Auditory Perception

Auditory Perception

Like the visual system, the human auditory system can be divided into two stages:

- Physical reception of sounds
- Processing and interpretation

Like the visual system, the human auditory system has both strengths and weaknesses:

- Certain things cannot be heard even when present
- Processing allows sounds to be constructed from incomplete information

The principal characteristics of sound - as perceived by the listener - are:

- Pitch
- Loudness
- Timbre

Pitch

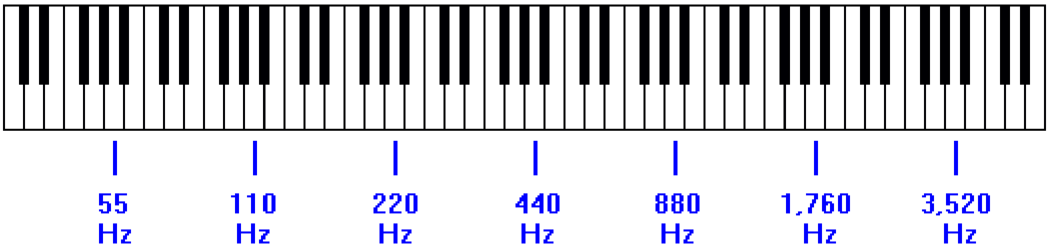
The human ear can detect changes in air pressure at rates from 25 to 16,000 cycles a second (approximately).

The faster the rate of change (i.e., the more cycles per second) the higher the *pitch* of the sound we hear.

- Most people have good *relative* pitch recognition
- Only a few people (mainly musicians) have good *absolute* pitch recognition

The relationship between the frequency with which the air-pressure varies and perceived pitch is not linear.

Each *doubling* of the frequency produces a unit increase in pitch (an *octave*).



The smallest pitch/frequency change that can be discerned (the Just Noticeable Difference or JND):

- is roughly constant across the frequency-range for pitch
- varies for frequency, with quite small changes being discernible at low frequencies, but only larger changes being discernible at higher frequencies.

Loudness

The perceived intensity of a sound depends upon:

- The sound pressure.
- The distance between the source and the listener
    - sound intensity does not decline linearly with distance.
- The duration of the sound
    - human beings are very poor at judging the loudness of sounds that are heard for less than about 0.2 seconds.
    - Such sounds usually appear much quieter than their intensity might suggest.
- Absorption and reflection of the sound by the air and by surrounding objects.
    - Different frequencies may be absorbed/reflected by different amounts.
- The frequency of the sound
    - the perceived loudness varies with frequency as well as amplitude.

Timbre

Factors determining the timbre of a sound include:

- The quantity of harmonics and their relative strengths
    - one sound might include lots of harmonics, some of which are almost as strong as the fundamental
    - another might include just the fundamental plus a few, relatively weak harmonics.
- The relationship of the harmonics to the fundamental
    - one sound might contain only *even-order* harmonics
(f_0 * 2, f_0 * 4, etc.)
    - another might include *odd-order* harmonics
(f_0 * 3, f_0 * 5, etc.).

The timbre of a sound is also determined by the way in which the amplitude varies over time.

This is known as the *amplitude envelope*. It helps us to distinguish between sounds which may have similar harmonic structures.

For example:

- Some sounds start at a high amplitude but then fade in intensity
(e.g., percussion instruments)
- Other sounds start at a low amplitude and build in intensity
(e.g., some wind instruments)

Localisation of Sound Sources

Our hearing system allows us to determine the location of sound sources with reasonable accuracy, subject to certain limitations.

- **Stereo hearing** allows us to locate the source of a sound by comparing the sound arriving at each ear.
- **Head movement** allows us to improve the localisation accuracy of stereo hearing.
- **Analysis of reflected versus direct sound** yields information about the route a sound has travelled to reach us.
- **Familiarity / Pattern-Matching** affects localisation accuracy - both ways.

Stereo hearing allows us to compare:

- **amplitude** (interaural intensity), although our sensitivity to amplitude changes is limited
- **time of arrival** (interaural delay) - we can recognise differences of 10 micro-seconds or less between the time of arrival of a sound at each ear.

Stereo hearing works in the horizontal plane only and is least effective in the middle range of audible frequencies.

Research has shown that human ability to locate sources of sound in the horizontal plane varies from:

- Around 1^{o} or less when the source is directly in front/behind the listener.
  However, *front-back reversals* are common.
- Around 15^{o} or more when the source is to the left or right of the listener.
- Sounds originating within an arc of approximately 70^{o} on either side of the head are localised with least accuracy.

Localisation performance is better for non-musical sounds than for musical tones.

Localisation errors at the sides of the head are typically:

- 8^{o} for clicks
- 5.6^{o} for other noises

Front-back reversals are also less common for clicks and noises.

Localisation accuracy varies with frequency:

Below 1000 cycles	good	Based on timing/phase differences
1000 to 3000	poor	Neither timing/phase nor intensity differences predominate.
Above 3000 cycles	good	Based on intensity differences

Stereo hearing operates only in the horizontal plane. Localisation of sound in the vertical plane is far less accurate.

Research has shown that the average listener can reliably distinguish only **three** vertical source locations.

Judgement of distance is based partly on intensity - the quieter the sound, the further away the source.

However, distance also affects:

- The audio spectrum of the sound - some frequencies travel better than others
- The balance between reflected and direct sound - the further the sound has travelled, the more likely it is to include a significant percentage of reflected components.

