# Speech Perception

How do we extract meaning from speech?

Early models assumed a 'bottom-up' approach, i.e.:

* Separate the stream of speech into words.
* Identify each word and determine its meaning.
* Determine the meaning of the whole utterance.

More recent models assume a 'top-down' approach, i.e.:

* Analyse the prosody to locate the key-words
* Identify the key-words and guess at the meaning of the utterance
* If unsuccessful, analyse more words until the meaning has been extracted

Note that:

* Most traditional (bottom-up) models were originally developed to explain perception of written rather than spoken language.
* Neither type of model can fully explain phenomena such as *context effects* and *frequency effects*.
* Both types of model assume that speech is held in a temporary store during processing
* Modern models fit well with the concept of humans as 'lazy processors'.

## Analysing the Content of Speech

Top-down models suggest we analyse speech by searching for key-words, but what are the key words in an utterance?

* **New Information versus Given Information**
  + Each spoken phrase is normally constructed around one word.

This word is deemed by the speaker to contain *new information* not previously known to the listener or inferable from the context.

* + The remaining words either:
    - Complete the grammatical structure or...
    - ...provide redundancy by presenting information already known to the listener or inferable from the context.

Such words are said to contain *given information*.

* **Content Words versus Function Words**
  + Some of the words within a phrase are essential, for example, verbs and nouns. The phrase cannot be understood if they are omitted or mis-heard.

Such words are known as *content words*.

* + Other words are not essential, for example, words like *a*, *the*, and *of*. The phrase can be understood even if they are omitted or mis-heard.

Such words are known as [function words](http://web.archive.org/web/20100410235115/http:/www.cs.ucc.ie/%7Eianp/CS2511/HSP_notes.html#funcWords).

## The Role of Prosody

The speaker provides *prosodic cues* to help the listener parse the speech steam.

Appropriate prosody helps the listener identify key words (new information, content words).

Several features of prosody are implicated in speech processing.

* Intonation
  + The pitch rises at the start of each phrase and falls at the end
  + Where several phrases follow in succession, each has a lower average pitch than its predecessor
  + Within each phrase, a sharp change in pitch marks the position of the new information
* Rhythm
  + Stressed syllables of key words (new information, content words) are spoken at regular intervals and may also be marked by an increase in volume.
  + Pauses are placed between words where necessary to support the rhythm.
  + Longer pauses are placed between phrases.

## Speech as a Co-operative Activity

Speech is a *co-operative activity* in which each speaker:

* Assesses what is relevant to the listener(s) at this point in an exchange
* Assesses what is known to the listener(s) (given information) and what is not known (new information)
* Constructs a phrase around the new information
* Presents the phrase with appropriate prosody

Successful speech communication relies on the maintenance of a *shared context* between speaker and listener(s).

In practice this shared context often breaks down.

Therefore successful speech communication also relies on mechanisms for *clarification and repair*.

## Designing Speech-Based Interfaces

Speech-based interfaces are becoming more widely used, but:

* Speech communication is very different from the visual/tactile communication used with GUIs.
* There are few models or sets of guidelines for designers of speech interfaces or speech-based systems.

Speech can be used:

* for input (speech recognition)
* for output (speech synthesis)

Some systems use speech for both input and output (e.g., voice-mail), while others use speech only for output (e.g., SatNav), and a few use speech input in conjunction with visual output.

## Speech-Recognition

Speech recognition involves two distinct processes:

* Word recognition
* Recovery of meaning.

Most of the problems associated with word recognition have been solved.

However, recovering the meaning from a sequence of words is far more difficult.

Speech-recognition is mainly used in applications that involve either:

* speech-to-text transcription, where recovery of meaning is not necessary.
* tightly-constrained problem domains, where vocabulary can be limited and every word given a fixed meaning.

Many problems must be overcome before speech recognition can be used in 'free' environments.

Even when the individual words are correctly recognised, speech is more difficult to analyse than written language.

There are a number of reasons for this:

* Speech relies heavily on non-grammatical sentence forms ([minor sentences](http://web.archive.org/web/20100410235115/http:/www.cs.ucc.ie/%7Eianp/CS2511/HSP_notes.html#majMin))
* There is no punctuation. Parsing cues are provided through a mixture of pausing, intonation and stress in ways that are complex and not fully understood.
* Fragmentary phrases are common.
* Repetition and re-phrasing are common
* Speech makes extensive use of *anaphoric* devices, e.g., he/she/him/her/it/etc., instead of referring to a person or object by name.
* Efficient speech communication relies heavily on other communication channels - gesture, facial expression, etc..
* Ambiguous words, phrases, etc., are often used in order to achieve brevity, with the correct interpretation being signalled through prosody.
* Speech relies on a 'shared context', a body of shared assumptions and knowledge relevant to a spoken exchange which determines what can and cannot be said.

The context is dynamic, changing as the exchange progresses. Modelling this context is extremely difficult.

## Speech Synthesis

The design of speech output can be broadly separated into two stages:

* Deciding what should be said and when.
* Choosing appropriate prosody for each utterance.

Some frameworks exist to assist in the design of dialogues, but much work still needs to be done.

Considerable work has been done on creating rules to generate valid prosody, but for best results the output must be 'hand tuned'.

How good are current speech-synthesisers and speech-based systems?

Studies suggest that people usually understand synthetic speech as well as natural speech.

However, there is evidence that understanding synthetic speech requires more effort than understanding natural speech.

For example, Nusbaum & Pisoni (1985) compared comprehension of spoken tracts under four conditions:

|  |  |
| --- | --- |
| Natural speech | meaningful content |
| "         " | meaningless content |
| Synthetic speech | meaningful content |
| "           " | meaningless content |

They found that:

* Subjects understood the meaningful content regardless of whether it was presented using natural or synthetic speech.
* Word-recognition for the meaningless content was far worse when the passage was presented using synthetic speech

This suggests that semantic processing was necessary to recognise words presented using synthetic speech.

Other studies (e.g., Luce, 1982) have shown that subjects recall more of the syntax and surface structure of synthetic speech than of natural speech.

This suggests that they had to listen more carefully to extract meaning from the synthetic speech.

These techniques are widely used to measure the quality of speech synthesis.

Studies show that, as speech synthesis techniques have developed, the gap between natural and synthetic speech has narrowed, but that synthetic speech remains more difficult to understand than natural speech.