

Engineering Part IIB: Module 4F11

Speech and Language Processing

Lecture 1: Overview / Introduction

Bill Byrne

Lent 2012



Cambridge University Engineering Department

<http://mi.eng.cam.ac.uk/~pcw/local/4F11/index.html>

Course Contents

| | | | |
|------------|------|----------------------------------------------------------------------|-----|
| 19 Jan | 1 | Overview / Introduction | WJB |
| 26/27 Jan | 2,3 | Acoustic Analysis | PCW |
| 2/3 Feb | 4,5 | ASR Introduction and Isolated Word Recognition | PCW |
| 9 Feb | 6 | Sub-word Acoustic Models | PCW |
| 10 Feb | 7 | Language Models | PCW |
| 16 Feb | 8 | ASR Search Issues | PCW |
| 17 Feb | - | Examples Class | PCW |
| 23/24 Feb | 9,10 | Weighted Finite State Transducers for Speech and Language Processing | WJB |
| 1 Mar | 11 | Introduction to Statistical Machine Translation | WJB |
| 2 Mar | 12 | SMT - Alignment | WJB |
| 8 Mar | 13 | SMT - Translation | WJB |
| 9 Mar | 14 | Text-to-Speech Synthesis | WJB |
| TBD | - | Examples Class | WJB |

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

consumer products – voice control and interaction

Google Mobile App for the iPhone – www.google.com/mobile/apple/app.html

[found-speech transcription/indexing](#)

Machine-Human Communication

[output from information systems](#)

intelligent assistant etc.

proof-reading machines

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



Voice Morphing¹

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

¹<http://svr-www.eng.cam.ac.uk/~hy216/VoiceMorphingPrj>



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.

Examples²:

John McCain: speaker-dependent models **trained** on fairly large amounts of speech

George Bush: speaker-neutral models **adapted** on a fairly small amount of speech

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 → English-to-Japanese Translation →
Personalized Japanese Speech Synthesis

²Examples produced by Junichi Yamagishi for EMIME project – emime.org



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.



Statistical Machine Translation

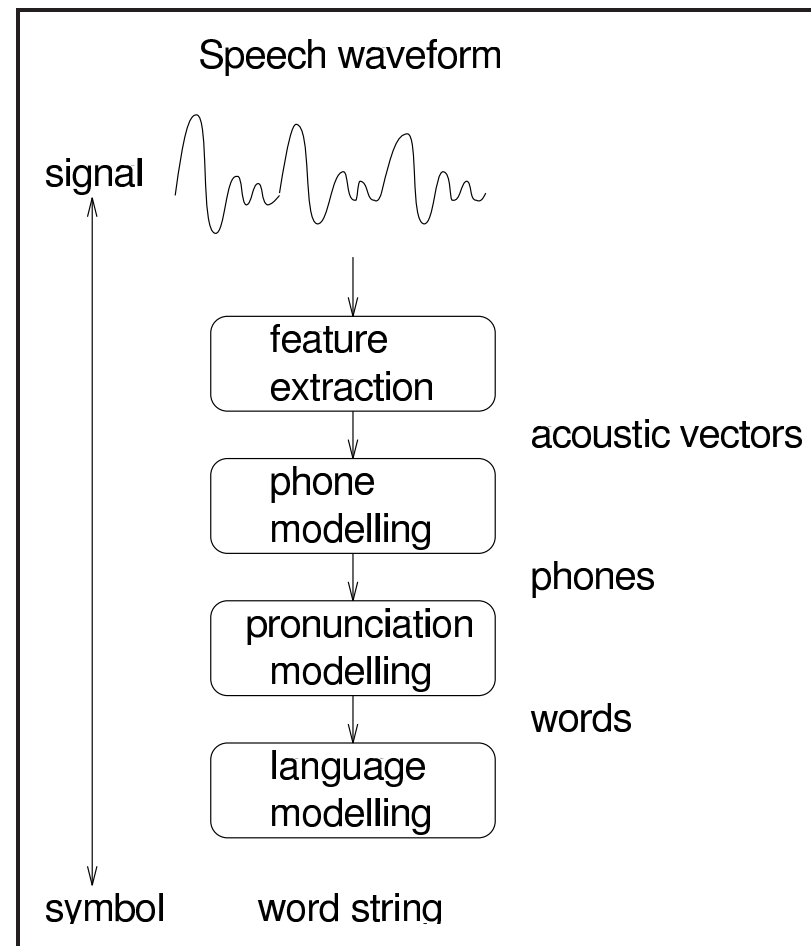
Techniques and approaches have been borrowed from speech recognition and applied to Statistical Machine Translation (SMT). Important aspects of the SMT problem draw heavily on engineering techniques. Approaches are very computational and rely on algorithms based on statistical models of how translation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

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



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.

