Sound Visualisation
as an Aid for the Deaf,
a New Approach

by

A.A. Soltani-Farani

Submitted for the Degree of
Doctor of Philosophy
from the
University of Surrey

Centre for Vision, Speech and Signal Processing
School of Electronic Engineering, Information Technology, and Mathematics
University of Surrey
Guildford, Surrey GU2 5XH, U.K.

August 1998

© A.A. Soltani-Farani 1998
To my patient wife
Abstract

Visual translation of speech as an aid for the deaf has long been a subject of electronic research and development. This thesis is concerned with a technique of sound visualisation based upon the theory of the primacy of dynamic, rather than static, information in the perception of speech sounds. The goal is design and evaluation of a system to display the perceptually important features of an input sound in a dynamic format as similar as possible to the auditory representation of that sound.

The human auditory system, as the most effective system of sound representation, is first studied. Then, based on the latest theories of hearing and techniques of auditory modelling, a simplified model of the human ear is developed. In this model, the outer and middle ears together are simulated by a high-pass filter, and the inner ear is modelled by a bank of band-pass filters the outputs of which, after rectification and compression, are applied to a visualiser block.

To design an appropriate visualiser block, theories of sound and speech perception are reviewed. Then the perceptually important properties of sound, and their relations to the physical attributes of the sound pressure wave, are considered to map the outputs from the auditory model onto an informative and recognisable running image—like the one known as cochleagram. This conveyor-like image is then sampled by a window of 20 milliseconds duration at a rate of 50 samples per second, so that a sequence of phase-locked, rectangular images is produced. Animation of these images results in a novel method of spectrography displaying both the time-varying and the time-independent information of the underlying sound with a high resolution in real time. The resulting system translates a spoken word into a visual gesture, and displays a still picture when the input is a steady state sound.

Finally the implementation of this visualiser system is evaluated through several experiments undertaken by normal-hearing subjects. In these experiments, recognition of the gestures of a number of spoken words, is examined through a set of two-word and multi-word forced-choice tests. The results of these preliminary experiments show a high recognition score (40-90 percent, where zero represents chance expectation) after only 10 learning trials. General conclusions from the results suggest: a potential quick learning of the gestures, language independence of the system, fidelity of the system in translating the auditory information, and persistence of the learned gestures in the long-term memory. The results are very promising and motivate further investigations.
Acknowledgement

I would like first to express my deep gratitude to my supervisor Dr Edward Chilton for his encouragement, guidance, and advice during the course of this study; and then to acknowledge the full financial support provided by the Ministry of Culture and Higher Education (MCHE) of my country, Iran.

I am grateful to Professor Josef Kittler and all the members of Centre for Vision, Speech and Signal Processing (CVSSP) for providing a nice and friendly environment by giving advice and help during my studies.

I am exceptionally grateful to Robin Shirley from the Psychology department for his generous helps and advice regarding the design of the experiments and the psychology aspects of this thesis and also during the course of my writing up.

I would like to thank my colleagues, especially Daniel Davies, whose voices have been recorded and used throughout the experiments. I also thank my colleagues and friends who have voluntarily taken part as subjects in the main experiment of this work.

Finally I would like to record my appreciation to my wife, Fatemeh, for her patience and support during the course of my studies and thank my dear son, Ali, and my dear daughter, Bushra for their patience, particularly when living away from me in Iran during the last year of my studies in the UK.
# Contents

1 Introduction .......................................................... 1
   1.1 Visual Aids .................................................. 1
      1.1.1 Spectrogram Display ................................ 2
      1.1.2 Lipreading ............................................. 3
   1.2 Objective of This Thesis .................................. 3
   1.3 Scope of This Thesis ....................................... 5

2 Literature Review ............................................... 7
   2.1 Time-Frequency Displays ............................... 8
   2.2 Speech-Feature Displays ................................. 11
      2.2.1 Speech Features as a Supplement to Lipreading 12
   2.3 Sound Visualisation ....................................... 14

3 Auditory Process and Modelling ............................. 17
   3.1 Human Ear ................................................. 18
      3.1.1 Outer and Middle Ears .......................... 19
      3.1.2 Inner Ear ............................................ 21
   3.2 Auditory Modelling ....................................... 29
      3.2.1 Modelling the Outer and the Middle Ears .... 30
      3.2.2 Modelling the Inner Ear ......................... 32
      3.2.3 Modelling Mechanical to Electrical Transduction 37
   3.3 Summary .................................................... 39
4 Auditory Based Visualiser

4.1 Sound Perception

4.1.1 Loudness Perception

4.1.2 Pitch Perception

4.1.3 Timbre Perception

4.2 Speech Perception

4.2.1 The Nature of Speech Sound

4.2.2 Special Mode of Speech Perception

4.2.3 Models of Speech Perception

4.2.4 Acoustic Cues and Redundancy in the Speech Wave

4.3 Static Spectrogram

4.3.1 Visualisation of Loudness

4.3.2 Visualisation of Pitch

4.3.3 Visualisation of Timbre

4.4 Correlogram Movie

4.4.1 Computation of Correlogram

4.4.2 Shortcomings of Correlogram

4.5 Dynamical Spectrogram

4.5.1 Phase Locking Evidences

4.5.2 Frame-Locking Procedure

4.5.3 Advantages of Dynamical Spectrogram

4.6 Speech Visualiser

4.7 Audvis, a Visualiser Tool

4.8 Summary

5 Experiments, Result, and Discussion

5.1 Methods

5.1.1 Speech Materials

5.1.2 Subjects
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2 Main Experiment</td>
<td>81</td>
</tr>
<tr>
<td>5.2.1 Test Procedure</td>
<td>81</td>
</tr>
<tr>
<td>5.2.2 Evaluation Measure</td>
<td>85</td>
</tr>
<tr>
<td>5.3 Results and Discussion</td>
<td>86</td>
</tr>
<tr>
<td>5.3.1 Results</td>
<td>86</td>
</tr>
<tr>
<td>5.3.2 Discussion</td>
<td>89</td>
</tr>
<tr>
<td>5.4 Other Experiments</td>
<td>93</td>
</tr>
<tr>
<td>5.4.1 Recognition of New Gestures Without Learning</td>
<td>94</td>
</tr>
<tr>
<td>5.4.2 Memory Decay for Visual Gestures</td>
<td>95</td>
</tr>
<tr>
<td>5.4.3 Short Term Memory</td>
<td>97</td>
</tr>
<tr>
<td>5.5 Summary</td>
<td>97</td>
</tr>
<tr>
<td>6 Conclusion and Future Work</td>
<td>100</td>
</tr>
<tr>
<td>6.1 Conclusion</td>
<td>100</td>
</tr>
<tr>
<td>6.1.1 Outline of the Problem Addressed</td>
<td>100</td>
</tr>
<tr>
<td>6.1.2 Summary of the Research</td>
<td>101</td>
</tr>
<tr>
<td>6.1.3 Contributions and Findings</td>
<td>102</td>
</tr>
<tr>
<td>6.2 Future Work</td>
<td>104</td>
</tr>
<tr>
<td>6.2.1 Psychological Investigations</td>
<td>104</td>
</tr>
<tr>
<td>6.2.2 Technical Developments</td>
<td>104</td>
</tr>
</tbody>
</table>
List of Figures

2.1 Visible Speech Translator, a sound visualiser device used at Bel Laboratories during 1940’s. This schematic diagram shows how speech sounds are converted into visible patterns. Form [75, page 17]. ......................... 9

2.2 Learning progress of the visible-speech training program conducted by Potter, et. al. during 1940’s. Case A pertains to a congenitally deaf engineer; cases B and C belong to two groups of two women each. Reproduced from [75, page 27]. .................................................. 10

3.1 Simple block diagram of the sound visualiser system. ...................... 17

3.2 Human ear (from Denes and Pinson, 1993 [19]). ......................... 19

3.3 The cochlea. (a) unrolled. (b) cross section. From [19]. ............... 22

3.4 Basilar Membrane ................................................................. 23

3.5 Amplitude of vibration occurring on different points of the basilar membrane due to a sine-wave stimulation activating the stapes. ............... 25

3.6 A schematic diagram of an inner hair cell showing the hairs and the synapses of auditory nerve fibres connected to it. The cell body sits on the basilar membrane. .................................................. 27

3.7 A typical function of rate-intensity for an auditory fibre at its best frequency. The threshold and saturation are higher for frequencies above or below this particular frequency. The dynamic range at frequencies above the characteristic frequency is a little greater [66]. ...................... 28
3.8 A simplified model of the peripheral part of the auditory system up to neural firing activity of the inner hair cells. This model is more accurate at moderate sound levels. .......................................................... 29

3.9 The frequency responses of the preemphasis filter used as a model for the cascade of the outer and the middle ears for different values of $\alpha$. The value of $\alpha = -0.6$ has been chosen in our auditory model. ........ 31

3.10 (a) Frequency response of the model of an auditory filter centred at 1 kHz (solid curve), compared to the one extracted from the formula suggested by Moore and Glasberg, [57], (doted curve). Note the similarity of the two curves in the pass-band area of the frequency. An optimal model, obtained from the gamatone function (dashed curve) has been also shown for comparison. (b) The impulse responses of the two models. ........ 36

3.11 Frequency responses of 10 successive filters of the simulated auditory filter-bank. As is evident, there is a high percentage of overlap (78 percent) between the successive filters. ................................. 38

4.1 Equal loudness contours. For each specified loudness, the intensity of a tone required to produce the given loudness has been depicted versus its frequency. Redrawn from [55] .................................................. 42

4.2 Pitch versus frequency of a pure tone. After [34]. ......................... 44

4.3 Auditory-based spectrogram of the utterance 'sky noise', uttered by a female speaker. Note the variations of the formants, especially the second formant (F2) which seems to represent the motion of the tongue during the speech. ................................................................. 48

4.4 Auditory-based spectrogram of the utterances 'boil', 'boys', and 'noise'. Compare the different patterns of the same phoneme /oi/ in association with different consonants. .............................................. 50

4.5 A 40-millisecond frame of the cochleagram (a static spectrogram) of the vowel /ai/ from the word 'sky', uttered by a female speaker. ........... 54
4.6 A number of successive, 20-millisecond frames of the correlogram of the vowel /æi/ from the word 'sky', uttered by a female speaker. When animated, these images display the motion of the periodicities of the sound for different frequency regions. Images have been displayed in the order of top to bottom. .......................................................... 62

4.7 Cochlear vibration (dotted curves) and firing pattern of auditory nerve fibres (solid curves), as represented by a very simplified model of the auditory system for a low CF channel (236 Hz) and a high CF channel (2930 Hz). .......................................................... 64

4.8 Dynamical spectrogram of speech sound /æi/ from the word 'sky', uttered by a female speaker. Note transitions of the formants F1 and F2—the lower and upper bands of enhanced energy (darkening) across groups of frequency channels, which move apart as the sequence of frames is descended. These transitions are clearly perceived when the images are displayed, one by one, on the same screen with the speed of 50 images per second. .......................................................... 67

4.9 Comparison between the correlogram and dynamical spectrogram of a synthesised harmonic sound having its odd components mistuned. Note the phase changes between successive frames of the dynamical spectrogram. The correlogram is unable to show the phase changes, despite the beat sensation given rise to during the auditory perception of the underlying sound. .......................................................... 70

4.10 Modified shape of the dynamical spectrogram. The sine-wave shape of its vertical sides act as markers which help the viewer to relate the vertical positions to their corresponding frequency. .......................................................... 71

4.11 Dynamical spectrogram of the utterance 'yes' spoken by a British male speaker. The frames are animated with a speed of 50 frames per second in the order of top-left to bottom-right, in vertical steps. These images, then, cover a duration of 360 milliseconds of the speech sound. .......................................................... 72
List of Figures

4.12 Block diagram of the Audvis software package. (a) Auditory model and image frame-locking. (b) Audvis, the audio-visualiser tool. This tool can be used for monitoring any sequence of audio-visual blocks (individual image frames with their corresponding audio frames). ................................... 73

4.13 The Audvis software tool. This can be used to monitor a sequence of blocks, each containing a frame of an auditory signal and the rectangular image assigned to it. Its schematic block diagram is shown in Figure 4.12b. ................................................................. 75

5.1 Layout of the interactive window of the test session, (a) at the start of the session, and (b) at the Learn stage of the first test. For the Practice and Question stages see Figure 5.2. ................................. 82

5.2 Layout of the interactive window of the test session, (a) at the Practice stage, and (b) at the Question stage of the first test. See also Figure 5.1. 83

5.3 Means and standard deviations of chance-adjusted, percentage correct scores for the ten two- to four-word, forced-choice tests of the main experiment. Averages have been computed over the 30 normal-hearing subjects in Table 5.4. In fact scores cannot exceed 100—that the error bars for tests T4 and T6-T8 appear to do so is an artifact that reflect the presence of a ceiling effect when the scores approach the upper bound. 89

5.4 Average of the scores obtained by the subjects in relation to their ages. The abscissae represents the age and the ordinate shows the average of the 10 chance-adjusted, percentage correct, scores of the tests undertaken by each subject. ................................. 92
# List of Tables

5.1 The full set of English words used in the experiments in the form of two- to four-word, forced-choice tests. ................................................. 80

5.2 Pronunciations of the translations of the words 'yes' and 'no' in three further languages, used in the tests of language independence. ............... 80

5.3 The words used in the multi-word, forced-choice tests and the conditions of their repetition in each test. ...................................................... 84

5.4 Percentage scores for the ten forced-choice tests for all subjects, together with their averages and their ages. Mean and standard deviation of the scores of each test over the subjects are also included in the table. Note that a score of zero represents chance expectation. For the conditions of the tests see table 5.3. ................................................................. 87

5.5 Characteristics of the subjects taking part in the main experiment. The average score obtained, the elapsed time, and the time of day at the start of the test session, have also been shown for each subject. ................. 88

5.6 Matrix of confusion between the four words 'south', 'west', 'north', and 'east' (a), and between the three words 'one', 'two', and 'three' (b). ................. 93

5.7 Results for a test involving the four words 'south', 'west', 'north', and 'east' without any prior learning stage. The words can thus have been recognised only because of their similarity to previously known words. Errors are shown in bold. ................................................................. 94

5.8 Results from an extended series of repeated recognition tests for the words 'zero' to 'nine', without any practice between tests. .......................... 96
Chapter 1

Introduction

Deaf people, particularly the congenitally deaf, are faced with two major problems in their communication with other people with normal hearing, namely speech perception and speech production—hearing and speaking. Aids for the deaf, then, can be divided into the categories of speech perception aids and speech training aids. In this thesis it is intended primarily to contribute towards the solution of the first problem—the hearing problem of the deaf. It may also be the case, although not specifically explained here, that the approach used here may have a bearing on the second problem, through its potential for use as a speech training aid.

In the electronic age, speech perception aids have exploited two different sensory routes—tactual and visual—depending on their use of the skin or the eyes, respectively, as an alternative to the ear. In these devices, which are also known as sensory aids, speech information is transmitted as tactual or visual signals which can be perceived by the skin or the eyes of a deaf user. Both approaches have been pursued since the beginning of this century without any real breakthrough (see chapter 2, Literature Review).

1.1 Visual Aids

Apart from direct speech to text conversion, a field which will not be considered in this thesis, visual aids have been developed mainly using two different forms of presentations: spectrogram displays and lipreading.
1.1.1 Spectrogram Display

The spectrogram-display technique was first introduced in the classic work of Potter, et al. [75] during 1940's. In this, the spectrogram of ongoing speech is presented in the form of a sliding, conveyor-like picture which displays the time-frequency patterns of the input sound at a relatively low resolution (see Figure 2.1). The time resolution of the spectrogram used in this method is inevitably limited by the speed of the moving belt which is, in turn, determined by the ability of the human eye to follow the running patterns.

There are two questions regarding this type of visual representation of speech:

1. Is the perceptually important information contained in the input sound recognisable through watching the output picture?

2. Is this information recognisable at a normal speed of speech.

After a considerable amount of research on spectrogram reading, the answer to the first question proved to be positive [28, 42, 75, 108]. But in response to the second question, it was realised that no one could manage to read the spectrograms with anything like the same effectiveness with which we understand the spoken words.

Two main explanations have been given for the lack of success with spectrogram reading aids. One argument attributed the relatively modest results of the experiments to the poor resolution of early displays, lack of appropriate training strategies, and short time of exposure to the displays—that is, that the techniques were too primitive (but might be improved). A more fundamental argument, on the other hand, holds that speech is not a form of simple alphabet to be readable by the eyes, but rather that it is a complex code for which the auditory system is a unique decoder, and therefore it is unrealistic to expect the visual system to be effective at taking over the role of the auditory system (see section 4.2). Referring to the precedent of lipreading as a classic example of at least partly successful visual perception of speech, however, they concluded that a more effective type of visual aid might be via a display of the vocal tract articulation.
1.1.2 Lipreading

The arguments against spectrogram reading together with the unsatisfactory results obtained from attempts to explore it shifted the main focus towards lipreading research. Various workers have reported projects to develop systems that produce cartoon-like animations of the lip, tongue, and jaws, etc., according to the speech features extracted from the input sound signals. But this approach, even if ideally implemented, has a major problem which is that the information available through lipreading contains many ambiguities. These ambiguities mainly arise from visually similar articulatory movements which produce different speech sounds. To avoid this problem, a number of workers have proposed speech-feature displays for use as a supplement to lipreading. For example an innovative, wearable aid referred to as eyeglass speechreader has been invented that presents five features of speech in the form of binary values, displayed on five distinct points of the eyeglass. This device, although used by its inventor for years, was limited in that it could not work properly in noisy environments.

In this type of aid, several speech features, e.g. prosody of speech, which cannot be perceived during lipreading, are automatically extracted from the speech signal and presented to the lipreader as a partial compensation for the ambiguities left unresolved by the lipreading process. A critical problem for this type of aid, then, is that the automatic recognition of speech features is subject to error, particularly in noisy environments. Although dramatic developments have been taken place in the field of automatic speech recognition, a full solution to this problem remains some way off.

1.2 Objective of This Thesis

The first type of visual aid, which presents spectrograms, is in fact a sound visualiser, whereas the second type, a lipreading aid, is a speech visualiser. An aid that is a sound visualiser has three major advantages over a lipreading aid:
1.2. **Objective of This Thesis**

1. A sound visualiser system can be used for perception of sound, a much more general source of information than speech. There are various kinds of informative sounds such as alarms, door bells, background noise, songs, music, etc., which a deaf user can potentially understand using an appropriate sound visualiser.

2. A sound visualiser system works independently of the language of the user, but a lipreading aid is strongly dependent on the language to which it is applied because of the need for robust extraction of the relevant speech features and the phoneme differences between languages.

3. Since a lipreading system involves automatic recognition of speech features, it is relatively vulnerable to noise, interference, or distortion. But when using a sound visualisation aid, the recognition task is transferred to the user's brain which is an amazingly powerful and robust system, although no specific tests for robustness to noise will be described here.

The main problem with the sliding spectrogram as a sound visualiser is that it is difficult to read at normal speech speeds. This problem, we shall argue, is due mainly to the static, rather than time-varying, nature of the information delivered by the spectrogram. The speech information, by contrast, has primarily a *dynamic* format [85] which, therefore, has its nature changed if it is illustrated by a sliding *frozen* image such as a spectrogram. By the same argument, the feasibility of lipreading as an aid for speech perception arises because of its time-varying characteristics. In keeping with this idea, there is an interesting discussion in [85] about the primacy of kinetic/dynamic, rather than static/pictorial, facial information in the perception of speech through lipreading. This is the key hypothesis that underpins the investigation started by this dissertation.

