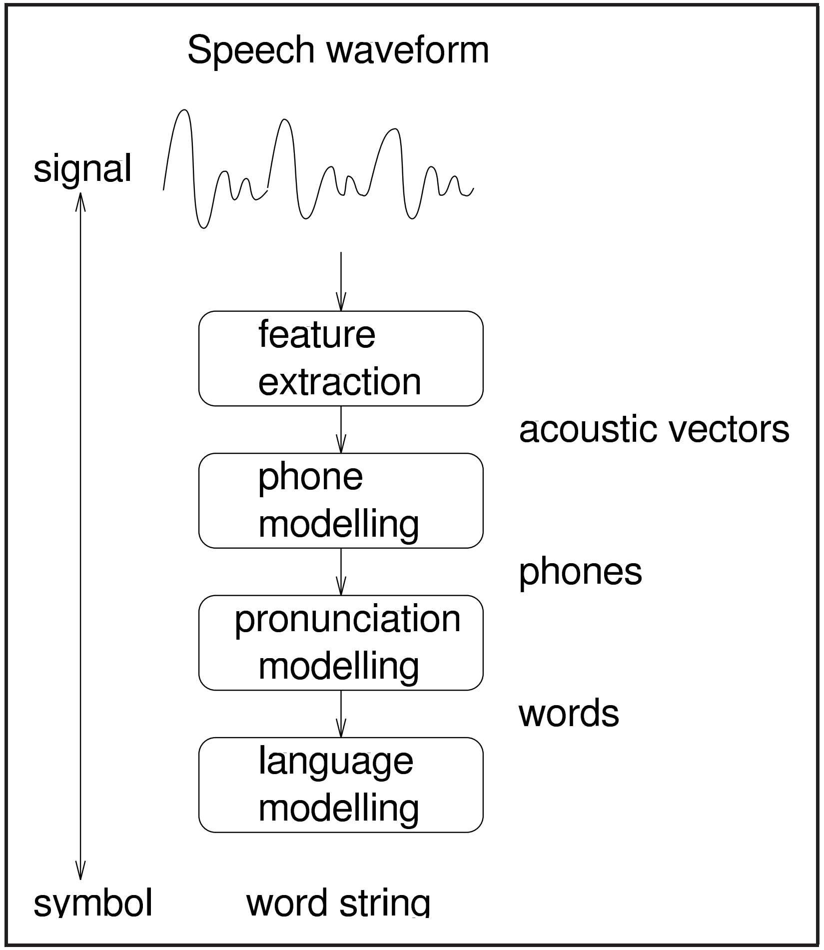
CUED Interactive Speech to Speech Translation

`http://mi.eng.cam.ac.uk/~wjb31/local/4F11/FinalVideo2.mov`

Oliver Lambert, 4^{th} year project, 2011/2012

Levels of Speech Processing

Much of speech processing concerns converting between the different levels of representation:



The Speech Waveform

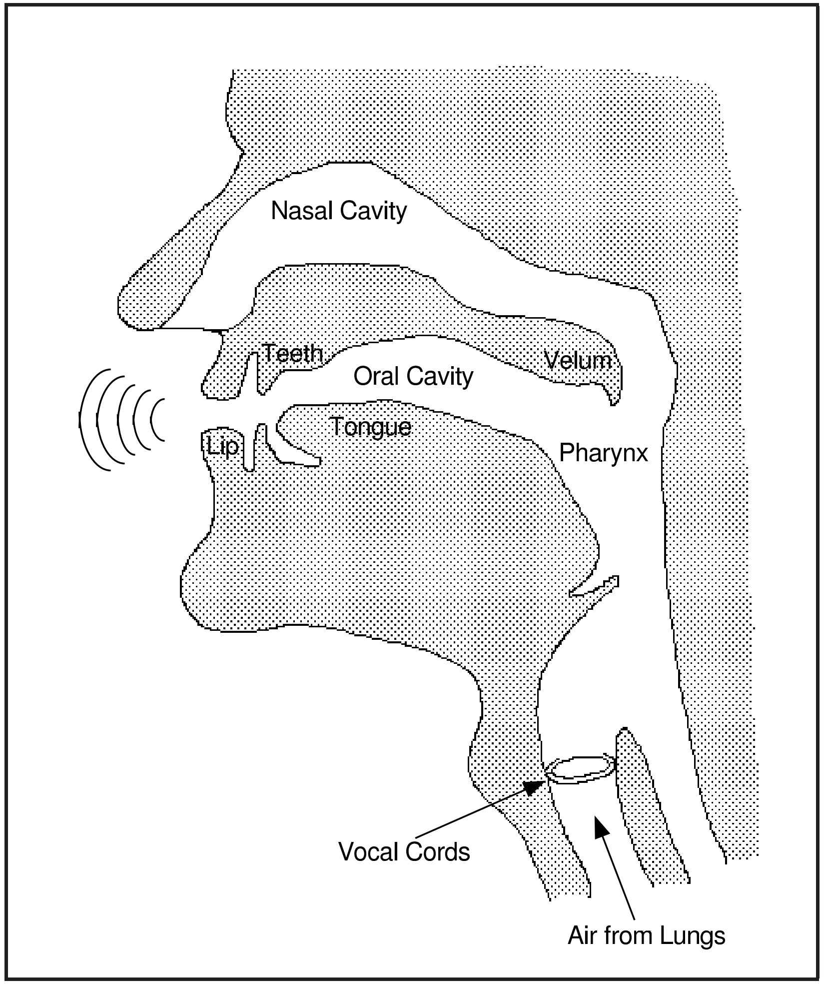
The speech signal is non-stationary and it contains a mix of pseudo-periodic and random components. Different classes of speech sounds have properties that depend on how they were produced.