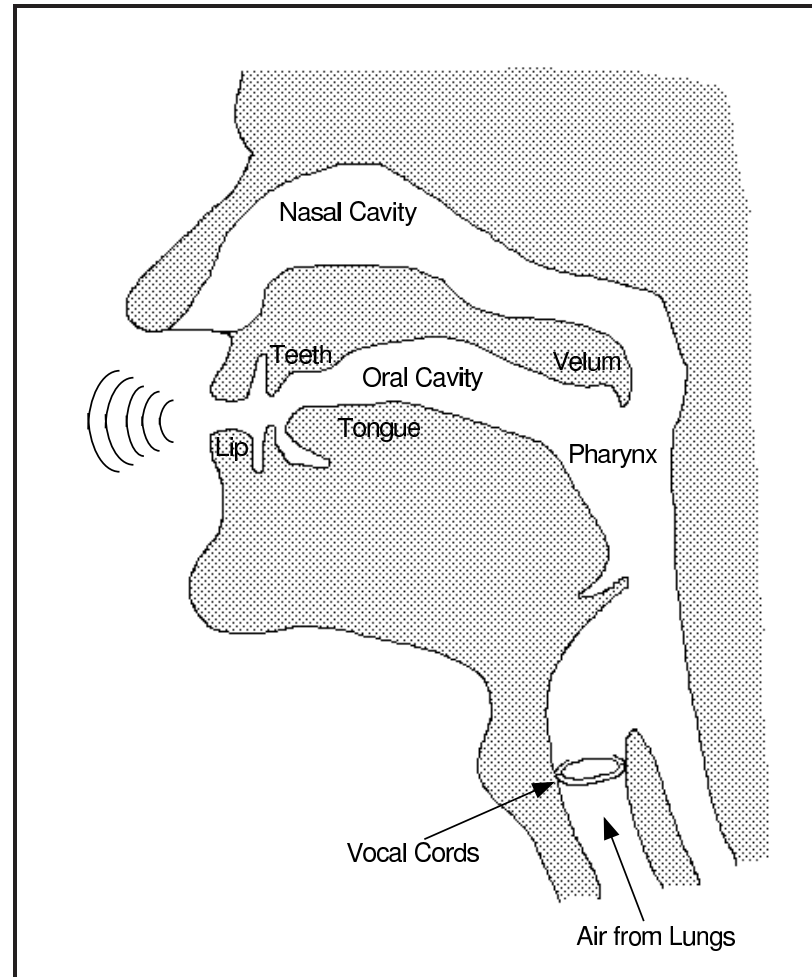
Analysis of the speech production mechanisms will allow us to formulate a simple **model** of the speech production system which we will use in speech analysis.

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.

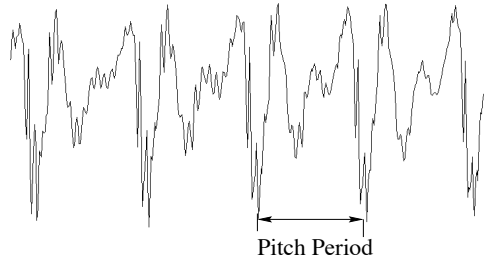
- The shape is varied by moving the three main **articulators**: lips, tongue and jaw. Changing the shape of this tube changes its transmission properties and hence modifies the spectrum of the emitted speech pressure wave.
- There is a secondary branch through the nasal cavity which is enabled by opening the velum. This gives rise to nasals such as in “make”, “run”, 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.