Taking this into account, the work presented here is aimed at discovering a recognisable dynamic format for visual information to represent sound. It is based on this belief that there must be some form of sound visualisation through which a
1.3 Scope of This Thesis

To achieve this goal the human auditory system has been studied to understand how it represents sound information and presents it to the auditory nervous system. Then the psychological and psychophysical issues of sound and speech perception have been examined to determine which qualities of sound are more important from the perception point of view and how they are related to the physical attributes of the sound pressure wave.

From these studies, a simplified model has been developed of the human ear, up to the firing activities of the auditory nervous system. A visualisation tool has then been created to demonstrate the output of this model in the form of a moving image. Finally the resulting system has been partly evaluated through a number of recognition tests.

Chapter 2 provides a brief account of previous work on visual aids reported in the literature. Recent findings about speech-feature displays and criticisms regarding spectrogram reading are described in more detail.

Chapter 3 contains two main sections. The first section contains a review of the anatomy and physiology of the human ear followed by a summary of the functions of different parts of the ear—the outer, middle and inner ears. The complex structure of the cochlea and the functions of the auditory nerve fibres connected to it are also schematically described in this section. The second section introduces a simplified model of the peripheral part of the auditory system, developed for representing auditory signals in approximately the same way as they are delivered to the auditory nervous system by the inner ear.

Chapter 4 is devoted to finding an appropriate method of visualisation for the output signals of this auditory model. It provides first a brief description of the human perception of sound and outlines the perceptually important qualities of speech sounds. This is followed by a short review of existing knowledge about speech
1.3. Scope of This Thesis

perception. Then considering the most perceptually important qualities of sound, particularly those of speech sounds, current theories of auditory representation in the human ear are employed to devise an appropriate method of sound visualisation. A comparison between our method of sound visualisation and other methods is also given in this chapter. Finally our implemented system of visualisation, is described in the form of two block diagrams.

Chapter 5 reports several experiments carried out for the evaluation of our system of sound visualisation. The results of experiments using normal-hearing subjects followed by discussions about the related findings are among the issues to be treated in this chapter.

Chapter 6 itemises the contributions and findings of this thesis. This chapter also outlines the scope for future work on the development of our sound visualiser system.

Some of the work described here has been already published or accepted for publication [93, 94]. An additional paper will be presented in November 1998:

A. A. Soltani Farani, E. H. S. Chilton, and R. Shirley. Dynamical spectrogram, an aid for the deaf. In The 5th International Conference on Spoken Language Processing (ICSLP'98), Sydney, Australia, 30th November - 4th December 1998.
Chapter 2

Literature Review

Research in the field of speech processing aids for the deaf has been going on since the beginning of this century. Aids for the deaf are divided into two main categories: speech-perception aids and speech-training aids. Although most of the speech perception aids, such as speech visualisation, can also be useful for speech training, in this review we focus on the perception aspect of the speech processing aids.

At first there was a great deal of interest in tactual aids. The most important outcome of this period of research was that it was recognised that deaf people could feel and remember the vibration patterns of different speech sounds [32]. Research in tactual aids declined at about the time of World War II, in spite of the enthusiasm raised during that early period. After about a decade, however, there was a revival of interest in the area, and since then it has remained, more or less, an active field of research alongside other domains of speech processing aids. The invention of electronic visual displays, during 1930's, opened a way into the field of speech visualisation as an aid for the deaf. So far visual and tactual aids have been progressing, in parallel, without any real breakthrough.

A collection of papers regarding different aids for the hearing impaired has been published by Levitt, et al. in 1980 [44], reflecting the contemporary knowledge about the subject. Also a brief account of the history of speech processing aids may be found in a paper recently published by Risberg [82].

The classic example of a visual aid for the deaf is lipreading, in which pro-
foundly deaf people attempt to perceive spoken words by watching the articulatory movements and gestures of the speaker. However, the perception of information by lipreading is limited by factors such as the similarity of articulatory movements for some different speech sounds (e.g. /p/ and /b/), and lack of information about the prosody of the speech signal. Feasibility of speech perception through lipreading together with its unfortunate limitations have motivated many researchers to investigate the usefulness of a number of automatic speech visualisation methods.

Research work on visual aids can be divided into two categories: time-frequency displays, and speech-feature displays.

2.1 Time-Frequency Displays

The invention of the cathode-ray tube raised excitement among those who had thought about the visualisation of speech waveforms in real time. But soon it was realised that speech waveforms were too perceptually complex to be useful for the deaf. Further hope was raised by the invention of the sound spectrograph, leading to the development of a real-time, sound-spectrum analyser. This experimental device, known as the Visible Speech Translator (VST), was extensively used at Bell Laboratories for experiments on speech perception aids, resulting in the classic book *Visible Speech*, by Potter et al. [75] first published in 1947.

A schematic diagram of the speech translator is shown in Figure 2.1. With this system, speech sounds are converted into a train of visible patterns that illustrate the spectrum of ongoing speech in the form of illuminated shapes on a moving belt of phosphorescent plastic. As shown in the figure, the input speech signal is spectrally analysed through a bank of twelve band-pass filters, their centre frequencies being uniformly distributed over a range of 3600 Hz, with the same bandwidth of 300 Hz.

The book reports a major experiment in which a group of young women and a congenitally deaf engineer learned to read large vocabularies of words. The graph of their learning progress has been depicted in Figure 2.2. Curve A represents 'hours of training' versus 'number of words' for the case of the deaf subject, and curves B
2.1. Time-Frequency Displays

and C represent those for the two groups of women. According to these curves the vocabularies of the subjects increased at the rate of about 3.5 words per hour of training. The authors also report:

"The visible speech class members were able to converse satisfactorily among themselves by talking clearly and at a fairly slow rate. Within the limits of their vocabulary, they were able to carry on conversations with about the same facility as a similarly advanced class in some foreign language."

The data confirm the feasibility of reading spectrograms even with such poor resolution as that of those early displays.

Despite the remarkably good results of these experiments, research on this type of visualisation as an aid for speech perception slowed down. During the 1960's there was a revival of interest in the Visible Speech Translator. An improved version of the speech translator was tested, by House, et al. [33], to evaluate the discriminability of visual translations of eight CV syllables, differing only in their vowels, by normal subjects. The results were not as satisfactory as the authors had expected before the
2.1. Time-Frequency Displays

Figure 2.2: Learning progress of the visible-speech training program conducted by Potter, et. al. during 1940's. Case A pertains to a congenitally deaf engineer; cases B and C belong to two groups of two women each. Reproduced from [75, page 27].
2.2. Speech-Feature Displays

experiments; however, they believed that according to the data their subjects could improve further. These relatively poor results were attributed to the characteristics of the display; but we believe they might also have been because of the subtle differences between the acoustic properties of some syllables under investigation which required a long time of exposure to learn their visual translations. The new version of the Visible Speech Translator was also used in several experiments as a speech-training aid, with relatively good results [32].

After about 50 years of research on speech visualisation with no real breakthrough [82] a paper was presented by Liberman et al. [46] with the title “Why Are Speech Spectrograms Hard to Read?” In the paper it was argued that speech is not a simple alphabet but is, rather, a complex code where the auditory system serves as a unique decoder; hence it might be unrealistic to expect the visual system to act effectively as a substitute. They concluded, however, that the best visual displays of speech for the deaf might be representations of articulatory muscle contractions. The counter argument [32] is that the relatively modest improvements obtained in the past can be attributed to the poor resolution of early experimental devices, lack of appropriate training strategies, and short time of exposure to the display. A paper by Klatt and Stevens [42] revealed that much of the speech code could be deciphered from visual reading of the spectrogram. By using a further improved version of the speech translator [97], in some spectrogram-reading experiments [10, 108], it was demonstrated by Victor Zue that a person can be trained to read spectrograms by eye alone, although at a rate much slower than that of normal speech.

2.2 Speech-Feature Displays

Although the art of reading spectrograms has considerably improved, no one can read spectrograms in the same way as we understand spoken words. Lack of take up of spectrogram reading aids supported the argument presented by Liberman and his colleagues about speech being a complex code [46]. They believed that the best way to present speech to a deaf person is to display some speech features which convey
2.2. Speech-Feature Displays

the articulatory movements. Subsequent research on speech processing aids, thus, has mainly shifted towards speech-feature displays.

2.2.1 Speech Features as a Supplement to Lipreading

Lipreading as a classic aid for the deaf has attracted many researchers in recent years [9]. The information available to the lipreader comes from the movements of several vocal tract organs such as lips, jaw, and tongue, as well as from the facial expressions that are visible. But this information involves many ambiguities, as there are many sets of different speech sounds which involve the same visible articulatory movements—e.g. /p/, /b/, and /m/. Speech-feature displays can be used as a supplement to lipreading, compensating for its ambiguities.

The first paper regarding this type of hearing aid, published by Upton in 1968 [103], describes a visual speech-feature display mounted on an eyeglass which is referred to as an eyeglass speechreader. This innovative, wearable aid presents five features of speech in the form of binary values displayed on five distinct points of the eyeglass. Preliminary results of the evaluation of this device was promising; hence, an improved version of it was tested several years later by Pickett et al. [65], who showed general improvement of lipreading with the use of the eyeglass speech reader, although with large individual differences. There were further research work on lipreading during the 1970's, as partly cited in [31].

Three major questions regarding speech-feature displays as a supplement to lipreading have been under investigation during the last two decades. The first of these is:

Which features provide the most important cues for speech-reading?

This question has been addressed by many researchers such as Erber [20], Risberg [80, 81], Breeuwer and Plomp [5, 6, 7], Grant, et al. [24], and Hnath-Chisolm, et al. [30]. According to these studies, several factors are almost equally important, in providing major support to lipreading. They are: sound-pressure variations, fundamental-frequency variations, formant transitions, and spectral balance. One
of the papers [30] also studied the amount of support normal-hearing subjects could obtain from various resolutions of fundamental frequency information as a supplement to lipreading. They reported a dramatic decrease in the amount of support obtained from a 12-level quantised fundamental-frequency compared with the unprocessed frequency of an electroglottograph signal, and a further decrease with smaller numbers of quantisation levels. These findings indicate the important role of the prosody of speech in its recognition.

The second question, therefore, is:

How should this information be extracted from the speech?

Researchers often think about the automatic extraction of features from the speech signal. A major problem in this regard is that the automatic recognition of the features to be displayed tends to be subject to error. Despite significant developments in different speech recognition techniques [2, 67, 86, 98, 99, 102] the accurate and robust recognition of speech features is yet to be achieved. As examples of speech-feature extraction through signal analysis techniques see [38, 59, 95, 106].

The third, and possibly the most important, question about speech-feature displays is:

How should the extracted features be presented to the deaf user?

There are two candidates for sensory presentation of the extracted features: tactile and visual presentations. An example of visual presentation has been mentioned above: the eyeglass speech reader. But the eyeglass speech reader is intended to supplement lip-reading, and so is less useful for the perception of speech sounds produced by non-human sources (such as telephones), or simply by a human whose face is not visible.

Some researchers have considered automatically converting speech into vocal tract movements in real time [43, 90, 76]. A paper by Lavagetto [43] describes about a device created to convert speech into lip movements, referred to as a multimedia telephone for hard of hearing people. A bank of six neural networks were used
to independently analyse the audio signal in order to extract different articulatory parameters, which are then used to generate the relevant animation displayed on a screen. The preliminary experimental results have been reported to be encouraging and proving the feasibility of their method. However, their method, even if successful, is not capable of conveying all the speech features in the form of facial gestures.

2.3 Sound Visualisation

We propose that the most useful way of presenting speech to the deaf might be displaying sound features, instead of speech features, as the human brain can perform much better than the artificial neural networks in extracting speech features from suitably presented sound qualities. The feasibility of speech perception through lipreading and of spectrogram reading (although very slow) suggests that it is not the auditory system as such that recognises speech (or decodes the auditory speech signal), but rather that speech understanding is carried out by deeper processes in the neural systems of our brains, programmed very early in our life. The ability of humans, even of adults, to learn a new language [41, 48, 49] attests the feasibility of learning a new system of coding in a reasonable period of time.

The Visible Speech Translator, as already discussed in section 2.1, is, in fact, an attempt to visualise sound. The idea was first discussed in a paper by Potter [74] (the author of the book Visible Speech [75]) with the title “Visible Patterns of Sound”, in which a number of different sounds such as hammer clicks, siren, music, and songs of birds, etc., were illustrated in the form of static spectrograms. The lack of success of the idea, in spite of both the favourable results reported in the book and further work carried out over the following four decades, is proposed here mainly to be due to the shortcomings of the spectrograms used.

Shortcomings of this type of visualisation, as we think, are as follows:
Spectrograms are static In all the work undertaken so far, the spectrograms under investigation have been static: even in real-time applications sliding spectrograms like the one in Figure 2.1 have been used. In a static spectrogram the temporal variations of a speech sound are mapped onto a spatial format, rendering quite a different scene. Moreover, reading of these sliding, frozen pictures demands a voluntary scanning of the eye, like ordinary silent reading, rather than more automatic actions such as in watching or listening. Reading a spectrogram is like reading a hand-written text in a new language having a huge number of letters. That is why a person needs a long time to learn them and to attain the required speed of reading.

Lack of information about prosody of speech Pitch and its variations which may deliver much information about the stresses, the speaker, and the voicing qualities, of an utterance are not illustrated clearly in the spectrograms. The role of the pitch variations in perception of a train of speech sounds is not less important than that of the intensity variations [7, 30, 84].

Low resolution Although recent experimental spectrograms [97] have more resolution, their frequency resolution is not yet sufficient considering the remarkably high resolution of the human auditory system. This may affect the perception of timbre of speech sounds as a major cue to discrimination of some speech sounds such as vowels. There are, of course, further different types of time-frequency distributions, recently developed, which provide arbitrarily high resolution in both time and frequency domains, but they have not been used for hearing aid investigations [4, 11, 47, 68, 92]; moreover, the time resolution of a sliding spectrogram is inevitably limited by its sliding speed which is, in turn, determined by the ability of the human eye to follow the running patterns.

We suppose the visual representation of the auditory-perceptible qualities of sound, with sufficient resolution, in a form as much similar as possible to the one represented by the ear, may be a solution to the problem. Sound is essentially a motion [85] and therefore its representation in the form of still patterns on paper is in
fact a deviation from its nature. Moreover, moving parts of a picture are more easily perceptible and more persistent than still ones. Therefore converting an auditory event into a synchronised, live animation seems likely to be much more effective than into a sliding, frozen image. Thus we shall commence this new attempt in the field of sound visualisation by reviewing the functions of the human ear in the analysis and representation of sound, and hence seek a more appropriate visual display for it.
Chapter 3

Auditory Process and Modelling

In seeking an informative, visual representation of sound it seems reasonable to begin by considering the way our ears work in extracting audible information from the acoustic signals. As researchers in different areas of speech and sound processing have concluded, the human ear is a unique system for analysing sound, with an amazing resolution in both the time and frequency domains and over an extremely broad dynamic range. That is why the mammalian auditory system has been closely studied by workers in such varied fields as communication engineering, psychology, physiology, neurology, etc., during recent decades. Using state-of-the-art knowledge regarding the operation of the ear, we hope to be able to design an effective system of sound visualisation that will be useful for deaf or hearing-impaired people.

This chapter, together with the next, is devoted to a different type of sound visualisation system for which a simple block diagram is shown in Figure 3.1. By this system the digitised input sound is applied to a computational model of the ear. Output from that model is then mapped onto a sliding image, known as the

![Figure 3.1: Simple block diagram of the sound visualiser system.](image-url)
3.1 Human Ear

Cochleagram image [92], which represents the time-frequency pattern of the input sound with a high resolution. Applied to the animator block of the system, this sliding image is transformed into a sequence of rectangular images which are displayed on a screen, one after another, at a rate of 50 images per second; that is, each image is depicted on the screen for 20 milliseconds. The result of this animation is a motion picture which is synchronised with the underlying sound and which can thus demonstrate its audible information in a visible form.

In the current chapter, the first block of the system, the auditory model, is introduced; the visualiser and animator blocks are discussed in the next chapter. In the first section as a brief description of human auditory system, the anatomy and the physiology of the ear are reviewed, followed by a short account of auditory nerve activities, according to recent studies. The second section of the chapter is concerned with the modelling of the ear.

3.1 Human Ear

The human auditory system is usually divided into two parts: the peripheral and the central. The peripheral part comprising the ears and the neural fibres connected to them has been investigated in detail by many researchers. Although there are some discrepancies between the explanations given by different workers regarding some aspects of the hearing process, most of the findings about the auditory periphery have generally been accepted. The central part, on the other hand, is still far from understood, because it is associated with the central nervous system which is less anatomically accessible and exhibits much less structural differentiation. Psychoacoustic data obtained from psychological experiments, however, have provided much information about the work of the brain which has been used to establish a reasonable theory of hearing. This section will give a brief summary of the human ear up to the firing activity of the cochlear nerves. For a more detailed description of the human auditory system, the reader is referred to the books cited in [19], [53], and [66].
3.1. Human Ear

3.1.1 Outer and Middle Ears

Figure 3.2 illustrates the anatomy of human ear. The ear from a functional point of view is divided into three main parts: the outer ear, the middle ear, and the inner ear. The outer ear consists of the externally visible portion of the ear, the pinna, and the ear canal. The ear canal, an air filled pipe of about 2 centimetres in length, is open to the outside world at one end and closed off by the eardrum at the other end. The outer ear, therefore, behaves as an acoustic resonator for the sound waves guided through it. The guided pressure waves amplified at frequencies near the resonance frequency (of about 3 kHz) set the eardrum into vibration. Although there are other functions such as sound localisation associated with the outer ear, its main role is to work as a bandpass filter.
3.1. Human Ear

The vibrations of the eardrum are transmitted through the middle ear by three small bones, the **ossicles**, to an opening in the bony wall of the fluid filled inner ear—the **cochlea**. This opening is called the **oval window**. The three small bones are called the **malleus**, **incus**, and **stapes**. The handle of the malleus is attached to the eardrum. Vibrations of the eardrum are passed by the malleus to the incus by which, in turn, they are transmitted to the stapes. The foot-plate of the stapes, connected to the membrane covering the oval window, transfers the motion into the fluids inside the cochlea. Thus the airborne pressure waves at the outer ear are transformed by the middle ear into the fluid vibrations at the inner ear (although air is technically also a fluid, the term “fluid” will be used in this account to imply “liquid”).

In this transformation, two functions are performed. First, amplification of the pressure waves entering the cochlea, and second, mechanical impedance matching between air and liquid, which decreases the loss of energy in between.

Another major function of the middle ear is to protect the inner ear against intensive sounds. This is believed to be accomplished through the action of the small muscles connected to the malleus and the stapes. Contraction of these muscles, known as the middle ear reflex, tends to decrease the flexibility of the chain of the ossicles and affects the impedance matching work of the middle ear. This substantially decreases the pressure variations transmitted to the inner ear at high sound levels, to protect the delicate structure of the cochlea. Unfortunately, activation of the reflex is too slow to work instantaneously against impulsive sounds such as gun shots, etc. and they can, therefore, cause permanent damage to our ears.

Two other functions have been suggested for the reflex. First is the reduction of the sound received from the listener’s own vocal tract during speech and the second, which is more relevant to our discussion, is a reduction of the masking of high frequencies by the lower ones. This is because the reflex mainly affects the low and middle frequency part of the auditory spectrum.

Transmission of sound through the middle ear is most efficient at the middle
frequencies (500–4000 Hz) [53].