Analysis of the speech production mechanisms will allow us to formulate a simple model of the speech production system which we will use in speech analysis.

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 in lecture 2 we will look at the speech signal in the frequency domain also.

Human Vocal Tract (Cross-section)



The Acoustic Tube

The oral cavity and pharynx form an acoustic tube.

- The shape is varied by moving the three main articulators: lips, tongue and jaw. Changing the shape of this tube changes its transmission properties and hence modifies the spectrum of the emitted speech pressure wave.

- There is a secondary branch through the nasal cavity which is enabled by opening the velum. This gives rise to nasals such as in “<u>m</u>ake”, “ru<u>n</u>”, etc.

The articulators are continually moving as each sound is produced. Often the target vocal tract state for each sound is never reached and the current sound is further modified by anticipation of the following sounds. The exact realisation of each individual sound is heavily dependent on previous and succeeding sounds. These co-articulation effects are an important source of variability in the speech signal.

Demonstrations

Side-view of the vocal tract

`http://mi.eng.cam.ac.uk/~wjb31/local/4F11/ArticulatoryPixs/1-midsagittal.gif`

Movie of the phrase ‘spot chick’ (no audio)

`http://mi.eng.cam.ac.uk/~wjb31/local/4F11/ArticulatoryPixs/2-spot_chick.qt`

Movie of the sentence ‘Try not to annoy her.’

`http://mi.eng.cam.ac.uk/~wjb31/local/4F11/ArticulatoryPixs/3-try_not.mov`

Labelled anatomy of the vocal folds

`http://mi.eng.cam.ac.uk/~wjb31/local/4F11/ArticulatoryPixs/5-Gray956.png`

Schematic operation of vocal folds

`http://mi.eng.cam.ac.uk/~wjb31/local/4F11/ArticulatoryPixs/6-vocalfoldoperation.pdf`
(from Taylor)

Stroboscopic video of vocal fold operation

`http://mi.eng.cam.ac.uk/~wjb31/local/4F11/ArticulatoryPixs/8-slowcords.gif`

MRI of the moving vocal tract (video clip linked to each still image)

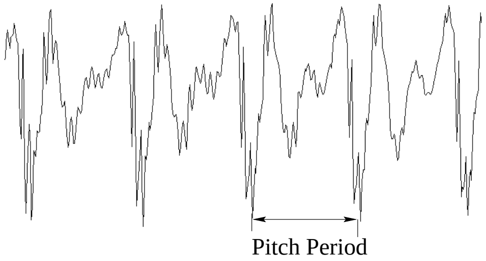
`http://www.phon.ox.ac.uk/mri`

Beat Boxing

`http://www.wired.com/underwire/2013/01/beatboxing-mri-study/`

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

For most practical engineering applications, 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. For example, “This is speech” consists of 9 phonemes in sequence

```
th ih s ih z s p iy ch
```

Notice that there is no explicit identification of word boundaries in continuous speech. This is one of the factors which makes speech recognition difficult.

Some sounds, particularly vowels, form a continuum and hence various choices of phone set are possible. However, for English around 40 are typically used. Table 1 lists a set commonly used for American English called ARPAbet.