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

Table 1: The ARPAbet American English Phone Set



Consonant Classification

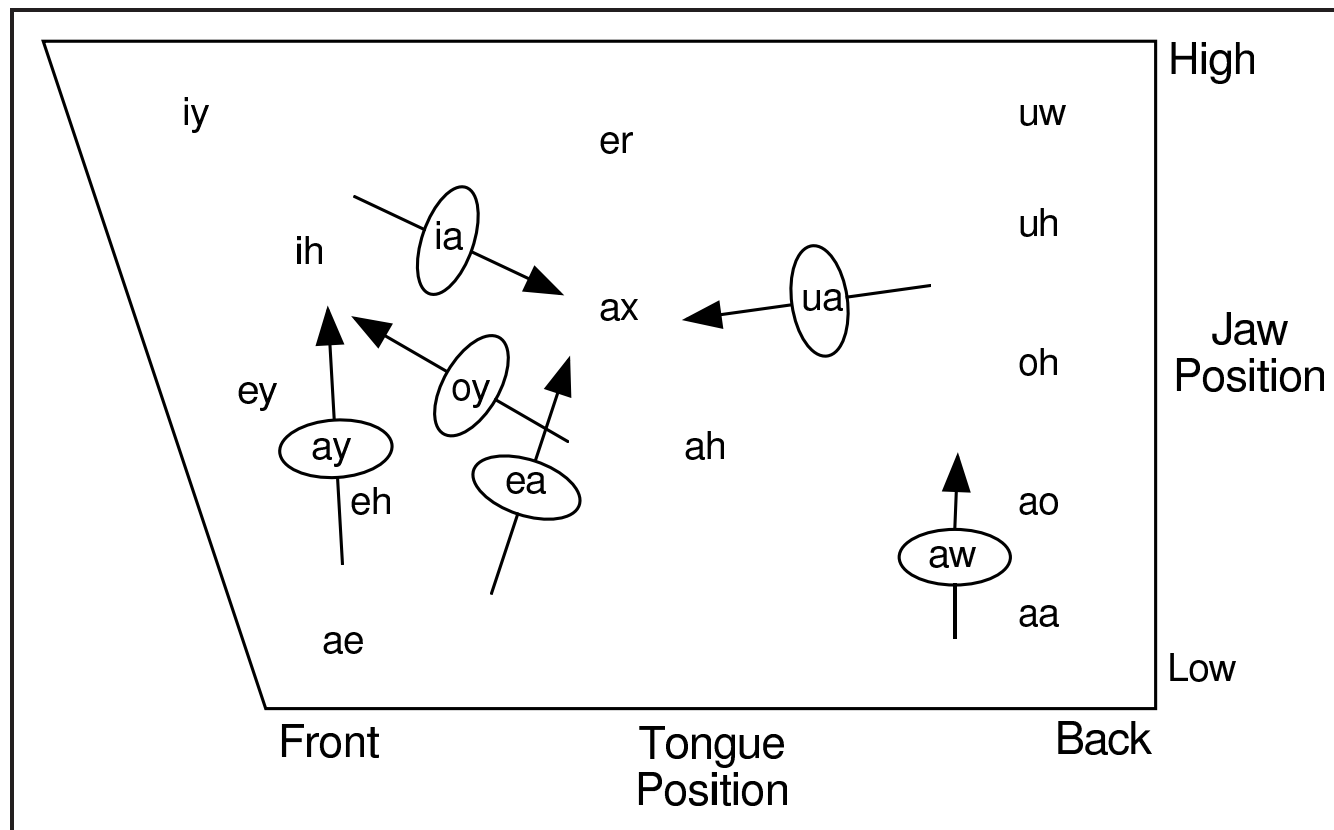
| | lip- lip | lip- teeth | teeth- tongue | alveolar- tongue | palate- tongue | velum- tongue |
|-----------|-------------|---------------|------------------|---------------------|-------------------|------------------|
| nasal | m | | | n | | ng |
| stop | p b | | | t d | | k g |
| fricative | | f v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