### 3.1.2 Inner Ear

The inner ear consists of two systems: the *semicircular canals* and the cochlea. Although these are connected together as a bony vessel of fluids in the temporal bone of the skull, they have two distinct functions. The semicircular canals act as a three dimensional rotation-detection system responsible largely for the orientation and balance of the body. The other system, the important one from our point of view, is the cochlea, a snail-like tube of 2 3/4 turns and 3.5 centimetres in length. To envisage how it would look, if unrolled, the cochlea has been schematically shown in Figure 3.3a as a straight tube. One end of the cochlea where it is connected to the middle ear is called the *base* and the other end, the tip of the tube, is called the *apex*.

The cochlea is divided along its length into three spiral tunnels separated by Reissner’s membrane and the basilar membrane. The two outer tunnels, the *scala vestibuli* and the *scala tympani*, are connected through the *helicotrema*, an opening at the apical end of the cochlea, where fluid can pass freely between them. Both chambers are full of an almost incompressible fluid. The scala vestibuli is linked to the middle ear through the oval window, the membrane-covered opening attached to the foot-plate of the stapes in the middle ear. The scala tympani is also connected to the middle ear through another membrane-covered opening known as the *round window*.

Between the two chambers lies the most intricate part of the inner ear, the *scala media*, which is also full of fluid but one which differs from that inside the other cavities. The scala media (also called the *cochlear duct*) is a spiral cavity between the basilar membrane and the Reissner’s membrane. Inside this cavity, attached to the basilar membrane along its entire length, there exists the *organ of Corti* (Figure 3.3b) which is an interface between the mechanical part of the ear and the auditory nervous system. The organ of Corti contains about 20000 sensory receptors known
3.1. Human Ear

Figure 3.3: The cochlea. (a) unrolled. (b) cross section. From [19].
3.1. Human Ear

Figure 3.4: *Basilar Membrane*

as hair cells which are innervated by the auditory nerves.

**Basilar Membrane**

Looking like a belt, the basilar membrane is very narrow (about 0.04 millimetres) at the basal end of the cochlea, near the oval window, and has its maximum width (about 0.5 millimetres) at the other end, the apex. There is a gradual transition between these extremes along the entire length of the cochlea. Furthermore its stiffness gradually decreases from the narrow end towards the wide end. Thus, the basilar membrane is narrowest, lightest, and stiffest at the basal end and widest, heaviest, and most elastic at the apical end of the cochlea (Figure 3.4).

The cochlear structure is excited through the oval window by the motion of the
stapes foot-plate activated by an input sound—or any other means. If the motion is slow, inward movement of the oval window displaces fluid towards the helicotrema through which the displacement is passed back along the other side to the middle ear, resulting in a corresponding outward movement of the round window. But if the oval window is set into vibration by an incoming sound, any pressure variation in the fluid on either side of the cochlear partition results in a pressure difference across the partition. The pressure wave travels very rapidly through the incompressible fluids of the cochlea applying the pressure difference almost simultaneously along the whole length of the basilar membrane. The pattern of motion on the basilar membrane, however, due to its mechanical properties takes some time to develop. A form of travelling wave, initiated from the base, progresses along the basilar membrane toward the apex. The amplitude of the wave increases at first but then after passing a point of maximum diminishes rather sharply.

The response of the basilar membrane to sine-wave stimulations, in a steady state, reveals the special pattern of the vibration along its length. For a sinusoidal input each point on the basilar membrane vibrates in a sinusoidal manner with the same frequency as that of the input sound. The amplitude of the vibration, however, is different for different points of the basilar membrane. The position of maximum amplitude depends strongly on the frequency of the stimulation. For high frequencies the points near the oval window, where the basilar membrane is lightest and stiffest, vibrate with maximum amplitude. For lower frequencies the peak of vibration moves toward the other end of the membrane, where it is wider and more elastic (see Figure 3.5). The form of the basilar membrane, therefore, brings about a spatial separation of the maximum sensitivity to different frequencies. Thus, for a complex sound, comprising tones of different frequencies, the spectral pattern of sound is mapped onto a spatial format.

The curves shown in Figure 3.5 do not fully reflect the remarkably high sensitivity of the human ear in the frequency domain. This is because they were obtained using an optical technique, during the classic work of von Békésy [3], in which too
3.1. Human Ear

Figure 3.5: Amplitude of vibration occurring on different points of the basilar membrane due to a sine-wave stimulation activating the stapes.
3.1. Human Ear

high levels of cochlear vibrations were used. Recent work [23, 40, 79, 87] suggests that the degree of frequency selectivity of the basilar membrane depends on the intensity of the sound applied, in such a way that at low levels of vibration it is much more selective than at high levels. Indeed, any point along the basilar membrane can be considered as a bandpass mechanical filter with a resonance frequency characteristic of its location. The bandwidth of the filters and the distribution of their characteristic frequencies along the length of the basilar membrane have been studied by many workers in different ways which will be discussed later.

Hair Cells

The mechanical motion of the basilar membrane needs to be converted into electrical signals that can be transmitted and processed by the central nervous system. The organ of Corti (Figure 3.3b) contains about 3500 sensory receptors known as inner hair cells which are innervated by the auditory nerves. Uniformly distributed in a single row along the basilar membrane, the inner hair cells convert the motion at their position on the basilar membrane into firing activity at the nerve fibres attached to them. Figure 3.6 shows a schematic diagram of an inner hair cell. The hairs, a bundle of fine filaments protruding from top of the cell body, are set into vibration as a result of the motion of the basilar membrane. This has been described by saying either [66] that the hairs, having contact with a fixed membrane at their other end, are bent as a result of the movement of their parent cell, or that [19] they vibrate due to vibration of the fluid surrounding them. Movements of the hairs induce some electrical voltage inside the cell which in turn sets the nerve fibres attached to it into action. The firing activities are transmitted in the form of electro-chemical pulses travelling along the nerve fibres towards the higher levels of the auditory nervous system.

It is the firing rate, i.e. the frequency of the pulses, which represents the intensity of the vibration of the basilar membrane. There are about 20 nerve fibres, of different diameters, attached to a single inner hair cell. An auditory fibre has some
3.1. Human Ear

Figure 3.6: A schematic diagram of an inner hair cell showing the hairs and the synapses of auditory nerve fibres connected to it. The cell body sits on the basilar membrane.

background activity, the spontaneous firing in the absence of sound stimulation, the rate of which ranges from close to 0 up to about 150 pulses per second [19]. When an inner hair cell is set into vibration any fibre attached to it will be activated provided that the intensity of the vibration exceeds the fibre's specific threshold. The fibres having higher spontaneous rate have lower threshold while, on the other hand, fibres of lower background activity need a more intensive stimulation to set them into firing.

Different workers [66, 83] have studied the firing response of single nerve fibres to various input sounds and their common results can be summarised, so far as they relate to our discussion, as follows:

- The firing rate of a fibre increases as the intensity of the stimulus increases. This increment is started after a specific threshold of sound intensity and stops at a saturation point (see Figure 3.7). The dynamic range between the threshold and the saturation point is typically about 30 dB.

- For a given fibre, the threshold and saturation points depend on the frequency
3.1. Human Ear

Figure 3.7: A typical function of rate-intensity for an auditory fibre at its best frequency. The threshold and saturation are higher for frequencies above or below this particular frequency. The dynamic range at frequencies above the characteristic frequency is a little greater [66].

of the stimulus. The lowest threshold is associated with the characteristic frequency of the position of the related hair cell on the basilar membrane. This frequency is also called the best frequency of the fibre.

- The temporal pattern of the firing activity in response to a single tone shows that the firing rate increases in only one half of the cycle of the input waveform. That is, the fibres show phase locking; so that the firing pattern of a fibre has a temporal regularity in response to some periodic stimulation. This phase locking occurs at lower frequencies below 4-5 kHz.

There are also several rows of hair cells, parallel to but separated from the inner hair cells, known as the outer hair cells. It has recently been discovered that they change shape in response to signals received from the central nervous system, and
3.2 Auditory Modelling

The amazing ability of human auditory system to distinguish and recognise sounds has motivated many researchers to study and model its functions of sound analysis in different areas of audio-acoustic processing, such as: speech recognition, speech coding, speech training, time-frequency representations, etc. [1, 22, 92]. Here the auditory modelling is approached from the viewpoint of sound visualisation as an aid for the deaf. From this point of view, we are interested in mimicking the work of the ear as an aid for translating auditory scenes into some sort of visible scenes, by which the sense of hearing is hoped to be presented to a deaf or hearing-impaired person with as much realism as possible. In this section our method of auditory modelling will be introduced, based upon the latest theories of hearing, as reviewed in the previous section.

Figure 3.8: A simplified model of the peripheral part of the auditory system up to neural firing activity of the inner hair cells. This model is more accurate at moderate sound levels.

affect the cochlear mechanics. They are thought to play a role in the remarkable sensitivity and frequency selectivity of the cochlea. There is not yet, though, much information in literature about the work of outer hair cells.

3.2 Auditory Modelling

The amazing ability of human auditory system to distinguish and recognise sounds has motivated many researchers to study and model its functions of sound analysis in different areas of audio-acoustic processing, such as: speech recognition, speech coding, speech training, time-frequency representations, etc. [1, 22, 92]. Here the auditory modelling is approached from the viewpoint of sound visualisation as an aid for the deaf. From this point of view, we are interested in mimicking the work of the ear as an aid for translating auditory scenes into some sort of visible scenes, by which the sense of hearing is hoped to be presented to a deaf or hearing-impaired person with as much realism as possible. In this section our method of auditory modelling will be introduced, based upon the latest theories of hearing, as reviewed in the previous section.
3.2. Auditory Modelling

3.2.1 Modelling the Outer and the Middle Ears

As discussed in the previous section, the outer ear behaves as a bandpass filter resonating at frequencies around 3 kHz, but the response of the middle ear to the input sound depends non-linearly on the energy of the applied signal. For the sake of simplicity here, it is assumed that the average power of the input to the system is of moderate intensity. The outer and the middle ears have been modelled together as a high-pass filter with the system function:

\[ H(z) = 1 + \alpha z^{-1} \] \hspace{1cm} (3.1)

This filter is often used in speech processing techniques [67] as a preemphasis filter. That is because hearing is most sensitive around the frequency of 3 kHz—as a result of the resonance action of the outer ear. Although this filter raises the frequencies above 5 kHz, the region which is in fact attenuated by both the outer and the middle ears, it has been used for two reasons. The first is that the high frequency regions of the output from the model will be attenuated within the visualiser block of the system—as will be discussed in the next chapter. The second is that the spectrum of voiced sections of the speech signal naturally have a negative slope because of the physiological characteristics of the human speech production system. These attenuations can be offset by the preemphasis filter.

Higher order filters, particularly adaptive ones, might be thought more useful than the above filter; however, its simplicity and the fact that in our experiments we are only interested in the frequency range of 200 Hz to 4 kHz have prompted the use of such a simple and common filter. Figure 3.9 shows the frequency responses of the preemphasis filter, according to relation (3.1), for different values of \( \alpha \). The value of \( \alpha = -0.6 \) was found empirically to be the best figure, based on the quality of the visible output from the system.
3.2. Auditory Modelling

Figure 3.9: The frequency responses of the preemphasis filter used as a model for the cascade of the outer and the middle ears for different values of $\alpha$. The value of $\alpha = -0.6$ has been chosen in our auditory model.
3.2. Auditory Modelling

3.2.2 Modelling the Inner Ear

Modelling the inner ear is the most important part of the auditory modelling as the most complicated analysis of the acoustic input and the mechanical to electrical transduction is carried out within this part. There are several nonlinear functions among the works of the inner ear which contribute to the remarkable sensitivity and the huge dynamic range of the ear. It is believed that the cochlea is an extremely adaptive system and its functions are supposed to be under the control of the central nervous system; however, for a normal situation where the energy of the input sound is at a moderate level, it can be modelled as a rather simple system. In this project it has been assumed that the input to the system is an auditory signal with a moderate level of energy filtered and sampled properly. Although the system has been designed to work for different bandwidths and sampling rates, the evaluating experiments have been carried out on a few spoken words, filtered by a 4 kHz low-pass filter, sampled at a rate of 8000 samples per second, and finally gain controlled.

The mechanical structure of the basilar membrane is modelled as a bank of filters. Any point along the basilar membrane is considered as a bandpass filter with a centre frequency and bandwidth characteristic of its position. The distribution of the centre frequency of the filters along the length of the basilar membrane have been studied by different workers [3, 29]. The general conclusion is that the characteristic frequency of a point is nearly exponentially proportional to its distance from the apex—the point with the lowest resonance frequency. Here, the mechanical displacement of the basilar membrane is modelled by a bank of filters comprising 100 bandpass filters for the range of 200 Hz to 4 kHz. This range of frequency, which resembles the frequency range of a public telephone channel, has been chosen due to the requirements of the experiments; of course the implemented auditory model is capable of covering any range of audible frequency (and also any number of filters) through changing some parameters.

The selected number of filters is much less than the number (about 3500) of inner hair cells sampling the vibration of the basilar membrane, but it seems quite enough
3.2. Auditory Modelling

to represent the mechanical vibration of the basilar membrane with a high resolution.
The frequency-position function suggested in [29] has been used to determine the
central frequency of each auditory filter:

\[ f_c = A(10^{ax} - 1) \]  \hspace{1cm} (3.2)

where \( f_c \) is the characteristic frequency in Hz and \( x \) is the normalised distance of
the corresponding point along the membrane (i.e., \( 0 < x < 1, x = 0 \) being the apex).
The two constants \( A = 165.4 \) and \( a = 2.1 \) have been reported [29] to be appropriate
for the human cochlea.

**Calculation of the Central Frequency of the Filters**

The position of a point, along the basilar membrane, with characteristic frequency
\( f_c \) can be obtained, according to (3.2), by:

\[ x = \frac{1}{\alpha} \log(1 + \frac{f_c}{A}) \]  \hspace{1cm} (3.3)

Assume \( N \) as the number of auditory filters uniformly distributed between the points
\( x = x_0 \) and \( x = x_N \) on the basilar membrane where the apex and the base are
considered to be at \( x = 0 \) and \( x = 1 \) respectively; then the place of the \( i \)th filter will
be:

\[ x_i = x_0 + i \times \frac{x_N - x_0}{N} \]  \hspace{1cm} (3.4)

Suppose \( f_c^{(0)} \) and \( f_c^{(N)} \) to be the characteristic frequencies of the points \( x_0 \) and \( x_N \)
respectively. Then, from (3.3), and (3.4) the characteristic frequency \( f_c^{(i)} \) of the \( i \)th
point can be calculated by the relation

\[ f_c^{(i)} = \alpha \beta^{N} - A \]  \hspace{1cm} (3.5)

where \( \alpha \) and \( \beta \) are defined as:

\[ \alpha = A + f_c^{(0)} \]

and

\[ \beta = \frac{A + f_c^{(N)}}{A + f_c^{(0)}} \]
In this project we have assumed $A = 165.4$ and $a = 2.1$, the appropriate parameters for the human cochlea [29], to obtain the characteristic frequencies of 100 points spread over the interval: [200, 3850] Hz. The use of (3.5) leads to the formula:

$$f_c^{(i)} = 365.4 \times (10.989)^{\frac{i}{100}} - 165.4$$  

(3.6)

where $0 \leq i < 100$. By (3.6) the centre frequencies of 100 band-pass filters are calculated and the filters are designed as follows.

**Auditory Filter Shape**

Auditory filter shapes have been estimated by several groups of workers in different ways [56, 57, 60, 62, 64]). It has been suggested that an auditory filter is a band-pass filter whose bandwidth depends on its centre frequency so that its equivalent rectangular bandwidth (ERB) is related to centre frequency by the formula:

$$ERB = 6.23f_c^2 + 93.39f_c + 28.52$$  

(3.7)

where $f_c$ is the centre frequency in kHz and ERB is obtained in Hz. The shape of the auditory filter as discussed in [63] can be simplified to the formula:

$$W(f) = (1 + pf)e^{-pf}$$  

(3.8)

where $W(f)$ is the frequency response of the filter and $p$ is a constant related to the bandwidth by:

$$p = \frac{4.0}{ERB}$$

and $f$ is the deviation in frequency from the centre frequency of the filter.

Equation (3.8) has been obtained from experiments carried out in ordinary sound pressure levels for which the filter is symmetric around the centre frequency; but this is not the case for higher sound pressure levels. It is believed [58] that as the input sound pressure increases, the shape of the auditory filter changes so that on the low frequency side the sharpness of the filter decreases while, on the other hand, the high frequency side becomes slightly more sharply tuned, resulting in an asymmetric filter shape with a wider pass-band.
3.2. Auditory Modelling

Here, the symmetric shape of the filter represented by (3.8) has been approximated by a digital, linear-phase, FIR filter. Each band-pass filter is obtained by modulation of a cosine signal with a Hanning window, as a low-pass, FIR filter of proper length. The low-pass filter shape used as a basis for all the auditory filters is defined by

\[ h_b(n) = \frac{1}{N-1} \left(1 - \cos \frac{2\pi n}{N-1}\right) \]  

(3.9)

\( h_b(n) \) is the unit sample response of a linear-phase low-pass filter with a delay of \( \frac{N-1}{2} \) points where \( N \) is an odd number.

Some workers have proposed the gamatone and gamachirp functions as the optimal fits to the symmetrical and asymmetrical shapes of the auditory filter, respectively \[36, 101\]. Looking for the best fit, for our purpose, to the shape of the auditory filter defined by (3.8), we concluded that the Hanning window should be the best according to our criteria, which will be discussed later in this section. To calculate the ERB of this filter we start from the definition of ERB:

\[ ERB = \frac{\int_{-\pi}^{\pi} |H(\omega)|^2 \, d\omega}{|H(0)|^2} \times \frac{f_s}{2\pi} \]  

(3.10)

By this formula the ERB is obtained in Hz. From (3.9) we obtain

\[ H(0) = \sum_{n=0}^{N-1} h_b(n) = 1 \]

and from the Parseval's relation \[77\], equation (3.10) reduces to

\[ ERB = f_s \sum_{n=0}^{N-1} h_b^2(n) = \frac{3f_s}{2(N-1)} \]  

(3.11)

For an auditory filter at characteristic frequency of \( f_c \) Hz the unit sample response of the filter is:

\[ h(n) = h_b(n) \cos \left(2\pi \frac{f_c}{f_s} \left(n - \frac{N-1}{2}\right)\right) \]

where \( f_s \) denotes the sampling frequency. Suppose, as an example, we want to model the auditory filter centred at 1 kHz. The bandwidth of the filter, according to (3.7), must be 128 Hz. For this bandwidth, the length of the unit sample response is obtained using (3.11) as follows:

\[ N = \frac{3f_s}{2 \times \text{ERB}} + 1 = 94.75 \]
3.2. Auditory Modelling

Figure 3.10: (a) Frequency response of the model of an auditory filter centred at 1 kHz (solid curve), compared to the one extracted from the formula suggested by Moore and Glasberg [57], (dotted curve). Note the similarity of the two curves in the passband area of the frequency. An optimal model, obtained from the gamatone function (dashed curve) has been also shown for comparison. (b) The impulse responses of the two models.

Rounding this value to the nearest odd number results in $N = 95$. Figure 3.10a shows the frequency response of our modelled filter (solid curve) compared with the shape obtained by the gamatone function (dashed curve) and the one obtained from relation (3.8), suggested by Moore and Glasberg [57] (dotted curve). The impulse responses of the low-pass bases of the two approximations have been shown in Figure 3.10b. Unlike their good similarity in the frequency domain, the two models are very different in time domain. The Hanning-window model has two advantages in comparison with the gamatone model. The first is that it has less
3.2. Auditory Modelling

time duration and has a symmetrical shape and, as a result, it can be implemented with much less computation. The second is its linearity in phase which keeps the phases of the input intact. However, the gamatone model has a low delay which is an advantage in respect to the Hanning window model.