Consonant sounds may be divided into 5 broad classes depending on the type of vocal tract constriction: plosives (stops), fricatives, affricates, liquids (semi-vowels) and nasals. Within each class the individual sounds are distinguished by the place at which the constriction occurs and whether or not there is voicing. Table 2 shows how each of the consonants in the ARPAbet are classified.

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

Table 1: The ARPAbet American English Phone Set

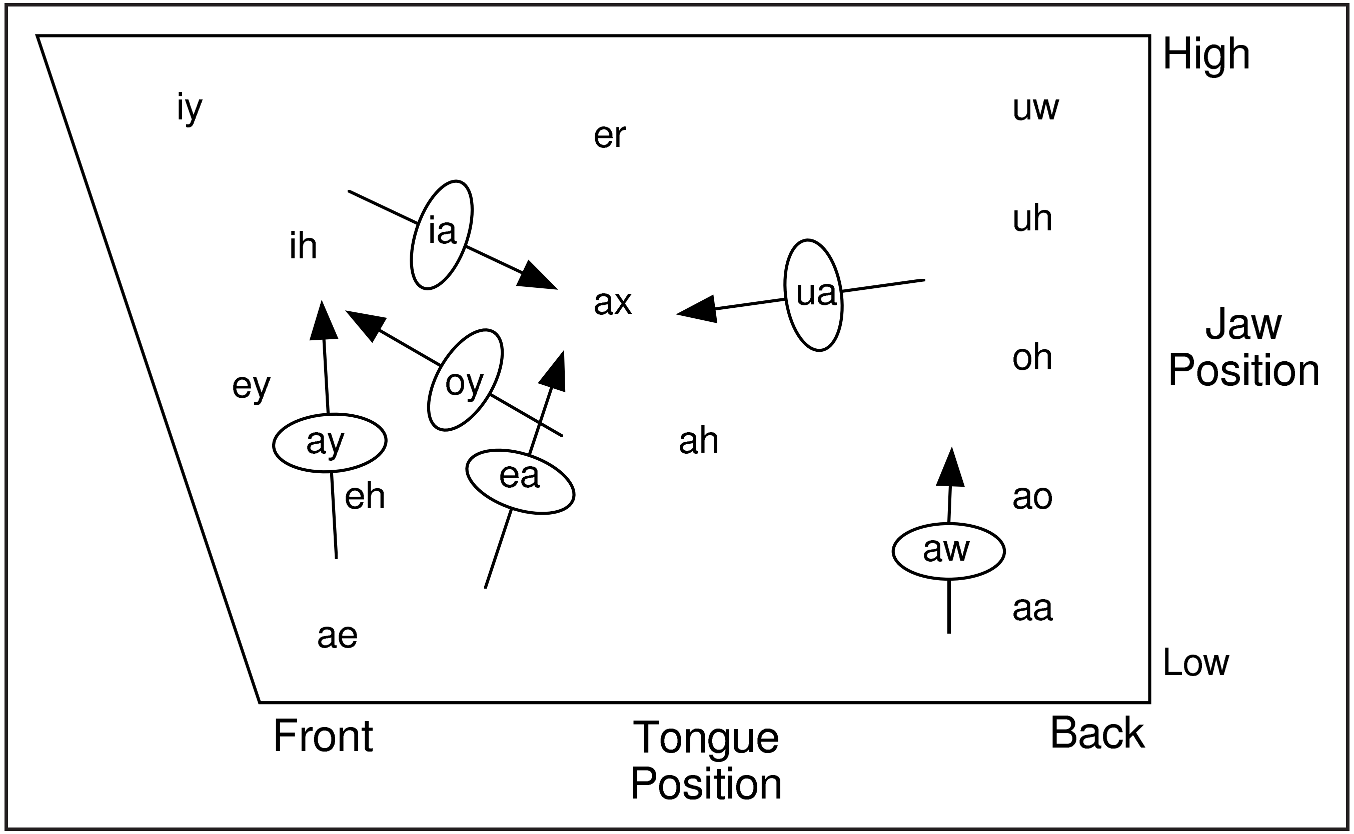
Consonant Classification

<table>
<tr><th></th><th>lip-lip</th><th>lip-teeth</th><th>teeth-tongue</th><th>alveolar-tongue</th><th>palate-tongue</th><th>velum-tongue</th></tr>
<tr><td>nasal</td><td>m</td><td></td><td></td><td>n</td><td></td><td>ng</td></tr>
<tr><td>stop</td><td>p b</td><td></td><td></td><td>t d</td><td></td><td>k g</td></tr>
<tr><td>fricative</td><td></td><td>f v</td><td>th dh</td><td>s z</td><td>sh zh</td><td></td></tr>
<tr><td>liquid</td><td>w</td><td></td><td></td><td>r l</td><td>y</td><td></td></tr>
<tr><td>affricate</td><td></td><td></td><td></td><td colspan="2">ch jh</td><td></td></tr>
</table>