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

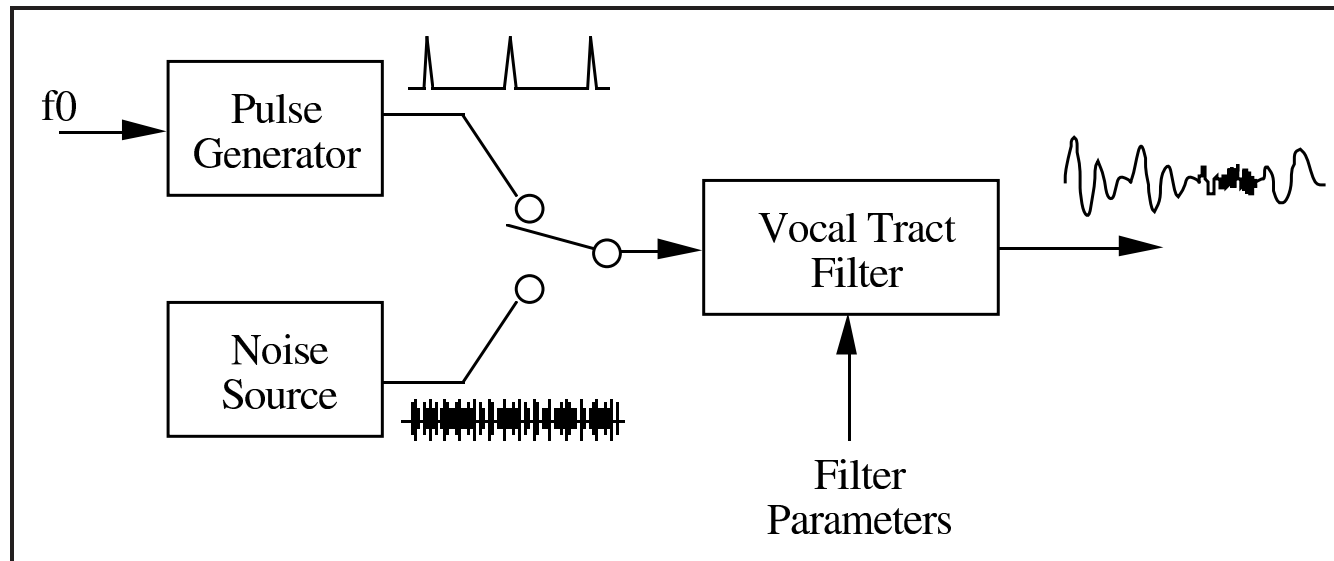
The human vocal tract is a complex time-varying non-linear filter which is excited by a number of different energy sources. Realistic model of the acoustic properties are immensely complex.

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.



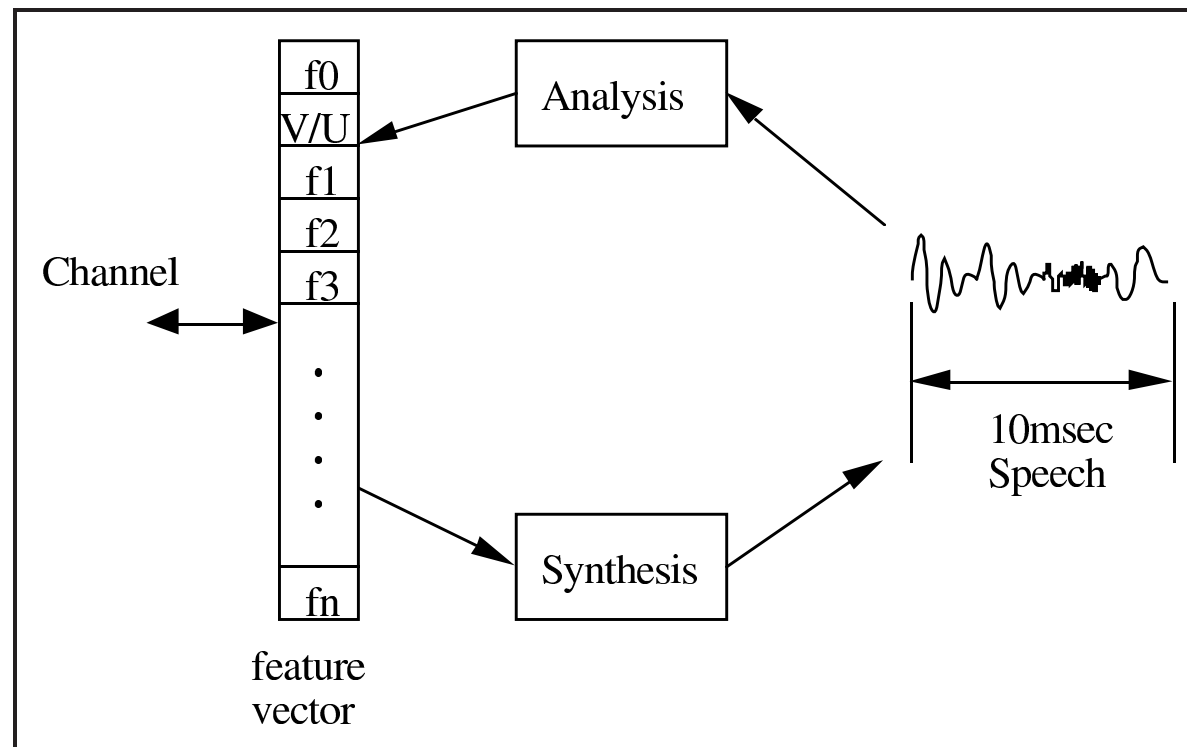


This model is widely used for both the analysis and the synthesis of speech. When used for synthesis, the filter parameters are updated every 10 msecs (or so). For voiced sounds, f_0 is set equal to the pitch frequency.

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.

W.J. Byrne
P.C. Woodland