The reasons behind the selection of relation (3.9) as the unit sample response of the auditory filter shape model are as follows:

- The resulted shape is reasonably similar to the one obtained through psychoacoustical and physiological experiments by various workers (Figure 3.10).

- Simplicity of the design of different band-pass filters in practice.

- Small time duration of this FIR filter which results in a higher speed of implementation and gives a lower delay to the signal.

- Linearity in phase. Although there is no evidence about the phase-linearity of the auditory filter, due to the shortage of empirical information in this respect, we have tried to keep the phase unchanged.

Figure 3.11 shows the frequency responses of 10 successive filters out of the simulated auditory filter-bank. This figure illustrates the high degree of overlap (about 78 percent) between the successive filters. The minimum number of filters covering a certain region of frequency can be estimated by the formula [57]:

$$ERB_{rate} = 11.17 \log \left| \frac{f + 0.312}{f + 14.675} \right| + 43.0$$

(3.12)

where $f$ is the upper limit of the frequency region in kHz, while the lower limit is taken as 0. According to this relation, the number of filters which cover the range of 200 Hz to 4 kHz is 22. This region has been simulated by 100 filters which corresponds to 78 percent overlap.

3.2.3 Modelling Mechanical to Electrical Transduction

An inner hair cell is modelled as a rectifier cascaded with a gain control. The intensity threshold of any fibre connected to an inner hair cell is characteristic of
3.2. Auditory Modelling

Figure 3.11: Frequency responses of 10 successive filters of the simulated auditory filter-bank. As is evident, there is a high percentage of overlap (78 percent) between the successive filters.

the fibre. Therefore for a single hair cell different fibres connected to it fire at different phases of the input signal (or do not fire if their threshold is higher than the signal’s maximum amplitude) and they also saturate differently. This is believed to be the cause of the remarkably high dynamic range of the auditory system. In our model, however, a hair cell is simply modelled as a half-wave rectifier. To obtain a sufficient dynamic range when visually representing the output, a gain control and a compressor have been cascaded with the rectifiers as discussed in the next chapter.
3.3 Summary

In this chapter a short description about the anatomy and physiology of the peripheral part of the auditory system, the ear, has been given. The functions of the outer, middle, and inner ears, as reported in current literature, have been reviewed. The complex work of the cochlea, particularly the role of the basilar membrane in spectral analysis of the input sound, has been discussed. According to the above discussions, the outer and the middle ears behave, together, as a band-pass filter. This filter, which is a wide-band filter centred around 3-kHz, can be considered as a high-pass filter for low frequencies (below about 8 kHz). The cochlea, in the inner ear, works like a bank of band-pass filters, each cascaded with a rectifier. Each channel of the filter bank involves several nonlinearities such as half-wave rectification, amplitude compression, bandwidth variability, and even variability of its resonance frequency. However, most of the nonlinearities can be ignored when the input signal is considered to be at moderate sound pressure levels.

Finally a simplified computational model of the ear has been introduced. By this model the works of the outer and middle ears, together, have been simulated by a high-pass, linear filter with a gain control. The length of the basilar membrane has been uniformly sampled to simulate the mechanical vibration of each sample point by a linear band-pass filter. The characteristic frequency and the bandwidth associated to each point are calculated using an empirical formulae suggested in the literature, and a bank of filters is created according to the parameters given to the model. The inner hair cells connected to the basilar membrane are modelled as half-wave rectifiers, and the amplitude compression function of the auditory fibres connected to them has been simplified to a compression formula.
Chapter 4

Auditory Based Visualiser

In the previous chapter a simplified human auditory model was introduced. Using the model, the auditory input is analysed through a bank of filters, the outputs of which are rectified and gain controlled. In this chapter we present a discussion about different methods of visualisation that may be applied to the set of rectified signals.

The goal of visualisation, here, is to demonstrate all the perceptually relevant information embedded in the underlying sound, in real time, in a way that the viewer can understand.

4.1 Sound Perception

What do we hear when we listen? The study of this question lies in the fields of psychophysics and psychoacoustics. In this section we are only interested in some measurable qualities of sound which can be linked to some physical properties of the signal. The most perceptually important qualities of sound are: loudness, pitch, and timbre. A brief review of them is given here.

4.1.1 Loudness Perception

Sounds can be characterised by differences in intensity across time or frequency or, usually, both. That is why intensity perception has been regarded as a basic concern of hearing research. The problem is how sound intensity is represented, or coded, in the human auditory system. One of the important qualities of a sound is its loudness which can be considered as a subjective quantity, substantially related to
4.1. Sound Perception

its physical intensity. The loudness has been defined [53] as:

"That attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from quiet to loud."

Our discussion of loudness is aimed at finding a reliable measure of this auditory sensation, in order that we can translate it into a proper visual representation.

Loudness versus Other Parameters

Psychoacoustic experiments show that the loudness of a sound of fixed physical intensity is affected by its other physical parameters such as its frequency and bandwidth. Also they have revealed how loudness is affected by factors intrinsic to the listener. In this project we are not interested in the differences between listeners or between different situations of the same listener regarding the loudness sensation. But, of course, we need to consider the relation of the frequency and bandwidth of a sound to its loudness.

The frequency dependence of loudness of pure tones has been described through a number of equal loudness contours, Figure 4.1, obtained from various types of experiments [55]. For example, they have been measured by asking listeners to match the loudness of a comparison tone of variable frequency to the loudness of a standard tone at 1000 Hz. The loudness level (in phons) of a tone at any frequency is defined as the intensity level (in dB SPL) of a 1000 Hz tone perceived as have the same loudness [53]. For example a 100 Hz tone which seems equal in loudness to a 60 dB, 1000 Hz tone is taken to have a loudness of 60 phons—although its intensity, as the contours say, is about 70 dB SPL. According to Figure 4.1, low frequency tones, below about 200 Hz, need to be more intense than high frequency tones (up to about 6 kHz) in order to have the same loudness. This phenomenon can be linked to the high-pass filtering of the outer ear as discussed in chapter 3.

Loudness of a sound is also affected by its bandwidth and its temporal factors. Experiments have revealed that loudness of a band of noise with a fixed total intensity increases as its bandwidth grows—despite the reduction of its spectral amplitude.
4.1. Sound Perception

Figure 4.1: **Equal loudness contours.** For each specified loudness, the intensity of a tone required to produce the given loudness has been depicted versus its frequency. Redrawn from [55]

This is discussed in the next section. For sounds of a short duration, such as tone bursts, their loudness also depends on their time duration [55]. It seems that over a certain range of duration, the ear behaves as if it were an energy integrator. That is, over this range of duration loudness of a sound is determined by its energy rather than its power.

**Loudness Scale**

Although the equal loudness contours illustrate the growth of loudness with intensity and, for example, show that the slope of this growth is higher for low frequencies than for high frequencies, they cannot provide a numerical measure of the slope. Different techniques have been used to scale the loudness [53], the most straightforward of which is *magnitude estimation*. By this technique the listener is simply asked to
4.1. Sound Perception

assign a number to each sound in a set of sounds having different intensities. It has been suggested [96], that loudness, \( L \), is a power function of intensity \( I \):

\[
L = k I^{0.3}
\]  

(4.1)

where \( k \) is a constant depending on the listener and the units used. This function implies that a 10 dB step in intensity corresponds to an approximately twofold change in loudness.

4.1.2 Pitch Perception

The second perceptually important attribute of sound is its pitch. Pitch has been defined as [53] :

"That attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from high to low."

Like loudness, pitch is a subjective attribute, which cannot be represented simply by physical means. It is related to the repetition rate of the waveform of a sound. For a pure tone, pitch corresponds to the frequency of the tone, but for a complex harmonic sound it is proportional to its fundamental frequency. [49]

Pitch of Pure Tones

The pitch of a pure tone depends on its frequency. The pitch-frequency relation has been estimated through various experimental methods [34, 53] resulting in the classical, mel scale, as shown in Figure 4.2. The mel unit has been defined so that the pitch of a 1000 Hz standard tone has a pitch of 1000 mels. A tone that sounds twice as high is assigned a value of 2000 mels, and a tone sounding half as high is given a value of 500 mels. According to the resulting curve the mel scale is similar to the frequency-position function of the basilar membrane (equation 3.3), or the critical-band scale for the frequency resolution in the inner ear [54]. This relation implies that pitch is determined by the centre of excitation activity along the basilar membrane.
Although pitch is substantially determined by the frequency of the stimulus, it is also affected by other parameters. The perceived pitch of a pure tone is changed if its intensity is changed. The scale of this change is different for different frequencies, but in any case, it is smaller than 20 percent of the perceived pitch, even for very high changes of intensity [34]. There are some other factors affecting pitch of a sinusoidal tone, such as the presence of other tones or noise in the background.

**Pitch of Complex Tones**

A complex tone is composed of several sinusoidal tones (harmonics) with frequencies which are multiples of the fundamental frequency of the complex tone—the repetition rate of its waveform. Pitch of a complex tone is perceived as equal to the pitch of a pure tone having the same frequency as the fundamental frequency of the complex tone. The question of which harmonics of a complex tone contribute to the sensation of pitch has long been a subject of debate. Although the conclusion
4.1. **Sound Perception**

From various experimental results reveals, to some extent, the solution, the problem is still under investigation. Related theories are discussed in more detail during the next section.

When we are exposed to a number of simultaneous complex tones, such as musical sounds, we can easily perceive pitches of different instruments at the same time. This is a remarkable feature of auditory sensation which is a matter of research in hearing theory.

### 4.1.3 Timbre Perception

The third perceptually important attribute of sound is *timbre*. Timbre has been defined [35] as: The quality of a sound by which a listener can tell that two sounds of the same loudness and pitch are dissimilar. Actually, this definition tells us *what timbre in not*, instead of, *what timbre is*. According to this definition, timbre depends upon many physical parameters of sound, such as the spectrum, the waveform, the frequency location of the spectrum, and the temporal characteristics of the stimulus. To simplify the issue we accept here the restricted definition of Plomp [71]:

> "That attribute of sensation in terms of which a listener can judge that two steady complex tones having the same loudness, pitch, and duration are dissimilar."

According to this definition, timbre depends on the relative magnitudes and phases of the harmonics of the tones.

Psychoacoustic experiments show [55] that timbre is primarily determined by the magnitudes of different harmonics, although their phase pattern also plays a small role [13, 61, 70, 72, 73].

**Dimensionality of Timbre**

The two attributes described so far, *i.e.* loudness and pitch, have been considered as unidimensional; that is, sounds of different pitches can be ordered on a single scale.
of pitch extending from low to high. Similarly, sounds of different loudnesses can be ordered on a single scale going from quiet to loud. For timbre, however, there is no single scale along which we can order timbres of different sounds. A steady complex sound can be defined as follows:

\[ s(t) = \sum_{n=1}^{N} a_n \cos(2\pi n f_0 + \phi_n) \]  

(4.2)

Generally speaking, \( s(t) \) may have \( 2N \) dimensions as \( a_n \) and \( \phi_n \) can be selected independently. Plomp [71] presents a thorough discussion about the number of dimensions sound may have according to (4.2) with the restrictions dictated by the nature of audible sounds and the functions of the auditory system. It is concluded [71] that, excluding very low and very high frequencies, the number of dimensions can be reduced to a value of about 15. For a restricted set of sounds, e.g. vowel-like sounds, the number of dimensions involved may be much smaller. The timbre differences between the vowels of speech sounds, for example, may be described satisfactorily by only 3 dimensions.

Our problem is how the timbre of a real sound can be demonstrated visually. This problem will be discussed in section 4.3.

4.2 Speech Perception

A given speech sound cannot be represented by a fixed acoustic pattern because its acoustic pattern varies in a complex manner according to preceding and following sounds. Even a missing part of a word may be 'heard', because of its high probability of existence in the context of a sentence; this is especially evident when we are listening in a noisy environment. On the other hand, unknown words, if clearly articulated, can be recognised in isolation, that is, by only acoustic cues. We shall concentrate here on the perceptual processing of patterns in the acoustic waveform.

4.2.1 The Nature of Speech Sound

Words are broken down into smaller units called syllables which can in turn be analysed in terms of phonemes—the speech sounds. Phonemes are defined in terms
4.2. *Speech Perception*

of perception, rather than in terms of acoustic patterns. Although some recent linguists and psychologists deny phonemes to be the basic units of speech perception [51, 105], the analysis of speech in terms of phonemes has been widely used and is helpful for our present purposes, so we shall use the concept for the purpose of discussion. A simple view would hold that speech is a composition of acoustic patterns which have a one-to-one correspondence with the phonemes. A review of the way in which speech sounds are produced reveals that, unfortunately, this view is not tenable.

Speech sounds are produced by the vocal organs, composed of the vocal folds and vocal tract. Vocal folds are put into vibration and affect the flow of air from the lungs, resulting in a complex tone with a relatively low fundamental frequency whose harmonic spectrum covers a wide range of frequencies. This spectrum is then modified by the vocal tract, which behaves like a filter imposing resonances called *formants* (or anti-resonances) at certain frequencies. The shape of the vocal tract determines the centre frequencies of the formants. They are numbered according to their order on a frequency scale: the first formant (F1), the second formant (F2), and so on. Speech sounds are, mainly, classified as: vowels, fricatives, stops, affricates and nasals. Except for vowels, which are relatively stable over time, other phonemes are mainly produced by a narrowing or constriction action of the vocal tract at some points. They differ in the place of narrowing, the degree of constriction, and the starting time of the vibration of the vocal folds. Voiced consonants are those accompanying the vocal tract vibration, such as /z/, /b/, /d/, and /g/, for which their unvoiced counterparts are: /s/, /p/, /t/, and /k/, respectively. In the case of the unvoiced sounds, vibration starts after the complete release of the vocal tract constriction.

The differences of the speech sounds are reflected in the acoustic characteristics of the speech wave. Hence, speech comprises acoustic patterns which vary in frequency, intensity, and time.

Figure 4.3 illustrates the temporal variations of the spectrogram of the utterance
4.2. Speech Perception

'sky noise', uttered by a female speaker. In this spectrogram, harmonics below about 1.5 kHz (up to the 6th harmonic) are resolved and four formants are distinguishable at some points of time; as an example, at time 0.25 seconds the four formants have been indicated in the figure. The horizontal black lines are the resolved harmonics. It is, generally, believed that the most important feature of a vowel is the set of its formant frequencies, and not its harmonics. Variations of the frequency of a certain formant or set of formants is a marked characteristic of some speech sounds. These variations, known as formant transitions, reflect the motion of the articulators during speech. Fluctuations of the tongue during the utterance 'sky noise', for example, seem to be reflected in the oscillation of its second formant, as can be seen in Figure 4.3. Indeed, for some speech sounds, such as stop consonants, the formant transitions are the most important cues for their identifications.

Acoustic properties of a speech sound are also influenced by the preceding and following sounds. Figure 4.4 shows three instances of the vowel /oi/ in the words 'boil', 'boys', and 'noise'. Note the coarticulation effect of /oi/ in connection with different consonants /b/, /l/, /n/, and /z/. It is the motion of the formants within the relatively long duration of the phoneme /oi/, in 'boil' for example, that is con-
sidered as an important cue for the identification of the short sounds /b/ and /l/.

### 4.2.2 Special Mode of Speech Perception

Some researchers have argued that the perception of speech differs from the perception of nonspeech sounds. They argue, for example, that the rate of 30 phonemes per second, which may occur in speech, is too fast for resolution in the auditory system, and conclude that speech must be a special code which needs a special decoding mechanism. They believe, further, that some phonemes such as vowels are not completely encoded, compared with consonants which are highly encoded. The opposite argument is that experiments show that listeners can learn to identify sequences of sounds of as short as 10ms duration, that is, potentially 100 units per second. Therefore speech can be considered as a sequence of acoustic patterns which although they might not all be perceived separately, the listener can learn to recognise their overall sound as a group of acoustic patterns representing, for example, a syllable.

There has also been a discussion about a *special mode* of perception of speech [53, 78]. It is suggested that when we listen to natural speech, our perception state is triggered to *speech mode* and then it becomes impossible to hear the sound sequences in terms of its acoustical characteristics; rather, we perceive a unified stream of speech sounds. The evidence for the existence of a special mode of perception, and the fact that certain areas of brain are specialised for dealing with speech sounds, support the idea that speech perception is special. However, listening in a specific perceptual mode is not unique to speech perception: for example, a harmonic complex tone can be perceived in two distinct ways. We may listen analytically or synthetically, *i.e.* perceive it as a number of partials with different pitches, or as a single pitch equal to its fundamental frequency, respectively [53].

### Audiovisual Interaction

It is believed that there are interactions between auditory and visual perception of speech [100], that is, that the movements of the speaker's face and lips may strongly
Figure 4.4: Auditory-based spectrogram of the utterances 'boil', 'boys', and 'noise'. Compare the different patterns of the same phoneme /oi/ in association with different consonants.
influence a listener's perception. An interesting series of experiments on this subject has been reported by McGurk and Macdonald [50]. One of the experiments, for example, was as follows: Articulation of two similar words such as 'tata' and 'mama', spoken by a person, were recorded on a videotape and then rearranged such that the audio recording of 'mama' was synchronised with the video recording of 'tata'. Then the subjects were asked to both listen and watch the audiovisual recording. Unaware of the conflict between auditory and visual cues, most observers said that they heard the sound 'nana'; they were surprised when they closed their eyes and heard 'mama'. It is said that the acoustical and optical information are combined in a complex manner which is not always easy to explain. These experiments confirm the idea that we make use of other sources of articulatory information in the perception of speech sounds.

4.2.3 Models of Speech Perception

There are many models of speech perception, the two most influential of which are mentioned here. The first model, known as the 'motor theory' [45, 53] of speech perception, holds that the listener perceives the articulatory gestures intended by the speaker when producing an utterance. This theory also holds that speech perception and speech production are closely related through an innately specified link. Thus, perception of the intended gestures takes place in a specialised speech mode in which a conversion is made from acoustic signal into articulatory gestures. This model, however, does not specify how this conversion is accomplished.

The second model, being very different, proposes that there is a sequence of stages of processing as follows: The speech signal is first analysed in the peripheral part of the auditory system. This analysis including functions such as filtering, rectification, adaptation and phase locking, emphasises some of the less variable characteristics of phonetic features and suppresses irrelevant variability [14, 15, 16, 17, 18]. The second stage is an array of acoustic property detectors such as onset detectors, spectral transition detectors, formant frequency detectors, and periodicity detectors.
4.2. Speech Perception

The third stage is an array of phonetic feature detectors that make use of acoustic property values to decide whether a specific phonetic feature, such as voicing or nasality, is available. This stage is language specific, that is, the feature detectors are tuned to the phonetic disparities of the language in question. Finally there are stages of lexical and semantic searches in which the context is used to compensate for the incorrect decisions at previous stages.

4.2.4 Acoustic Cues and Redundancy in the Speech Wave

The theory underlying the analytic model above assumes that it is possible to find a relatively invariant mapping between acoustic patterns and perceived speech sounds. According to this theory, even consonants, which are said to be highly encoded, are accompanied by some invariant cues. Moreover, there are some context-dependent cues present in all consonant-vowel syllables which serve to identify the phonemes involved.