Sound localisation (in both the horizontal and vertical planes) can be improved by tailoring the sound distribution.

This is done using **Head-Related Transfer Functions** (HRTFs).

Ideally, HRTFs should be tailored to suit the individual. However, this is complex and costly.

Researchers are currently trying to develop *non-individualised HRTFs* which will give a useful improvement in localisation accuracy for a substantial percentage of the population.

Sensory Memory for Audio

As with the other senses, it appears that we there is a sensory memory associated with the hearing system - the *Echoic Memory*.

It stores the last few seconds of incoming sound, in its raw form.

Researchers disagree as to the length of the store. Estimates range from as little as one or two seconds to as much as 60 seconds.

However, there is significant evidence for the existence of such a store.

The existence of this auditory store explains some of the following effects:

- **Recall of Un-attended Material**
    - A number of studies have been conducted in which subjects were asked to listen to several simultaneous streams of speech or sound, then recall the content of ONE of the streams.
    - They were either not told in advance which stream they would have to recall, or were deliberately told to concentrate on the wrong stream.
    - Subjects were able to recall the last few seconds of sound from any of the streams, but could only recall earlier material from the stream to which they had consciously listened.
    -
- **The Recency Effect**
    - If someone listens to a voice reciting a list of digits (or characters, etc.), and is then asked to repeat the digits, he or she will recall the last few digits more reliably than the earlier ones.
    - Typically the last 3-5 digits are recalled.
    - The number of digits recalled is roughly constant: if the list is made longer, more digits will be forgotten from the earlier parts of the list, but roughly the same number of digits from the end of the list will be recalled.

- **The Auditory Suffix Effect**
    - The recency effect (see above) is most noticeable when the speech or sound is followed by a period of silence.
    - If a further sound occurs after (e.g.) a list has been spoken, recall is impaired.
    - Conversely, if speech or sound is followed by complete silence, the period for which the last few seconds of it can be recalled extends significantly.

In short, the human hearing system behaves as if it incorporates a 'tape-loop' that can store a few seconds of sound:

- Sounds are recorded onto the loop as they are heard.
- New sounds are recorded over older sounds, but...
- if no new sounds are heard, the previous recording remains.

New sounds impair recall regardless of their type, but speech (or sounds that are interpreted as speech) cause greater impairment than non-speech sounds.

Research suggests that the human hearing system responds differently to speech and non-speech sounds.

Speech appears to make greater demands on mental resources. Consider, for example:

- The Auditory Suffix Effect (described above).
- Studies which show that an ambiguous sound causes more disruption to recall and other processes when it is interpreted as speech than when it is interpreted as a musical sound.

Summary

It appears that human beings are good at...

- Detecting changes in pitch, and distinguishing between differing successive pitches.
- Recognising and distinguishing between rhythmic structures.
- Recognising and distinguishing between familiar timbres.
- Localising the source of low-pitched and high-pitched sounds in the horizontal plane.

...and bad at...

- Recognising absolute pitches, or distinguishing between different pitches presented at significantly different times.
- Detecting changes in loudness (unless the changes are gross).
- Recognising and distinguishing between *un*familiar timbres.
- Localising the source of mid-pitched sounds in the horizontal plane.
- Localising the source of all sounds in the vertical plane.

Some Applications

In mainstream computing, sound is rarely used as a primary means to communicate information.

It is used mainly for:

- simple warnings (success, failure, etc).
- to make educational applications more engaging
- in entertainment applications (e.g., games)
- for branding (e.g., start-up tones)

Sound is used more extensively in a number of specialised fields, including:

- applications for blind and visually-impaired people.
- hands-free/eyes-free applications

Sound has been used in interfaces in a number of ways.

Synthetic speech is easy to use, and its meaning is immediately obvious.

However, speech is a relatively slow method of presenting information and places a heavy load on cognitive resources.

Therefore, many auditory interface developers have opted to use ***non-speech sound***.

Two different approaches have emerged:

- Auditory Icons (Gaver, 1989) are based on natural sounds, and are intended to be instantly recognisable to the user.

  However, it can be difficult to find appropriate sounds to represent many functions.

- Earcons (Blattner, Sumikawa & Greenberg, 1989) are musical motifs, etc., which are structured so as to convey information.

  This overcomes the problem of associating sounds with functions, but the user has to learn the meanings of the Earcons in each application.

A recent development is the use of ***Spearcons***:

- Spearcons use speech which has been speeded-up until it is only just recognisable.
- Users can initially identify the meaning of a Spearcon by listening carefully to the speech.
- However, as they learn the meaning of each Spearcon they can ignore the speech content and treat it as non-speech sound. This will involve much less cognitive effort.

Gaver (1991) demonstrated the potential of sound with the ARkola simulation.

ARkola simulated a drink-bottling plant, in which many processes had to be monitored to ensure efficient operation.

Gaver used complex, predominantly life-like, sounds to represent the processes in place of the traditional, visual control panel.

The study found that operators soon learned to interpret the sounds and could monitor processes effectively, even whilst performing other tasks.

The TIDE Maths project used sound to help blind people work with mathematical equations.

- 3D sound projection was used to place each term of a polynomial expression at a unique position in space.

- To avoid overloading the user with sounds, each expression had a characteristic 'background' sound which it made when not selected.
- When selected, the term would be spoken out, with non-speech sound used to indicate parentheses, grouping, etc..
- In one implementation, the terms could be manipulated using a data-glove.