Table 2: Consonant Classification

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 Source-Filter Model

The human vocal tract is a complex time-varying non-linear filter which is excited by a number of different energy sources. Realistic model of the acoustic properties are immensely complex.

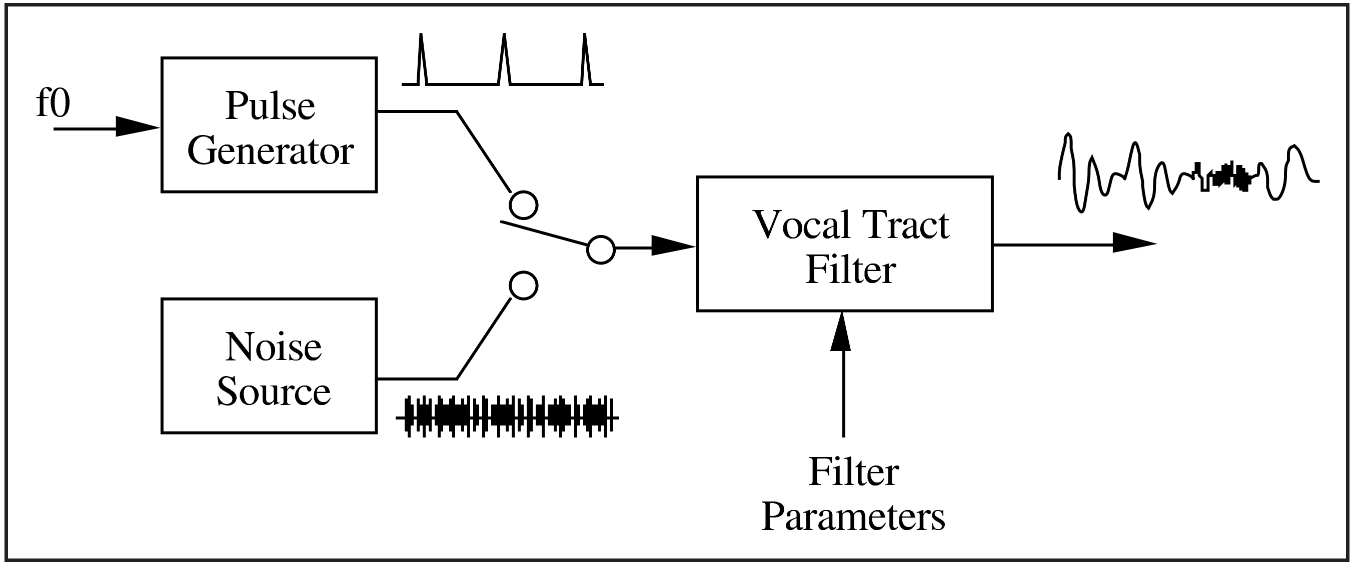
The use of models in engineering applications requires low complexity and computational cost. In order to produce usable models a number of simplifying assumptions can be made:

1. The vocal tract can be represented as a single lossless linear time variant filter with a single input.

2. The excitation is either a periodic pulse train or noise, depending on basic sound classes.

3. The filter and excitation characteristics are stationary over periods of the order of 10 msecs.

These assumptions lead to the *source-filter model* of speech production as illustrated next.