It is believed that any given consonant can be uniquely defined in terms of a number of features. Some classes of sounds are characterised by a rapid change in spectrum. Rapid change in amplitude is also a characteristic of some classes of phonemes. A number of speech sounds such as the stop consonants /p, t, k, b, d, g/ are characterised by rapid changes in both their spectra and amplitudes. Another important, invariant cue serving to distinguish between certain pairs of speech sounds is voicing, that is, the presence of periodicity in their waveform. For example, /b/ is voiced while /p/ is unvoiced. The gross shape of the spectrum is another distinguishing property of stop consonants, although dynamic changes in their spectra are believed to be more important [39]. Waveform envelope of the speech signal is also a source of information which can serve as a useful cue to distinguish voiced from unvoiced sounds, consonants from vowels, or unvoiced stops /p, t, k/ from other consonants.

In fact, in the perception of speech the listener uses many different types of information which are available in the speech wave. The multiplicity of acoustic
4.3 Static Spectrogram

Cues allows for a high level of redundancy in the speech wave. Several different acoustic cues are available for a given phoneme when just one or two of them might be sufficient for recognition. This redundancy together with the effects of context serve to overcome the ambiguities inherent in speech. That is why speech is so remarkably resistant to many kinds of quite severe distortions, such as background noise, spectral filtering, peak clipping, and interference from other sources of sound [53, 55].

4.3 Static Spectrogram

In chapter 3, *Auditory Process and Modelling*, a simple model of the human auditory system was introduced. Output from the model, as shown in Figure 3.8, is a collection of rectified, narrow-band signals. Each signal is considered as a simplified model of the neural activity of an auditory fibre connected to a certain point along the basilar membrane. The modelled points have been uniformly selected along the length of the basilar membrane, resembling the form of distribution of the cochlear hair cells.

Now the question is how to make an image, using the output signals, to visualise (by visualise we mean here to show visually), as accurately as possible, all the perceptually important features of the input sound. In this section, we consider a limited time interval of the input sound and try to make a static image using the corresponding output. Techniques for visualisation of the temporal variations of this image, in real time, will be discussed in later sections. Theories of hearing, together with the current knowledge of sound perception, as discussed in the previous section, are considered for making an informative image of the input sound. Of course, the design of a good, visually perceptible image of sound requires a good knowledge of the human auditory system together with a reasonable understanding of the human visual system, detailed discussions of which are beyond the scope of this project.

A static spectrogram, as discussed here, is a rectangular image which shows the auditory information in three dimensions: time, frequency, and energy. This
4.3. Static Spectrogram

A 40-millisecond frame of the cochleagram (a static spectrogram) of the vowel /ai/ from the word 'sky', uttered by a female speaker.

Visualisation of the three most perceptually important features of sound have been carried out as follows.

4.3.1 Visualisation of Loudness

The intensity of sound has been displayed as a gray level scale in the spectrogram image. Loudness in auditory perception is considered as the counterpart of brightness (or darkness) in visual perception [37]. Therefore, it is quite reasonable to represent the sound intensity, the source of loudness, by a gray level, as is also the case in traditional spectrograms.

Intensity is coded, in the auditory periphery, by the firing rate of the auditory fibres innervating the inner ear [69]. Any hair cell mounted on the basilar membrane behaves like a rectifier. The rectified signal, which is in effect an electrical voltage induced in the hair cell [12], activates any nerve fibre attached to the cell, provided that the amplitude of the voltage exceeds the specific threshold of the fibre. An auditory fibre is activated at a specific threshold and its firing rate increases as the...
amplitude of the signal increases, but saturates after a specific, higher level of the stimulating signal (see Figure 3.7). Different fibres attached to a single hair cell are activated at different thresholds and saturated at different levels. Hence, the central auditory nervous system has sufficient information to estimate the rectified signal induced in any hair cell.

In this model, the intensity is encoded independently for different frequency channels (i.e. different hair cells). We have depicted this information, for each frequency, in the form of a horizontal line which represents the rectified signal, if present at that channel, by a gray level scale. As shown in Figure 4.5, the overall darkness of the image, when perceived by the visual system, represents the loudness of the underlying sound.

The processes of threshold and saturation of a nerve fibre have been modelled by the relation:

\[ y = \frac{m}{1 + de^{-ax}} \]  

where \( x \) represents the action potential induced in the hair cell, \( y \) is the firing rate of the fibre (number of spikes per second), \( d \) is the dynamic range of the output and \( m \) is the maximum amplitude of the output signal. In this model \( a \) is determined by

\[ a = \frac{\log d}{x_s} \]

where \( x_s \) is defined as the value of \( x \) for which the output equals half of its maximum.

In fact this simple model is considered as including the way in which the rectified output of a hair cell is coded by the overall contributions of different fibres connected to it. In the static spectrogram, as shown in Figure 4.5, the darkness of any point is determined by the value of \( y \) in (4.3), where \( x \) is the corresponding channel's output at the corresponding time.

Another type of compression, attributed to the discharge activity of auditory nerve fibres, is related to the mechanical nonlinearity of the basilar membrane [25]. We will refer to this later.
4.3. Static Spectrogram

4.3.2 Visualisation of Pitch

As discussed in previous section, pitch is related to the periodicity of the speech waveform. Several explanations have been proposed as to how the human auditory system perceives the pitch of a single or complex tone. Here three classes of theories are outlined:

Rate/Place\(^1\) Representation

It has been suggested [88, 89] that a spatial profile (i.e. frequency profile) of the average firing rate over all classes of auditory-nerve fibres could be a well defined measure of the spectral pattern of a speech sound at low sound pressure levels. But for higher sound pressure levels, as the high spontaneous-rate fibres saturate, only the low and medium spontaneous-rate fibres can be considered as the basis of spectral representation. According to this model, therefore, the intelligibility of speech is predicted to decrease for higher sound pressure levels, as a result of the reduction in the number of contributing fibres, but yet it improves according to intelligibility tests [26]. Moreover a rate/place representation may fail to show a sufficiently high resolution in frequency.

According to this theory, the pitch of a complex tone must be obtained from the spatial (i.e. frequency) intervals between the harmonic peaks along the speech spectrum (e.g. along a vertical line at a given time of the cochleagram depicted in Figure 4.5). But this model cannot explain how the pitch of a complex sound could be extracted from the unresolved, high frequency portion of the spectrum, as is actually the case for high-pass-filtered speech sounds [53]. In the case of high-frequency, single tones, however, it seems that this type of representation is the only model to account for the perception of their pitch.

\(^1\)“Place” here means position along the basilar membrane of the cochlea—i.e. characteristic frequency (see section 3.1.2)
4.3. Static Spectrogram

Temporal/Place Representation

In this model the temporal pattern of firing activity of each fibre is analysed. For this analysis, the model requires that the system have some knowledge of the characteristic frequency of the fibre in question [26]. Dynamic activity in terms of rate is considered only for those fibres for which their activity is synchronised to their characteristic frequency [88, 91]. Pitch perception, by this model, can be more accurate than through using only the rate/place model, because the way firing activity changes with time is also included in frequency computations. For example, Figure 4.5 shows that the periodic firing pattern of the high frequency fibres alone contains sufficient temporal information to indicate the period of the input speech waveform uniquely.

Temporal/Non-place Representation

Some researchers [21, 22] have suggested that the temporal activity of an auditory fibre regardless of its place information (i.e. its characteristic frequency) contains enough information about the periodicities inherent in the signal. At medium and high sound pressure levels, particularly, temporal information of intensive spectral peaks, such as formants, will be distributed over a wide range of frequency channels. This large amount of redundancy regarding the formant frequency could prove useful in the presence of noise. This model, however, fails to resolve high frequency channels (more than about 4 kHz), because neural phase-locking falls off as frequency increases. By this model, the overall periodicity of a sound waveform, which is typically less than 1 kHz (and less than 400 Hz for the human voice), can be inferred from nearly all the active channels (see Figure 4.5).

None of these theories can solely explain the amazing ability of human pitch perception in different situations. It seems that the brain works differently for different types of auditory stimuli [53, 55] when perceiving their pitch. When a single tone is in question, for low-frequency sounds the temporal information is used, while for high frequency stimuli only the place (cochlear position) information
4.4. Correlogram Movie

seems to be informative. But for a complex tone, depending on its pitch period and the distribution of energy over its spectrum, either the temporal or the spatial information or both of them could be used to estimate the pitch period.

Perception of pitch by the auditory system is in some ways like perception of space by the visual system [8, 37]. In a static spectrogram, both temporal and spatial information are mapped onto spatial scales. Temporal information is depicted horizontally, and place (i.e. frequency) information has been shown vertically. Distribution of frequency along the vertical axis of the spectrogram image is similar to the distribution of characteristic frequencies of auditory nerve fibres along the basilar membrane, and this in turn is in accord with the mel scale measure of pitch perception as obtained through psychoacoustic experiments (Figure 4.2). The horizontal presentation of temporal information is, however, uniform. But this is also similar to the relationship between perceived pitch and changing frequency for low pitch sounds (below 1 kHz).

4.3.3 Visualisation of Timbre

Timbre as defined in section 4.1.3 is a multidimensional attribute of sound. The gray level pattern of the spectrogram image reflects all of the timbre information of the sound. Harmonics are depicted as dark lines in the image and formants show in the form of maximally intensive harmonics. Also the temporal pattern across different frequency channels can give information about timbre. In the work that follows, the effects of timbre will be included in the general tests.

4.4 Correlogram Movie

The static spectrogram is an informative image, rich enough to illustrate nearly all the perceptually relevant information embedded in a speech sound. But it has a problem in that it cannot display the temporal variations of the input sound in a temporal manner. Most speech processing systems assume that the perceptual characteristics of a speech sound, such as pitch and formant frequencies, are relatively
4.4. Correlogram Movie

stable within a time interval of 20 milliseconds. Upon this assumption a reasonable method for visual representation of sound, in real time, could be sampling its perceptual information every 20 milliseconds (i.e. 50 samples per second) and displaying them, frame by frame, synchronised with the input sound. The rate of 50 images per second is equal to or faster than the frame rate of standard movie films and television systems which, therefore, is well established as an acceptable compromise for our visual systems.

A very interesting technique for displaying some features of sound, called the correlogram, has been proposed by Slaney and Lyon in [92]. A correlogram represents sound as a three-dimensional function of time, frequency, and periodicity. A correlogram, as proposed in [92], is a sequence of rectangular, digital images displayed one after another, resulting in a motion picture. The images are constructed from the autocorrelation functions of the outputs from an auditory model.

We have applied this technique to our model of the auditory system as follows. The output from each cochlear channel, that is from each rectifier and gain control (Figure 3.8), is applied to a 1-kHz, low-pass filter. At a given point of time, short time autocorrelations of the output from the filters are computed and then mapped onto level functions which are depicted as horizontal lines of a digital image. The updating of this image, every 20 milliseconds, leads to a cinematograph type image which has been called the correlogram movie [92].

4.4.1 Computation of Correlogram

The autocorrelation function of $x(t)$ is defined as:

$$ R_x(\tau) = \int_{-\infty}^{+\infty} x(s)x(s-\tau) \, ds $$

Here, however, we are interested in a short segment of the signal ending at time $t$, that is, a window of the signal between $t-T$ and $t$, defined by

$$ y(s,t) = w(t-s)x(s) $$
where, \( w(t) = 0 \) for \( t < 0 \) and \( t \geq T \). Therefore

\[
R_y(\tau, t) = \int_{-\infty}^{+\infty} y(s, t) y(s - \tau, t) \, ds = \int_{-\infty}^{+\infty} w(t - s) x(s) w(t - s + \tau) x(s - \tau) \, ds
\]

This can be rewritten as

\[
R_y(\tau, t) = \int_{0}^{T} w(s) x(t - s) w(s + \tau) x(t - s - \tau) \, ds 
\]  

(4.4)

Likewise, the short time autocorrelation of a discrete signal, say \( x(m) \), at time \( n \), can be computed by the formula:

\[
R_x(j, n) = \sum_{m=0}^{N-1} w(m) x(n - m) w(m + j) x(n - m - j)
\]

(4.5)

where \( R_x(j, n) \) represents the autocorrelation of a window of \( x(n) \), windowed by \( w(n) \), in terms of \( j \) —the autocorrelation lag. The window \( w(n) \), here, is a rectangular window of length \( N \).

Equation (4.5) has been applied to the output from each channel of the auditory model, at equal time intervals of \( n = kN \). Then the level function of the \( i \)th line of the \( k \)th image has been computed using the corresponding normalised autocorrelation function

\[
C_k(j, i) = \frac{R_x(j, kN)}{R_x(0, kN)} , \quad 0 \leq j < N, \; k = 0, 1, 2, \ldots
\]

(4.6)

where \( C_k(j, i) \) is the normalised autocorrelation of the \( i \)th channel for the lag of \( j \).

To map the relative power of each channel in a way that the loudness scale (4.1) is satisfied, the autocorrelation function has been scaled by a power of the energy as

\[
g_k(j, i) = E_{ki}^{0.3} C_k(j, i)
\]

(4.7)

where \( E_{ki} = R_x(0, kN) \) is the energy of the \( i \)th channel for the \( k \)th frame of the signal and \( g_k(j, i) \) is the gray level of the point \((j, i)\) of the \( k \)th frame. This method of scaling also reduces the dynamic range of energy which is required for a proper display. To find the best scaling factor, we need also some knowledge about the intensity perception of the human visual system which, however, has not been considered here.
4.5. **Dynamical Spectrogram**

Experimental results show that a symmetric image is much easier to perceive than the image produced by using (4.7) for $0 \leq j < N$. Therefore the autocorrelation lags have been selected within the symmetric interval of $\left(-\frac{N}{2}, \frac{N}{2}\right)$ that results in a symmetric image of the periodicity of each channel. To improve the display, the level obtained by equation (4.7) has been compressed according to relation (4.3).

Figure 4.6 shows a sequence of correlogram images, computed by this technique, of the sound /ai/ the static spectrogram of which has been depicted in Figure 4.5. The images have been produced at a rate of 50 images per second and each image represents a 20-millisecond frame of the input sound.

### 4.4.2 Shortcomings of Correlogram

Loudness and pitch can be fully represented by a correlogram but timbre is not illustrated in full detail. Only the periodicity of the output from each channel is displayed through the correlogram; that is, relative phase information is totally lost. Thus the relative phase of resolved, low-frequency harmonics as well as the fine time structure of the outputs from high frequency channels are lost. Psychoacoustic experiments, however, have shown [61] that the the relative phases of the harmonics of a complex tone affect its timbre. Hence a comprehensive representation of sound is expected to show the effect of phase changes in the spectrum of an ongoing sound. To create a more informative visualisation, we have devised a new technique of animation which will be discussed in the next section.

### 4.5 Dynamical Spectrogram

In this section we are looking for a technique to display all the information of the static spectrogram in a dynamic form synchronised with the input sound. The output from the auditory model is a sliding image, called the cochleagram, which is like a moving conveyor belt synchronised with the ongoing input sound. But this belt moves too fast to be followed with the eye. Suppose, for example, the image depicted in Figure 4.5 is moving with a speed of 2.5 metres per second from right
4.5. Dynamical Spectrogram

Figure 4.6: A number of successive, 20-millisecond frames of the correlogram of the vowel /ai/ from the word 'sky', uttered by a female speaker. When animated, these images display the motion of the periodicities of the sound for different frequency regions. Images have been displayed in the order of top to bottom.
4.5. **Dynamical Spectrogram**

to left and we want to catch the fine structure of the image; it is very difficult to pick out its resolved information properly. Moreover a sliding image is less easy to watch than a fixed one, because the eyes need to lock and relock to the image, so that frequent jumps of the eyes are required when watching a long, moving image of this type. Iterative locking of the image within a fixed image, instead, can help the viewer to watch without any eye movement.

As discussed in section 4.3.2 the temporal patterns of the activities of the auditory nerve fibres as reported by different workers [27] together with the evidences obtained from psychoacoustic experiments [55, 61, 70, 72] suggest that a fully informative representation of sound must be able to show the fine time structure of the outputs from different auditory channels. In this project, therefore, we have tried to apply processing principles that are motivated by observed properties of auditory nerve activities to develop a new method of representation which we call *dynamical spectrogram*.

**4.5.1 Phase Locking Evidences**

There is an important difference between the properties of the firing patterns of low frequency and high frequency fibres. At low frequency channels (say, below 1 kHz), harmonics are precisely resolved and neural discharges of the connected fibres are phase locked to the corresponding harmonics. At high frequency channels, on the other hand, due to the mechanical properties of the basilar membrane the frequency resolution is reduced; and due to the firing properties of auditory-nerve fibres, phase-locking of the discharges is also reduced. The instantaneous rate of firing, however, delivers the temporal information with fine time resolution [14, 15, 16, 17, 18, 89, 107].

Figure 4.7 shows the outputs from two low and high frequency channels of the auditory model, one at characteristic frequency (CF) of 236 Hz and the other at 2930 Hz. For the low CF channel the neural firing activity has been shown as a half wave rectified sine wave (solid curve) which is in synchrony with the basilar membrane
4.5. Dynamical Spectrogram

Figure 4.7: Cochlear vibration (dotted curves) and firing pattern of auditory nerve fibres (solid curves), as represented by a very simplified model of the auditory system for a low CF channel (236 Hz) and a high CF channel (2930 Hz).
vibration (dotted curve). For a high CF channel, however, the firing action is too slow to follow the high frequency vibration and then only a nonuniform sampling of the envelope of the wave is delivered by related nerve fibres. This behaviour has been modelled by low-pass filtering with a cut-off frequency of 800 Hz. Therefore the output from a high frequency channel of the model (solid curve) represents the temporal variations of the intensity of the basilar membrane vibration.

Because of the wide bandwidth of the auditory filters at high characteristic-frequency channels, their outputs contain more than one harmonic of the input sound and consequently their waveforms are like amplitude modulated signals. The envelope pattern of such a signal depends on the relative amplitudes and phases of its individual components. Therefore, successive auditory channels in a given area of high frequency may deliver different patterns. Psychoacoustic experiments confirm the idea that the envelope patterns of different channels together with their relative phases determine the perceived timbre of a sound [61, 71, 72].

Therefore, a more informative visualisation of sound, compared with the correlogram movie, may be a frame-locked animation of short time frames of its static spectrogram—as introduced in section 4.3. This type of animation leads to a still picture when the input to the visualiser is a steady-state sound. This picture, of course, shows all the information of the static spectrogram in the form of a rectangular image. For a changing sound, however, its long-term variations will be mapped onto a temporal dimension, resulting in a visual event which is very similar to the underlying auditory event.

4.5.2 Frame-Locking Procedure

Utilising the periodicity of the image, we can make an animation of a series of frame-locked, short-time pictures properly taken from the sliding image—the cochleagram. The animation is created as follows: A 20-millisecond frame of the cochleagram is shown on a screen for a duration of 20 milliseconds. The screen is iteratively updated every 20 milliseconds—50 iterations per second. At each iteration a new frame of the
cochleagram is selected out of its most recent 40 millisecond duration in a way that the selected frame, which is going to be depicted on the screen, has the maximum similarity with the current image on the screen. The measure of similarity, a value between 0 and 1, will be discussed later.

Assume $G(x,y)$ as the gray level of a point $(x, y)$ of a rectangular image constructed from the most recent 40 milliseconds duration of the cochleagram image (as in Figure 4.5) containing $M$ rows and $2N$ columns. Now consider an image of 20 milliseconds duration defined as:

$$g_i(x, y) = \begin{cases} G(x, y), & i \leq x < N - i \\ 0, & \text{otherwise} \end{cases}$$

and consider $g_p(x, y)$ as the image currently depicted on the screen. Let us define $\rho_i$ as the value of similarity between $g_p$ and $g_i$. The next image to be depicted on the screen, then, is $g_i$ for the value of $i$, $0 \leq i < N$, in terms of which $\rho_i$ is maximum. This image will be shown on the screen for 20 milliseconds of time after which $G(x,y)$ will be updated for the next iteration.

Figure 4.8 shows a few successive frames of the dynamical spectrogram of speech sound /ai/ from the word 'sky', spoken by a female speaker. The three perceptually important qualities of sound, such as loudness, pitch, and timbre, demonstrated with a high resolution, are vividly perceptible when they are animated on a screen. Particularly, the formant transitions of this phoneme, that is the divergence of formants F1 and F2, a major cue to the recognition of this phoneme, is clearly presented to the viewer.

**Similarity Measure**

The absolute similarity between two images may be defined by the correlation between them as follows:

$$\text{corr}(g_1, g_2) = \frac{2 \sum_{x,y} g_1(x, y) g_2(x, y)}{\sum_{x,y} g_1^2(x, y) + \sum_{x,y} g_2^2(x, y)}$$

where $g_1$ and $g_2$ are the gray level functions of the images defined for any point $(x, y)$ of the respective image. $\text{corr}(g_1, g_2)$, the normalised correlation, is 1 (its maximum)
4.5. Dynamical Spectrogram

Figure 4.8: Dynamical spectrogram of speech sound /ai/ from the word 'sky', uttered by a female speaker. Note transitions of the formants F1 and F2—the lower and upper bands of enhanced energy (darkening) across groups of frequency channels, which move apart as the sequence of frames is descended. These transitions are clearly perceived when the images are displayed, one by one, on the same screen with the speed of 50 images per second.
4.5. Dynamical Spectrogram

only when the two images, \( g_1 \) and \( g_2 \), are the same.

For our purpose, however, if two images are not the same but their gray level functions are proportional, that is,

\[
g_1(x, y) \propto g_2(x, y)
\]

they are still assumed maximally similar. Therefore, we have defined another measure of relative similarity (Pearson correlation) as follows:

\[
sim(g_1, g_2) = \frac{\sum_{x,y} g_1(x, y) g_2(x, y)}{\sqrt{\sum_{x,y} g_1^2(x, y) \sum_{x,y} g_2^2(x, y)}} \quad (4.8)
\]

This value, is 1 when \( g_1(x, y) \propto g_2(x, y) \).

4.5.3 Advantages of Dynamical Spectrogram

In contrast with the correlogram movie, the dynamical spectrogram demonstrates both the fine time structure at high-frequency channels, and the relative phases of the resolved harmonics at low-frequency channels. An interesting signal to show the ability of phase perception of the ear is a harmonic, complex tone, one component of which has a continuous running phase [70]. Such a signal may be represented by the relation

\[
x(t) = \cos(2\pi ft) + \cos[2\pi(2f + \Delta f)t] + \cos(2\pi 3ft)
\]

giving a continuous phase shift of \( 2\pi \Delta f \) radian per second. This type of sound gives rise to a beat sensation during perception, which is considered to be caused by timbre fluctuations. We have produced a harmonic signal comprising 9 harmonics (2nd till 10th) with its odd harmonics having a phase shift of \( 4\pi \) radian per second. The correlogram and the dynamical spectrogram of this synthesised sound have been shown in Figure 4.9. As one descends the dynamical spectrogram sequence, the phase shift of the resolved odd harmonics together with the changes in the time patterns of high frequency channels are clearly shown. The correlogram, however, does not show any change in phase of the resolved harmonics at low-frequency channels; nor is the change of time patterns of high frequency channels visible. When
displaying the correlogram in real time only some slight fluctuation show in the high frequency area of the image.

Another advantage of frame-locked presentation of the cochleagram is the ability also to show non-harmonic periodicities when these are present in the input sound. This is because the animated image is phase-locked to the dominant periodicity of the signal and, as a result, non-harmonic sounds show in the form of sliding features in the background.

4.6 Speech Visualiser

To create a more easily perceptible display of speech, the dynamical spectrogram has been modified as follows:

- On the low frequency lines of the cochleagram image (Figure 4.5) the negative amplitudes have been shown with white points. This means that half the points on a line have no information on them and the energy of those lines are not depicted meaningfully. Thus, the low-frequency lines look less intense than high-frequency ones. To compensate for this and so make the spectrogram movie more visually perceptible, the half-wave rectifiers in the model have been replaced by full-wave ones. Although this decreases the phase information on the low frequency lines of the image, this loss is not significant when interpreting speech.

- Other modifications have been applied to the shape of the image. As shown in Figure 4.10, the vertical sides of the rectangular image have been changed into sine-wave shapes. This acts as fixed calibration markers which help the viewer to more easily associate the vertical positions of the cochleagram to their corresponding frequencies.

- For the purpose of the experiments described in the next chapter, we have set the frequency domain of the visualiser to the range of 200 Hz to 3800 Hz. Also
4.6. Speech Visualiser

Figure 4.9: Comparison between the correlogram and dynamical spectrogram of a synthesised harmonic sound having its odd components mistuned. Note the phase changes between successive frames of the dynamical spectrogram. The correlogram is unable to show the phase changes, despite the beat sensation given rise to during the auditory perception of the underlying sound.
the outputs from the rectifiers have been smoothed by a 500-Hz low-pass filter, instead of 800-Hz, for a better visual display.

Figure 4.11 shows 18 successive frames of the speech specific spectrogram of the word 'yes', uttered by a British male speaker. The images have been displayed with a rate of 50 images per second, resulting in an animation of 360 milliseconds duration. As is evident in the figure, the high-frequency energy of the spectrum of the consonant /s/, the major cue to its recognition, does not show in the spectrogram movie. This is because of the low-pass filtering of the input sound, which is like the word being spoken over a telephone line, which in turn is anticipated to be an important context for the application of this process. Although this substantially decreases the intelligibility of the visual translation of the word, the formant transitions during the voiced parts of the word remain helpful for distinguishing this word from, say, the word 'no'.

4.7 Audvis, a Visualiser Tool

For implementation of the static spectrogram, the correlogram movie, and the dynamical spectrogram, several software packages have been developed. The dynamical spectrogram package consists of two parts: An auditory model and a multimedia tool which we call, here, Audvis (for its work of audio-visual presentation of sound).
Figure 4.11: Dynamical spectrogram of the utterance 'yes' spoken by a British male speaker. The frames are animated with a speed of 50 frames per second in the order of top-left to bottom-right, in vertical steps. These images, then, cover a duration of 360 milliseconds of the speech sound.
4.7. Audvis, a Visualiser Tool

Figure 4.12: Block diagram of the Audvis software package. (a) Auditory model and image frame-locking. (b) Audvis, the audio-visualiser tool. This tool can be used for monitoring any sequence of audio-visual blocks (individual image frames with their corresponding audio frames).
4.7. Audvis, a Visualiser Tool

Figure 4.12 shows a schematic diagram of these two parts. The auditory model software (Figure 4.12a) translates the input audio signal into a sequence of audio-visual frames, which is called here *audvis signal* or *audvis sequence*.

Output from the Auditory Model, the audvis signal, is a train of audio-visual blocks with a rate of, say, 50 blocks per second; this rate is determined by the model. Each block consists of a certain number of audio samples followed by the data of a rectangular image pertaining to the time-frequency information of the accompanying audio frame. To be more precise, as a result of the action of frame-locking described in the previous section, the image does not exactly pertain to the accompanying audio frame, but rather it belongs to a frame shared between it and the previous frame. However, as the frame-locking is accomplished only for periodic frames, e.g. voiced segments, it follows that due to the similarity between successive frames of a periodic sound, the information embedded in the visual part of a block is nearly the same as that embedded in its audio part.

The Audvis tool (Figure 4.12b) accepts an audvis sequence as its input and, after separating audio frames from their visual counterparts, monitors the visual images on a screen and downloads the audio frames into the selected audio media. The rate of the animation, which is synchronised with the audio output, is normally determined by the input stream but can also be selected manually.

This software, being an interactive tool, has many other applications such as for audio signal editing, recording any number of successive images, drawing the waveform and the spectrogram of any frame of the signal, etc. Figure 4.13 shows the layout of Audvis. With this tool an audvis sequence, recorded in a special file format, can be loaded in and then displayed in real-time synchronised with the corresponding audio signal. This tool can be used in any type of sound visualisation research, as it can animate any sequence of audio-visual blocks recorded in the appropriate format.
4.7. Audvis, a Visualiser Tool

Figure 4.13: The Audvis software tool. This can be used to monitor a sequence of blocks, each containing a frame of an auditory signal and the rectangular image assigned to it. Its schematic block diagram is shown in Figure 4.12b.
4.8 Summary

In this chapter a brief discussion of sound perception has been given, followed by a short account of present knowledge about speech perception. Keeping the most perceptually important qualities of speech sounds in mind, current theories of auditory representation in the human ear have been employed to devise an appropriate method of sound visualisation. Finally, besides examples of some other types of sound visualisation, our implemented system, Audvis, has been described in the form of a block diagram.

The most perceptually important qualities of sound are: loudness, pitch, and timbre. Loudness, the perceptual intensity of a sound, depends on physical parameters such as its amplitude, bandwidth, and time duration. For a pure tone, experimental results show a tonotopical relation between loudness and physical intensity of the sound; this has been suggested to be a power function. Pitch, the perceived periodicity of a sound, mostly depends on its frequency in the case of a pure tone, and on the repetition rate of its waveform for a complex sound. Timbre, a multidimensional quality of a sound, depends on the number, frequency, intensity, and relative phases of its harmonics. An appropriate visual representation of sound must present these three main qualities with their temporal transitions to the viewer with sufficient resolution.

Speech perception is argued not to be simply the perception of a train of speech sounds, the phonemes, but rather to be a more complex process of the human brain that is specialised for this task. According to this theory, speech is a special code, which is produced by the vocal tract articulation, and which is decoded by a highly specialised system in the listener's brain. The counter argument holds that the speech signal is analysed through a sequence of stages, using different properties of speech sounds and their relations, and reaching to a reasonable decision by taking context into account. According to this view, a number of invariant cues can be identified for any speech sound if the signal is analysed appropriately. There are also
available some visual cues such as facial movements during the speech production, which may affect the resulting perception.

According to the current knowledge about the representation functions of the ear, the outputs from our auditory model are depicted in the form of a belt-like image—the static spectrogram. A static spectrogram, also referred to as a cochleagram, is a rich, informative visual representation which shows nearly all the perceptually important qualities of sound with a good resolution. Loudness is reflected by the overall darkness of this image, pitch is shown as its spatial periodicity, and timbre is illustrated by its gray level pattern. The only problem with the static spectrogram is the way it maps the temporal transitions of an ongoing sound onto a spatial format, leading to a sliding image which is difficult to be perceived and followed by the eye. Thus the temporal character of a speech signal, one of its most important features, becomes transformed into a different format which makes it very difficult to perceive.

Two solutions have been proposed for this problem. The first is the correlogram, proposed by Slaney, et. al., which displays the periodicities inherent in the outputs from the auditory channels. In this model the temporal nature of an auditory event remains intact, but with substantially loss of phase information. By our technique of frame-locked sampling of the cochleagram, however, not only is phase information preserved, but also the fine time structure of auditory channels are displayed vividly. A software package comprising two separate parts has been developed. The first part contains an auditory model followed by a frame-locking block. The second is an audio-visual monitoring tool which can be used to display any sequence of audio-visual frames that have been supplied in a specified simple format.
Chapter 5

Experiments, Result, and Discussion

In the last two chapters, the theories and techniques behind the development of a sound visualisation package consisting of an auditory model and a visualiser tool, Audvis, were described. In this chapter our visualisation method is evaluated by a number of experiments with normal-hearing subjects.

The main experiment of this dissertation involves novice subjects who are totally unfamiliar with the type of visual gestures created by Audvis as translations of corresponding spoken words. Thus any uncontrolled oral communication between a subject and the experimenter or previous subjects might influence the way he interprets the gestures and, as a result, affect the uniformity of conditions across different subjects (subjects will be referred to here as “he”, since all but one were male). To avoid this, a software tool containing several interactive windows has been developed for running the experiments automatically under computer control. This program directs the subject, step by step, throughout a test session, by providing the necessary information and prompts at each step. The software has been designed so that instructions, test materials and the structure of every test considered for a specific session can all be loaded into it from a resource file. The structure and management of each experiment will be described in the following sections.
5.1 Methods

Our goal here is to investigate the potential usefulness of the sound visualiser as an aid for speech recognition, at least for the limited vocabularies described below. As discussed in a previous chapter (section 4.2, *Speech Perception*), some linguists and psychologists have proposed that speech is not just a train of characteristic speech sounds, called phonemes, but rather a complex combination of sounds coded into syllables and whole words. They argue that phonemes cannot be considered as 'the basic units' of speech. In the light of this argument and seeking a realistic way for the evaluation of the system, we decided to use simple words in a controlled setting for the preliminary tests of the system.

One of the critical problems of the deaf is that they cannot make use of a telephone—one of the commonest channels for modern human communication. The imagined scenario for evaluation experiments, then, is that a deaf or hearing-impaired user is presented with words spoken over a telephone line and then displayed as animated shapes on a screen, and tries to recognise them. To limit the possible responses, the words to be used are organised into sets of two, three, and four words and tested using standard forced-choice tests.

5.1.1 Speech Materials

The speech materials employed throughout these experiments consist of a selection of single words (e.g. 'yes', 'no', 'left', 'right', etc.) that would most frequently be needed during a typical telephone conversation in which a person is trying to obtain directions. The words used are listed in table 5.1. The words were spoken at a normal speed by three male and one female speakers. The spoken words were recorded in a room with ordinary acoustics and then low-pass filtered in order to simulate the conditions of an ordinary telephone line—a low-pass channel of 4 kHz bandwidth. Each word was uttered at least ten times with slightly different accents and stresses. The low-pass filtered words were sampled at a rate of 8000 samples per second and then input to the visualiser system. The resulting outputs from the system, in the
5.1. Methods

Table 5.1: The full set of English words used in the experiments in the form of two-to four-word, forced-choice tests.

<table>
<thead>
<tr>
<th>yes</th>
<th>no</th>
<th>left</th>
<th>right</th>
<th>go</th>
<th>stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>zero</td>
<td>one</td>
<td>two</td>
<td>three</td>
<td>...</td>
<td>nine</td>
</tr>
<tr>
<td>south</td>
<td>west</td>
<td>north</td>
<td>east</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.2: Pronunciations of the translations of the words 'yes' and 'no' in three further languages, used in the tests of language independence.

<table>
<thead>
<tr>
<th>Language</th>
<th>Pronunciation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persian</td>
<td>baleh kheir</td>
</tr>
<tr>
<td>French</td>
<td>wee no</td>
</tr>
<tr>
<td>Czech</td>
<td>ano ne</td>
</tr>
</tbody>
</table>

A form of sequences of images, were then recorded in a special format in computer files which can be processed by Audvis—the visualiser tool. The recorded words have a typical duration of about 400 milliseconds.

To evaluate the language independence of the system, we have also used the translations of the words 'yes' and 'no' in three further widely different languages: Persian (Farsi), French, and Czech. The French words were spoken by a female native speaker and the Persian and the Czech ones by two male native speakers. These words have also each been uttered ten times. Table 5.2 shows their pronunciations.

5.1.2 Subjects

For our preliminary experiments we selected normal-hearing, rather than deaf or hearing-impaired, subjects for two main reasons. The first was that finding a fairly large number of deaf or hearing-impaired subjects with similar levels of intelligence and age presented practical difficulties. The second was the potential difficulty in communicating with deaf people. Moreover, normal-hearing subjects would find themselves in the same position as deaf persons in our experiments, where there is no sound to hear. Of course, there must remain differences between deaf and
hearing people, from psychological points of view, even with respect to non-auditory representations of sound. However, the results of tests with normal subjects may still give much information about the potential usefulness of the system for the deaf.

The subjects were a number of students, mostly postgraduate, who volunteered to participate in the experiments. They all had normal or normally corrected vision and were aged 21 to 46 years. All the subjects were completely unfamiliar with the system of sound visualisation under investigation.

5.2 Main Experiment

The main experiment has been designed so that it can evaluate the ability of novice users of the system to learn the gestures as the visual translations of spoken words.

5.2.1 Test Procedure

In the main experiment the same test session, comprising ten separate parts, was offered to each subject. The subjects were asked to give up about 45 minutes of their time to sit down by a computer workstation and to carry out the experiment. Each part of the session, which is called here a test, is involved with only two, three, or four specified words as shown in Table 5.3.

Throughout a test session the subject, under automatic direction from an interactive computer program, proceeds step by step by pressing an appropriate highlighted button at each step. The visual translation of a word will start to be animated in a window on the screen two seconds after the subject presses the button labelled 'Next Word'. The time duration of the animation of a particular word is exactly the same as that of the same spoken word in real time. After that, the window is cleared—representing silence. Thus, only the start time of the animation of a word is under the control of the subject. When an answer to a question is required, the subject is allowed unlimited time to decide and then select the answer.

Each test (e.g. that using the words 'yes' and 'no'), consisted of three following stages (see Figures 5.1 and 5.2).
5.2. Main Experiment

You are going to take a test for assessment of a system of sound visualization. Each time you press the "Next word" button you will be shown an animation related to a known word. You are asked to watch the animation and try to learn it as the visualization of the corresponding word.

After some training trials, including learning and practice, you will be asked to link some randomly displayed animations to their corresponding words.

It is not very difficult. Just relax and watch the whole image. You need to be careful to watch how it emerges, moves, and vanishes. Please do not miss out the instructions provided at each stage throughout the test.

Now press "continue" to go on.

a) Instructions given to the subject at the start of the session.

Test No: 1

Learning

Here you learn the animation of the words: "yes" and "no".

Each time you press the "Next word" button you will be shown the name of a word followed by its animation. Repeat the action and try to learn the specific visualisation of each word.

b) Layout of the Learn stage.

Figure 5.1: Layout of the interactive window of the test session, (a) at the start of the session, and (b) at the Learn stage of the first test. For the Practice and Question stages see Figure 5.2.
5.2. Main Experiment

Practice
Here you practice through some examples to answer the questions. Press "Next word", watch the image, and decide which word was displayed. Then press the corresponding button and look at the message window to see the correct answer.

(b) Layout of the Practice stage.

Questions
Now try again without being told the correct answer.

(b) Layout of the Question stage.

Figure 5.2: Layout of the interactive window of the test session, (a) at the Practice stage, and (b) at the Question stage of the first test. See also Figure 5.1.
Table 5.3: The words used in the multi-word, forced-choice tests and the conditions of their repetition in each test.