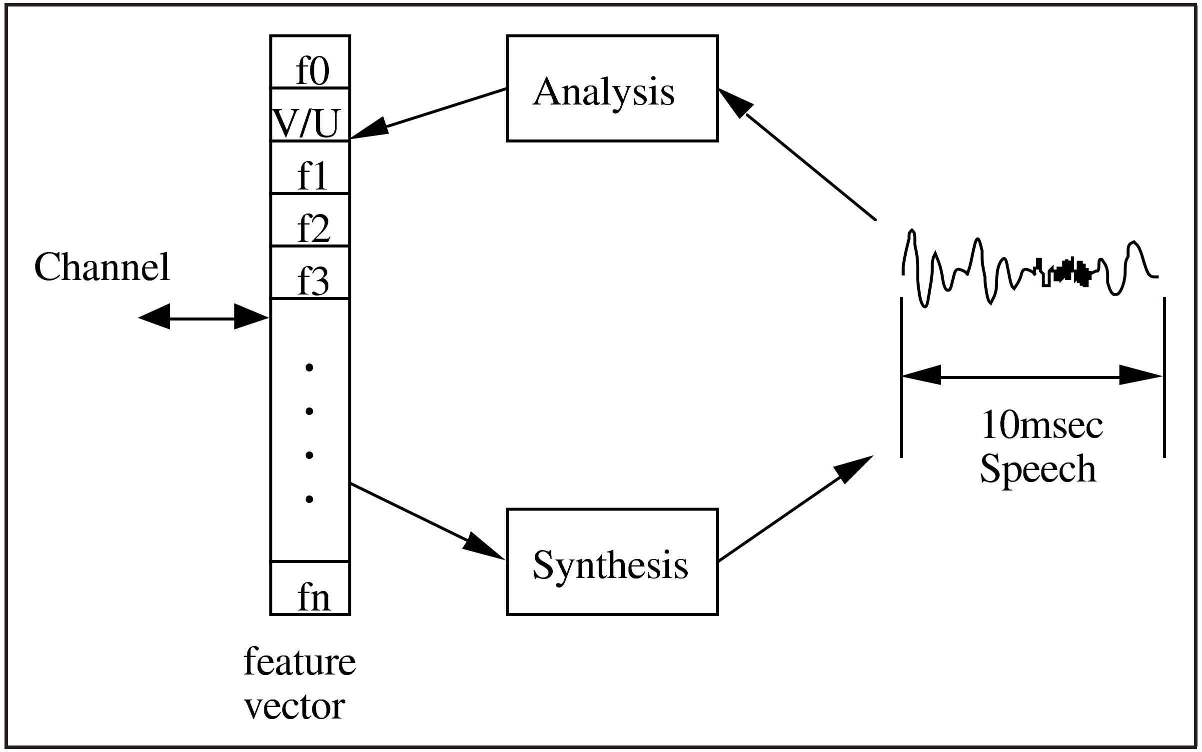


This model is widely used for both the analysis and the synthesis of speech. When used for synthesis, the filter parameters are updated every 10 msecs (or so). For voiced sounds, f0 is set equal to the pitch frequency.

When used for analysis, the speech is divided into segments of typically 10-25 msec called frames. For each frame, the set of filter parameters are determined which minimise the difference between the speech which would be generated by the model and the actual speech.

Source-Filter Model Applications

The source-filter model represents the speech signal as a stream of parameters.



The primary applications of speech processing are synthesis, coding and recognition, and the source-filter model provides the basis for all of these.

Analysis

Each frame of speech is analysed using the source-filter model and the corresponding parameters are stored as an acoustic feature vector. For recognition, sequences of feature vectors are compared with stored patterns in order to identify the spoken sounds or words.

Synthesis

Acoustic feature vectors are either pre-stored on disk or they are generated automatically from text. Each acoustic vector is then converted to a waveform to generate the required speech.

Coding

This combines analysis and synthesis. Speech is analysed into acoustic feature vectors, transmitted down a channel and re-synthesised at the other end

This course will deal with Analysis and Synthesis.

Course Contents

16 Jan	1	Overview / Introduction	WJB
17 Jan		**No Lecture**	
23/24 Jan	2,3	Acoustic Analysis	PCW
30/31 Jan	4,5	ASR Introduction and Isolated Word Recognition	PCW
6 Feb	6	Sub-word Acoustic Models	PCW
7 Feb	7	Language Models	PCW
13 Feb	8	ASR Search Issues	PCW
14 Feb	-	Examples Class	PCW
20/21 Feb	9,10	Weighted Finite State Transducers for Speech and Language Processing	WJB
27 Feb	11	Introduction to Statistical Machine Translation	WJB
28 Feb	12	SMT - Alignment	WJB
6 Mar	13	SMT - Translation	WJB
7 Mar	14	Text-to-Speech Synthesis	WJB
TBD	-	Examples Class	WJB

No lecture on Friday 17 January 2014!

W.J. Byrne
P.C. Woodland