<table>
<thead>
<tr>
<th>test number</th>
<th>words involved</th>
<th>conditions of repeated trials</th>
<th>number of questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>yes, no</td>
<td>the same utterance</td>
<td>10</td>
</tr>
<tr>
<td>T2</td>
<td>yes, no</td>
<td>different utterances, the same speaker</td>
<td>10</td>
</tr>
<tr>
<td>T3</td>
<td>yes, no</td>
<td>different speakers</td>
<td>10</td>
</tr>
<tr>
<td>T4</td>
<td>left, right</td>
<td>different utterances, the same speaker</td>
<td>10</td>
</tr>
<tr>
<td>T5</td>
<td>go, stop</td>
<td>different utterances, the same speaker</td>
<td>10</td>
</tr>
<tr>
<td>T6</td>
<td>baleh, kheir</td>
<td>different utterances, the same speaker</td>
<td>10</td>
</tr>
<tr>
<td>T7</td>
<td>wee, no</td>
<td>different utterances, the same speaker</td>
<td>10</td>
</tr>
<tr>
<td>T8</td>
<td>ano, ne</td>
<td>different utterances, the same speaker</td>
<td>10</td>
</tr>
<tr>
<td>T9</td>
<td>one, two, three</td>
<td>different utterances, the same speaker</td>
<td>12</td>
</tr>
<tr>
<td>T10</td>
<td>south, west, north, east</td>
<td>different utterances, the same speaker</td>
<td>12</td>
</tr>
</tbody>
</table>

Learn Stage

At this stage, concerned with the learning of the visual translations of the words involved, each spoken word (i.e. its visual gesture, without any sound) is displayed ten times. The word is also printed in a small window, immediately after the 'Next Word' button is pressed and two seconds before the animation is started, so that the subject is aware of the word to be displayed each time (see Figure 5.1b). The words appear during the learning stage in the fixed order of: 'yes', 'yes', 'no', 'no', 'yes', 'yes', ... so that each word is displayed exactly ten times in all. At this stage, every two successive utterances and hence displays of a word are identical; other instances of the same word in this stage and the subsequent stages are different as indicated in Table 5.3.

According to this design the subject is given a predefined, fixed number of chances to watch the gestures of the words involved, after which he is automatically directed to the next stage. This is in order to control the number of learning trials so as to be better able to compare the abilities of subjects at learning the visual gestures, in so far as this is reflected in their subsequent success at identifying them.
5.2. **Main Experiment**

**Practice Stage**

This stage was included to give the learner a chance to evaluate his progress and obtain more familiarity with the format of the test. At this stage a gesture, from among those learned, is randomly selected to be presented (but not identified) to the subject upon his click on the 'Next Word' button. Then he is asked to decide which word was displayed. When the subject reaches a decision, he presses the appropriately labelled button and the correct answer is displayed as feedback to let him evaluate his decision (Figure 5.2a). This sequence of practice questions continue until each word has been displayed three times.

**Question Stage**

This stage is carried out like the Practice stage except that the viewer is not given the correct answer. The number of questions given in this stage is 10 for all the two-word tests and 12 for the other tests. All the words involved in a particular test enter equally into this stage, that is, the number of gestural representations of each word displayed during the series of questions is the same for all words, although the subject is not made aware of this. Note that in all the tests, except for the first one (test T1, in Table 5.3), repeated presentations of a certain word are not the same, but rather are the gestures from different utterances of the same word.

It is the results from this stage, that is, the number of correct decisions recorded, which is used to evaluate the learning ability of the subject for the corresponding words. All the questions, together with the corresponding answers given by the subject, are recorded in a file. This file is subsequently processed by software to obtain the final results.

**5.2.2 Evaluation Measure**

The number of correct answers is used to assign a score to each test. To compensate for the effect of chance in evaluating multi-word, forced-choice tests, the
usual formula \([104]\):

\[
P_c = \frac{C - W}{\frac{n}{n-1}} \times 100
\]

was used. In this equation:

- \(P_c\) = the chance-adjusted, percentage correct response (the score)
- \(T\) = total number of questions in a test
- \(C\) = number of correct responses
- \(W\) = number of wrong responses, that is, \(T - C\)
- \(n\) = number of choices (e.g. 3 for a three-word, forced-choice test)

By this equation a score is obtained which will be 100 if all the answers are correct, and an expectation value of zero if all questions are answered randomly.

5.3 Results and Discussion

5.3.1 Results

The chance-adjusted percentage correct response scores of 30 subjects for every test are presented in Table 5.4. The data have been sorted in descending order of the average scores obtained by the subjects. Table 5.5 records some further information relating to the subjects and the conditions of the tests. These are: the sex and native language of each subject as well as the time of day at the start and the duration of the test session.

The last two rows of Table 5.4 represent the mean and standard deviation of the scores for each test across all subjects. According to equation (5.1) a score of less than or equal to zero stands for misidentification or random answering respectively. Bearing this in mind, in computing the means and variances, all negative scores have been taken as zero, because a negative score does not mean a worse response than a zero score. For example, a score of \(-80\), as for test T7 of subject S28 (Table 5.4), represents that the subject has been able to distinguish between the words 'wee' and 'no', but that he has identified the visualisation of 'wee' as that of 'no' and vice versa. Therefore in a sense this may be judged as actually better than a zero score, since the latter would merely represent a totally chance set of responses.
5.3. Results and Discussion

Table 5.4: Percentage scores for the ten forced-choice tests for all subjects, together with their averages and their ages. Mean and standard deviation of the scores of each test over the subjects are also included in the table. Note that a score of zero represents chance expectation. For the conditions of the tests see table 5.3.

<table>
<thead>
<tr>
<th>Subject</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>T6</th>
<th>T7</th>
<th>T8</th>
<th>T9</th>
<th>T10</th>
<th>Mean</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>S0</td>
<td>100</td>
<td>100</td>
<td>60</td>
<td>100</td>
<td>40</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>90</td>
</tr>
<tr>
<td>S1</td>
<td>100</td>
<td>80</td>
<td>60</td>
<td>80</td>
<td>80</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>85</td>
<td>100</td>
<td>88.5</td>
<td>35</td>
</tr>
<tr>
<td>S2</td>
<td>100</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>100</td>
<td>100</td>
<td>80</td>
<td>100</td>
<td>87</td>
<td>86.7</td>
<td>33</td>
</tr>
<tr>
<td>S3</td>
<td>100</td>
<td>60</td>
<td>60</td>
<td>80</td>
<td>60</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>85</td>
<td>82.5</td>
<td>36</td>
</tr>
<tr>
<td>S4</td>
<td>80</td>
<td>100</td>
<td>80</td>
<td>60</td>
<td>80</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>75</td>
<td>44</td>
<td>81.9</td>
<td>21</td>
</tr>
<tr>
<td>S5</td>
<td>100</td>
<td>60</td>
<td>60</td>
<td>80</td>
<td>60</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>60</td>
<td>40</td>
<td>100</td>
<td>76</td>
</tr>
<tr>
<td>S6</td>
<td>60</td>
<td>20</td>
<td>60</td>
<td>100</td>
<td>20</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>75</td>
<td>100</td>
<td>73.5</td>
<td>36</td>
</tr>
<tr>
<td>S7</td>
<td>80</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>40</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>11</td>
<td>73.1</td>
<td>29</td>
</tr>
<tr>
<td>S8</td>
<td>100</td>
<td>0</td>
<td>-20</td>
<td>100</td>
<td>60</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>60</td>
<td>72</td>
<td>32</td>
</tr>
<tr>
<td>S9</td>
<td>100</td>
<td>100</td>
<td>40</td>
<td>0</td>
<td>80</td>
<td>100</td>
<td>100</td>
<td>55</td>
<td>33</td>
<td>70.8</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>S10</td>
<td>-20</td>
<td>100</td>
<td>60</td>
<td>80</td>
<td>20</td>
<td>100</td>
<td>100</td>
<td>75</td>
<td>89</td>
<td>79.4</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>S11</td>
<td>80</td>
<td>80</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>60</td>
<td>100</td>
<td>63</td>
<td>56</td>
<td>69.9</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>S12</td>
<td>0</td>
<td>80</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>100</td>
<td>100</td>
<td>80</td>
<td>70</td>
<td>47</td>
<td>69.7</td>
<td>33</td>
</tr>
<tr>
<td>S13</td>
<td>40</td>
<td>20</td>
<td>100</td>
<td>80</td>
<td>-20</td>
<td>100</td>
<td>100</td>
<td>85</td>
<td>47</td>
<td>67.2</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>S14</td>
<td>80</td>
<td>40</td>
<td>20</td>
<td>40</td>
<td>80</td>
<td>100</td>
<td>100</td>
<td>83</td>
<td>67</td>
<td>67</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>S15</td>
<td>40</td>
<td>0</td>
<td>40</td>
<td>60</td>
<td>60</td>
<td>100</td>
<td>100</td>
<td>88</td>
<td>67</td>
<td>65.5</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>S16</td>
<td>20</td>
<td>20</td>
<td>100</td>
<td>100</td>
<td>80</td>
<td>60</td>
<td>100</td>
<td>75</td>
<td>67</td>
<td>64.2</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>S17</td>
<td>20</td>
<td>80</td>
<td>20</td>
<td>30</td>
<td>20</td>
<td>100</td>
<td>100</td>
<td>88</td>
<td>67</td>
<td>61.5</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>S18</td>
<td>20</td>
<td>40</td>
<td>20</td>
<td>100</td>
<td>20</td>
<td>100</td>
<td>100</td>
<td>70</td>
<td>73</td>
<td>60.3</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>S19</td>
<td>40</td>
<td>0</td>
<td>20</td>
<td>80</td>
<td>60</td>
<td>100</td>
<td>100</td>
<td>25</td>
<td>73</td>
<td>59.8</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>S20</td>
<td>0</td>
<td>0</td>
<td>60</td>
<td>80</td>
<td>60</td>
<td>100</td>
<td>100</td>
<td>75</td>
<td>22</td>
<td>59.7</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>S21</td>
<td>-20</td>
<td>80</td>
<td>0</td>
<td>80</td>
<td>40</td>
<td>80</td>
<td>100</td>
<td>100</td>
<td>63</td>
<td>44</td>
<td>68.7</td>
<td>37</td>
</tr>
<tr>
<td>S22</td>
<td>0</td>
<td>40</td>
<td>40</td>
<td>100</td>
<td>40</td>
<td>100</td>
<td>100</td>
<td>40</td>
<td>63</td>
<td>56</td>
<td>57.9</td>
<td>38</td>
</tr>
<tr>
<td>S23</td>
<td>20</td>
<td>-20</td>
<td>60</td>
<td>60</td>
<td>40</td>
<td>80</td>
<td>100</td>
<td>75</td>
<td>44</td>
<td>55.9</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>S24</td>
<td>40</td>
<td>40</td>
<td>100</td>
<td>100</td>
<td>40</td>
<td>80</td>
<td>100</td>
<td>38</td>
<td>22</td>
<td>54</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>S25</td>
<td>40</td>
<td>80</td>
<td>-40</td>
<td>100</td>
<td>100</td>
<td>-20</td>
<td>60</td>
<td>13</td>
<td>67</td>
<td>52</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>S26</td>
<td>0</td>
<td>60</td>
<td>40</td>
<td>60</td>
<td>20</td>
<td>60</td>
<td>100</td>
<td>75</td>
<td>0</td>
<td>51.5</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>S27</td>
<td>0</td>
<td>-40</td>
<td>20</td>
<td>100</td>
<td>20</td>
<td>80</td>
<td>100</td>
<td>38</td>
<td>33</td>
<td>47.1</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>S28</td>
<td>40</td>
<td>-20</td>
<td>40</td>
<td>60</td>
<td>40</td>
<td>80</td>
<td>-80</td>
<td>0</td>
<td>70</td>
<td>20</td>
<td>35</td>
<td>40</td>
</tr>
<tr>
<td>S29</td>
<td>20</td>
<td>-20</td>
<td>40</td>
<td>20</td>
<td>0</td>
<td>80</td>
<td>40</td>
<td>0</td>
<td>25</td>
<td>33</td>
<td>25.8</td>
<td>42</td>
</tr>
</tbody>
</table>

Mean: 46.7 45.3 49.3 80 39.3 86.7 91.3 84.7 68.4 57.6 64.9
SD: 39.4 37.9 31 21.7 29 21.2 22.7 27.5 23.3 28.8 14.6
### Table 5.5: Characteristics of the subjects taking part in the main experiment.

The average score obtained, the elapsed time, and the time of day at the start of the test session, have also been shown for each subject.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Average</th>
<th>Language</th>
<th>age</th>
<th>sex</th>
<th>Elapsed time (minutes)</th>
<th>Start time</th>
</tr>
</thead>
<tbody>
<tr>
<td>S0</td>
<td>90</td>
<td>English</td>
<td>31</td>
<td>m</td>
<td>36</td>
<td>9:34</td>
</tr>
<tr>
<td>S1</td>
<td>88.5</td>
<td>Sindhi</td>
<td>35</td>
<td>m</td>
<td>42</td>
<td>15:07</td>
</tr>
<tr>
<td>S2</td>
<td>86.7</td>
<td>Persian</td>
<td>33</td>
<td>m</td>
<td>43</td>
<td>18:58</td>
</tr>
<tr>
<td>S3</td>
<td>82.5</td>
<td>Persian</td>
<td>36</td>
<td>m</td>
<td>47</td>
<td>12:33</td>
</tr>
<tr>
<td>S4</td>
<td>81.9</td>
<td>English</td>
<td>21</td>
<td>f</td>
<td>31</td>
<td>14:06</td>
</tr>
<tr>
<td>S5</td>
<td>76</td>
<td>Persian</td>
<td>39</td>
<td>m</td>
<td>48</td>
<td>19:09</td>
</tr>
<tr>
<td>S6</td>
<td>73.5</td>
<td>Persian</td>
<td>36</td>
<td>m</td>
<td>48</td>
<td>15:34</td>
</tr>
<tr>
<td>S7</td>
<td>73.1</td>
<td>Persian</td>
<td>29</td>
<td>m</td>
<td>42</td>
<td>11:16</td>
</tr>
<tr>
<td>S8</td>
<td>72</td>
<td>Persian</td>
<td>32</td>
<td>m</td>
<td>56</td>
<td>14:09</td>
</tr>
<tr>
<td>S9</td>
<td>70.8</td>
<td>Persian</td>
<td>35</td>
<td>m</td>
<td>50</td>
<td>9:42</td>
</tr>
<tr>
<td>S10</td>
<td>70.4</td>
<td>Persian</td>
<td>29</td>
<td>m</td>
<td>31</td>
<td>19:02</td>
</tr>
<tr>
<td>S11</td>
<td>69.9</td>
<td>Persian</td>
<td>31</td>
<td>m</td>
<td>40</td>
<td>16:03</td>
</tr>
<tr>
<td>S12</td>
<td>69.7</td>
<td>Persian</td>
<td>33</td>
<td>m</td>
<td>38</td>
<td>13:50</td>
</tr>
<tr>
<td>S13</td>
<td>67.2</td>
<td>English</td>
<td>29</td>
<td>m</td>
<td>49</td>
<td>10:52</td>
</tr>
<tr>
<td>S14</td>
<td>67</td>
<td>Persian</td>
<td>36</td>
<td>m</td>
<td>38</td>
<td>15:32</td>
</tr>
<tr>
<td>S15</td>
<td>65.5</td>
<td>Persian</td>
<td>35</td>
<td>m</td>
<td>120</td>
<td>14:04</td>
</tr>
<tr>
<td>S16</td>
<td>64.2</td>
<td>Persian</td>
<td>42</td>
<td>m</td>
<td>52</td>
<td>18:37</td>
</tr>
<tr>
<td>S17</td>
<td>61.5</td>
<td>Persian</td>
<td>46</td>
<td>m</td>
<td>46</td>
<td>18:05</td>
</tr>
<tr>
<td>S18</td>
<td>60.3</td>
<td>Persian</td>
<td>30</td>
<td>m</td>
<td>38</td>
<td>9:51</td>
</tr>
<tr>
<td>S19</td>
<td>59.8</td>
<td>Arabic</td>
<td>28</td>
<td>m</td>
<td>52</td>
<td>19:00</td>
</tr>
<tr>
<td>S20</td>
<td>59.7</td>
<td>German</td>
<td>28</td>
<td>m</td>
<td>39</td>
<td>14:48</td>
</tr>
<tr>
<td>S21</td>
<td>58.7</td>
<td>Persian</td>
<td>37</td>
<td>m</td>
<td>42</td>
<td>15:22</td>
</tr>
<tr>
<td>S22</td>
<td>57.9</td>
<td>Persian</td>
<td>38</td>
<td>m</td>
<td>54</td>
<td>15:11</td>
</tr>
<tr>
<td>S23</td>
<td>55.9</td>
<td>Persian</td>
<td>36</td>
<td>m</td>
<td>35</td>
<td>15:11</td>
</tr>
<tr>
<td>S24</td>
<td>54</td>
<td>Persian</td>
<td>39</td>
<td>m</td>
<td>50</td>
<td>19:41</td>
</tr>
<tr>
<td>S25</td>
<td>52</td>
<td>Persian</td>
<td>40</td>
<td>m</td>
<td>39</td>
<td>15:36</td>
</tr>
<tr>
<td>S26</td>
<td>51.5</td>
<td>Persian</td>
<td>38</td>
<td>m</td>
<td>36</td>
<td>15:37</td>
</tr>
<tr>
<td>S27</td>
<td>47.1</td>
<td>Persian</td>
<td>38</td>
<td>m</td>
<td>81</td>
<td>13:56</td>
</tr>
<tr>
<td>S28</td>
<td>35</td>
<td>Persian</td>
<td>40</td>
<td>m</td>
<td>44</td>
<td>18:48</td>
</tr>
<tr>
<td>S29</td>
<td>25.8</td>
<td>Persian</td>
<td>42</td>
<td>m</td>
<td>42</td>
<td>15:18</td>
</tr>
</tbody>
</table>
5.3. Results and Discussion

Figure 5.3: Means and standard deviations of chance-adjusted, percentage correct scores for the ten two- to four-word, forced-choice tests of the main experiment. Averages have been computed over the 30 normal-hearing subjects in Table 5.4. In fact scores cannot exceed 100—that the error bars for tests T4 and T6-T8 appear to do so is an artifact that reflect the presence of a ceiling effect when the scores approach the upper bound.

The age of each subject is also reported in the last column of Table 5.4.

5.3.2 Discussion

The results are summarised in Figure 5.3. The mean and standard deviation for each test, over all the subjects, are plotted versus the number of the test, which represents its temporal position in the sequence of administration from T1 to T10.
5.3. Results and Discussion

Progress During the Test Session

The graph indicates that the subjects, from being initially totally unfamiliar with the system, progressed gradually during the test session. Although the mean scores of the first three tests are nearly equal, the increasing trend of their difficulty from T1 to T3 confirms the progress of the subjects in learning the visual gestures. Note that in the first test (T1) all the 10 questions have been selected from the same utterances of the words ‘yes’ and ‘no’, whereas the second test involves different utterances of the same pair of words and therefore a lower score for T2, compared to T1, might be expected. The third test (T3) has a higher score than the second one, in spite of the fact that the corresponding words have been spoken by different speakers. This reflects the improved ability of the subjects in learning to link slightly differing gestures to the same word, while distinguishing them from those linked to the other word. The high scores (about 90 percent) for tests T6, T7, and T8 show that after a relatively short period of learning time (about 20 minutes), subjects became sufficiently familiar with the system of visualisation and then only needed to see a few instances of pairs of new words to learn to discriminate between them. The presence of one-SD error bars which apparently extend beyond the maximum possible score of 100 in the case of tests T4 and T6-T8 is an artifact, reflecting the presence of a strong ceiling effect as the scores approached 100 and a correspondingly negatively skewed distribution of scores.

The low score for test T5, despite the familiarity of the subject with the system by the time, is due to the accidental similarity of the gestures for the words ‘left’ and ‘right’. This can be understood because they also seem audibly to be less discriminable than the other sets of words. A few scores of 80 percent or more for this test in Table 5.4 proves that these words are in fact recognisable, although only for a minority of the subjects.

The lower scores for the last two tests, T9 and T10, is interpreted as having two causes: the first is that the subject is now concerned with multi-word, rather than two-word, sets which are thus more difficult to learn; the second is the problem of
subject fatigue near the end of the test session, due to the efforts involved in learning what is, in effect, a new language.

**Language Independency of the System**

The words used in tests T6, T7, and T8 have been chosen from very different languages: Persian, French, and Czech respectively. They are the translations of the words 'yes' and 'no' and have been marked with their pronunciations in Figure 5.3. The high mean recognition scores on these tests suggest that the system of representation is independent of the language involved. The language independency of the system was of course predictable because it is the sound, not the speech, qualities which are translated by the system into the visual attributes of the displayed images.

**Individual Differences**

The high value of the standard deviations, especially for the first two tests, as shown in Figure 5.3, indicates large individual differences in learning ability after a limited number (10) of learning trials. These individual differences, however, decrease for the middle tests, suggesting that with a further amount of training, the words would become recognisable for all subjects.

There seems to be a relation between the average score and age of the subjects, as shown in Figure 5.4. The number of subjects and range of their ages presented in this figure is insufficient to establish a reliable relationship. However, the subjects are well matched in several respects (e.g. their level of education, their unfamiliarity with the system before, and the instructions given to them during the test session) so that the data do broadly confirm, as one might expect, that younger users (below about 35) are typically better at learning the gestures than older ones.

**Matrices of Confusion**

Table 5.6 shows the matrices of confusion between the words used in the last two tests (the three-word and four-word cases). Table 5.6a shows that among the four words 'south', 'west', 'north' and 'east', the greatest confusion occurs between 'south'
5.3. Results and Discussion

Figure 5.4: Average of the scores obtained by the subjects in relation to their ages. The abscissae represents the age and the ordinate shows the average of the 10 chance-adjusted, percentage correct, scores of the tests undertaken by each subject.

and 'west'. These two words are arguably also audibly more confusable than the other pairs of words. Table 5.6b indicates that among the three words 'one', 'two', and 'three', most confusion has occurred between 'two' and 'three', which again are somewhat audibly more confusable than the other pairs. The agreement between visual confusion of the gestures and the auditory confusion of the corresponding utterances supports the fidelity of the visualiser system in translating the auditory information. Further investigations are, of course, needed to evaluate further this fidelity which is a desirable property for the system.

General Findings

- All the words under examination are recognisable through their visual translations, at least by some of the subjects.
5.4 Other Experiments

Table 5.6: Matrix of confusion between the four words 'south', 'west', 'north', and 'east' (a), and between the three words 'one', 'two', and 'three' (b).

<table>
<thead>
<tr>
<th></th>
<th>south</th>
<th>west</th>
<th>north</th>
<th>east</th>
</tr>
</thead>
<tbody>
<tr>
<td>south</td>
<td>20</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>west</td>
<td>29</td>
<td></td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>north</td>
<td>4</td>
<td>12</td>
<td></td>
<td>17</td>
</tr>
<tr>
<td>east</td>
<td>0</td>
<td>10</td>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>one</th>
<th>two</th>
<th>three</th>
</tr>
</thead>
<tbody>
<tr>
<td>one</td>
<td>9</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>two</td>
<td>8</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>three</td>
<td>4</td>
<td>22</td>
<td></td>
</tr>
</tbody>
</table>

- After a small amount of familiarity with the system, typical users are able to learn to distinguish between small groups of words by watching their gestures displayed in real time only 10 times. This appears to be approximately an order of magnitude faster than previously reported using speech spectrograms [75], although differences in test procedures preclude any more detailed comparisons.

- There is no evidence for dependency of the system upon language applied.

- There are many individual differences at the start of learning, but this appears to decrease as familiarity with the system increases (although this is complicated by ceiling effects as proficiency improves).

- There is an apparent relation between the age and the learning ability of the users, so that the older the user the lower his ability to learn from a limited number of training trials. This decline was observed for ages above about 35 years and for 10 learning presentations of each word.

5.4 Other Experiments

In this section two further experiments which have been undertaken by the author are discussed.
Table 5.7: Results for a test involving the four words 'south', 'west', 'north', and 'east' without any prior learning stage. The words can thus have been recognised only because of their similarity to previously known words. Errors are shown in bold.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>north</td>
<td>north</td>
<td>11</td>
<td>north</td>
<td>north</td>
</tr>
<tr>
<td>2</td>
<td>south</td>
<td>south</td>
<td>12</td>
<td>south</td>
<td>south</td>
</tr>
<tr>
<td>3</td>
<td>south</td>
<td>south</td>
<td>13</td>
<td>west</td>
<td>west</td>
</tr>
<tr>
<td>4</td>
<td>west</td>
<td>west</td>
<td>14</td>
<td>east</td>
<td>east</td>
</tr>
<tr>
<td>5</td>
<td>north</td>
<td>north</td>
<td>15</td>
<td>west</td>
<td>west</td>
</tr>
<tr>
<td>6</td>
<td>east</td>
<td>east</td>
<td>16</td>
<td>west</td>
<td>west</td>
</tr>
<tr>
<td>7</td>
<td>east</td>
<td>east</td>
<td>17</td>
<td>north</td>
<td>north</td>
</tr>
<tr>
<td>8</td>
<td>south</td>
<td>north</td>
<td>18</td>
<td>south</td>
<td>south</td>
</tr>
<tr>
<td>9</td>
<td>east</td>
<td>east</td>
<td>19</td>
<td>west</td>
<td>west</td>
</tr>
<tr>
<td>10</td>
<td>south</td>
<td>north</td>
<td>20</td>
<td>east</td>
<td>east</td>
</tr>
</tbody>
</table>

5.4.1 Recognition of New Gestures Without Learning

The goal in this test is to evaluate how learning of a number of words may help visual recognition of unseen words, that is, whether the words having similar auditory cues are indeed also visually similar. In this experiment the author, being by then sufficiently familiar with the visual translations of some of the words from Tables 5.1 and 5.2, carried out, as a subject, a test involving the four words 'south', 'west', 'north', and 'east' (which were not among those previously encountered). This test was in the form of a four-word, forced choice test like test T10 in Table 5.3. Ten different utterances of each word were translated into visual gestures resulting in a sum of 40 gestures. The test procedure was like the one described in section 5.2.1, except that it contained only the Question stage (i.e. without learning and practice) with 20 questions. That is, a random selection of 20 of the 40 unseen gestures were presented, to the subject, who was asked if he could recognise them. The results are shown in Table 5.7. The chance-adjusted, percentage correct response of this test was 86.7 percent which is a very good recognition score.

As is evident from the table, only questions 8 and 10, both involving the word 'south', have been answered incorrectly. The completely correct recognition of the
other three words ‘west’, ‘north’, and ‘east’ was possible because of the similarity of their gestures to those of the words: ‘left’, ‘four’, and ‘wee’ respectively, which were already known visually to the subject. This experiment suggests that just as one can recognise a previously unheard word by analysing it into a sequence of familiar phonemes, there are cues in the dynamical spectrograms, the gestures, of a word that permit a similar analysis, perhaps (as may well be the case with phonemes) through correlation with the form and movements of the vocal tract during the articulation of that word. Thus learning a small number of phonetically different words may help the user to recognise new utterances which are similar in some respects to the learned words, and hence to speed up his learning procedure.

This is of course only the case for a normal-hearing or a partially deaf person who has some preliminary knowledge about the relations between the auditory cues and articulatory movements. For those profoundly deaf from birth, however, the problem is different, because they are unfamiliar with the auditory cues except via touch, and thus have little or no knowledge about any correlations with the utterances. For them, therefore, the learning task is like learning a completely new language, demanding much more time and effort than for others.

For those among the deaf who know lipreading, of course, there exist visual cues linked to, say, lip shape or lip movements that might be of use. For example, someone familiar with the spectrogram gestures for ‘four’, ‘or’, ‘door’, etc., is predicted to be able to distinguish the gesture of ‘more’ from that of ‘less’ without any specific learning trial. This is because the similarity of the known gestures for the words ‘four’, ‘or’ and ‘door’ can be linked, as a cue, to the analogy of the lip shape during their articulation, and hence the gesture for ‘more’, which is expected to contain the same cue, can be recognised relatively easily.

5.4.2 Memory Decay for Visual Gestures

Although the results of the main experiment demonstrated the feasibility of learning a set of gestures in a short period of time, this is not in itself a sufficient condition
5.4. Other Experiments

Table 5.8: Results from an extended series of repeated recognition tests for the words 'zero' to 'nine', without any practice between tests.

<table>
<thead>
<tr>
<th>Test no.</th>
<th>Month in which test carried out</th>
<th>Days without practice since previous test</th>
<th>Score (percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>July</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>August</td>
<td>27</td>
<td>94.4</td>
</tr>
<tr>
<td>3</td>
<td>October</td>
<td>58</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>January</td>
<td>86</td>
<td>94.4</td>
</tr>
<tr>
<td>5</td>
<td>May</td>
<td>128</td>
<td>83.3</td>
</tr>
</tbody>
</table>

for the sound visualiser system to be useful for the deaf. The gestures learned must also be able to be retained in one's memory for a relatively long period of time. The experiment described below has been aimed at evaluating how persistent such learned gestures are.

Having by then become well experienced in the recognition of the words 'zero', 'one', ..., and 'nine', the author carried out a ten-word forced-choice test involving 20 questions randomly selected from a pool of 100 gestures for these words (10 gestures for each). This test was repeated 4 times over the following period of 10 months without any intervening practice during this period. The results are shown in Table 5.8. For durations of less than 3 months there was no significant memory decay, as the percentage score remained well above 90 percent. For the last test, carried out after the subject had not been exposed to those gestures for more than four months, there was a slight decrease in the score (to 83.3 percent). Thus it seems that such gestures can persist in the memory for extended periods without refreshing.

This experiment, of course, needs to be repeated using many more subjects to obtain a reliable amount of data. But it can at least be inferred from this experiment that people do exist who can keep the data in memory for an adequate amount of time. Such persistence of the gestures is a necessary prerequisite for the system to be of practical use to deaf and hearing impaired users.
5.4.3 Short Term Memory

Another necessary attribute for the success of visual translations of speech sounds is their ability to be retained in short term memory for a sufficient time, say for the duration of a sentence. That is, the user must be capable of holding a few words in working memory at the same time, in order to relate them with the next incoming word and hence extract the meaning of the phrase or sentence in real time. A simple experiment to evaluate this attribute of the gestures might be to take a test of recognition of, say, multi-digit numbers such as telephone numbers—for different numbers of digits. This experiment has been left to be carried out in the future, as the present version of the software developed here does not yet have the ability to support this task.

However, a very simple test of recognition of two-digit numbers, undertaken by the author, where the digits were limited to 'zero', 'one', 'two', 'three', and 'four' (with a score of 100 percent correct) suggests, as one might expect [52], that no difference exists between identification rates for one-digit numbers and those for two-digit numbers. For longer numbers, of course, further investigation would be needed.

5.5 Summary

In this chapter a preliminary evaluation of our developed visualiser system has been reported. This system converts speech sounds into visual gestures which are hoped to be recognised by eye. Several experiments in the form of multi-word forced-choice tests using normal-hearing subjects have been reported. The words employed for these experiments were selected from those likely to be used in response to requests for directions, etc., in the course of a typical telephone conversation, mostly in the English language.

The main experiment, a set of 10 two- to four-word tests, was carried out by 30 subjects, all educated adults aged 21-46 years, in a typical duration of 45 minutes. Each test (of the 10) involved three stages, namely: Learn, Practice, and Question
5.5. Summary

stages. In Learn stage, the subject was presented with 10 gestures of each previously specified word and tried to learn them. In the Practice stage, the subject was given an opportunity to reinforce his learning through practice and feedback, and, finally, in the Question stage he was tested by a number of randomly selected forced-choice questions.

The results of the 10 tests, corrected for guessing, have been recorded as a set of 10 percentage scores for each subject. The means and standard deviations of these scores across the subjects were plotted against the order in which the tests were taken. A plot of these results indicated that the subjects, from initially being totally unfamiliar with the system, gradually improved in their recognition skills. The results also revealed that all the sets of words were recognisable, at least by some of the subjects, after only 10 learning trials. Typical percentage scores were between 40% to 91%, where zero percent means chance expectation.

Individual differences were high for the first two tests but reduced to a standard deviation of about 25 percent for subsequent tests. A plot of the average scores of each subject (over the tests) against their age suggested a relation between performance and age. The graph showed a decline of the scores after an age of about 35 years for this relatively small group of subjects.

Three of the sets of words (for 'yes' and 'no') were selected from three widely different languages: Persian (Farsi), French and Czech in order to evaluate the language independence of the system. The high scores for these tests indicated that the system was, as expected, independent of the language used.

The matrices of confusion for the last two tests (the three-word and the four-word ones) revealed that the less audibly confusing words were also less visually confusing. This supports the fidelity of the visualiser system in converting auditory cues into their visual equivalents. A four-word forced-choice test without any preceding learning stage showed that when a user is already familiar with the gestures for a number of words, he can relatively easily recognise the gestures of unseen words which are audibly similar to the learned words. This again supports the fidelity of
Finally, to investigate memory retention, a ten-word test involving well-learned words was repeated (four times) during a period of 10 months with no further learning trials between. The high scores in this experiment are evidence for the persistence of the learned gestures.
Chapter 6

Conclusion and Future Work

6.1 Conclusion

6.1.1 Outline of the Problem Addressed

This thesis is concerned with the problem of providing speech processing aids for the deaf. Of the two main categories of sensory aids for speech perception, that of visual aids, formed the subject of this research. A review of the relevant literature indicated that two principal methods of visual aid have been under investigation from the beginning of this century. The first and oldest of these is the spectrogram-display aid. By this technique, a running spectrogram of ongoing speech is displayed in real time and a deaf user is expected to recognise the spoken words through watching the display. The problem with this type of aid is that learning to read such a spectrogram is very difficult, particularly in real time.

The second method is a type of dynamic display based on lipreading by which a cartoon-like animation of articulatory organs of speech is displayed, based on the input utterance. A number of researchers have studied the presentation of various speech features as an aid to lipreading. Since this type of aid utilises automatic speech recognition techniques to extract speech features, it has a poor performance in the presence of noise; moreover it is not independent of the language of the user.

A third technique, proposed in this thesis, is the visual presentation of the perceptually important features of an input sound, rather than specifically those of speech, in a dynamic format, as similar as possible to the auditory representation
of that sound. This idea is based upon the theory of the primacy of kinetic, rather than static, information in the perception of speech sounds [85]. This aid is thus a sound visualisation system which does not need to be concerned with the problem of automatic speech recognition, and as a result is independent of the language of the user.

Based on studies of the human auditory system and on the theories of hearing, sound perception and speech perception, it was decided that a potential solution to the problem of providing a practical, visual speech perception aid might be the *dynamical spectrogram*, a new technique of spectrography which is introduced by this work [93, 94].

### 6.1.2 Summary of the Research

To find an appropriate visual representation of sound, the human auditory system, as the most effective system of sound representation, was first studied. Then, based on the latest theories of hearing and techniques of auditory modelling, a simplified model of the human ear was developed. In this model, the outer and middle ears together have been simulated by a gain-controlled high-pass filter, and the inner ear has been modelled by a bank of band-pass filters with an approximately logarithmic distribution of their centre frequencies and with about 78 percent overlap of their pass bands. The outputs from all channels of this filter bank, after rectification and compression, are applied to a visualiser block.

For the design of this visualiser block, theories of sound perception and speech perception were studied. The perceptually important properties of sound, and their relations to the physical attributes of the sound pressure wave, were then considered, to produce an informative and recognisable image capable of representing the relevant instantaneous properties of the incoming sound. The format to be used for this image was finally decided to be a modification of the auditory-based spectrogram, also known as the cochleagram.

Then, a review of techniques for the visualisation of speech signals and current
knowledge about auditory nerve representation led to the production of a frame-locked sequence of 20-millisecond frames sampled from the cochleagram image. An animation of this sequence at the rate of 50 images per second displays both the time-varying and the time-independent information in the underlying sound at high resolution in real time. The resulting system translates a spoken word into a visual gesture, and displays a still picture when the input is a steady state sound.

Finally the implementation of the sound visualiser system was evaluated through several experiments undertaken by normal-hearing subjects. In these experiments, recognition of the gestures of a number of words, selected from those likely to be used in the course of seeking directions via a telephone conversation, was examined through a set of two- to four-word forced-choice tests.

6.1.3 Contributions and Findings

An outline of the contributions made by this thesis is as follows:

- A new type of visual aid for the deaf based on the primacy of dynamic information in the perception of speech has been proposed.
- A novel general method for visualising sounds, called an auditory-based dynamical spectrogram, has been developed.
- A novel technique for sound visualisation as an aid for the deaf has been developed, implemented, and examined.
- A linear-phase FIR filter has been used for the first time in this context, to simulate an auditory filter involving a much lower time duration than the optimum filter.
- An extended computational model of the auditory system involving variable parameters has been created.
- A software tool (Audvis) for animating any sequence of sound-and-image frames, recorded in a specific simple file format, has been developed and im-
6.1. Conclusion

implemented.

- An interactive software tool has been developed for running multiple-word, forced-choice tests involving blocks of sound-and-image sequences recorded in the file format for Audvis. This software can be set up through a resource file to run any number of tests involving any number of choices in three stages—learning, practice, and question.

The findings of the experiments carried out on the recognition of the visual gestures for a number of spoken words are as follows:

- All the gestures were recognisable, at least by some of the subjects.

- Typically, a subject can learn to distinguish between a small set of gestures after watching the gestures for only 10 instances of each word.

- The system is language independent.

There were also indications that:

- A subject was able to recognise gestures of unlearned words which have some correlation with previously learned words.

- There is some evidence for a relationship between the age and the learning abilities of users: the older the user the slower at learning the gestures. The resulting fall off was observed for ages between 35 and 46 years (the highest included in the study).

- There are relatively large individual differences in the ability to learn to distinguish visual gestures after only ten learning and practice trials, but such differences decrease among users more familiar with the system.

- A limited series of tests, involving a single subject over a period of 10 months without any intervening reinforcement, showed little or no loss of recognition ability for visual gestures, suggesting that they may persist strongly in long-term memory.
6.2 Future Work

Future work on this project can be divided into two categories: psychological investigations and technical developments.

6.2.1 Psychological Investigations

From a psychological point of view many questions remain to be investigated as follows:

- Is it possible to communicate using a system of this type in real time at a normal speed of conversation?
- Can profoundly deaf people learn this visual language in a reasonable time and with a reasonable effort?
- Is this system useful for providing speech training to the deaf?
- If the answers to these questions are positive, then what is the best way of teaching this 'language' to the users?

Discussion of these questions is beyond the scope of this thesis and needs to be undertaken by psychology researchers.

6.2.2 Technical Developments

Real-time Translation

From a technical point of view, the first and possibly most important problem is that the software developed so far is not capable of converting sound into visual gestures in real time. Although the system is able to display a previously converted and recorded stream of sound in real time, it is not as yet capable of working as a real-time translator. Indeed, the system has been developed in two parts: the translator and the animator (see Figure 4.12). Only the animator works in real time. Therefore the first aim must be to develop new software for the translator so that it too can work in real time.
At present the audvis display sub-system is capable of real-time display of dynamical spectrograms on a 50MHz Sun SPARC 5 workstation using approximately 1.5% of the available capacity (i.e. 0.05 Mflops). The sound capture and conversion has been designed for accuracy rather than speed, and runs off-line in approximately 80 times real time. With faster algorithms, such as FFT, this might be speeded up by a factor of perhaps 10 times. Thus of the order of 25 Mflops might be needed in a production system to perform both sound capture and conversion concurrently in real time.

Rate of Animation

Another problem for the sound visualiser, in the case of general sounds, relates to its inability to show sound periodicities of between 25 and 50 cycles per second. The rate of animation is 50 images per second, therefore periodicities of less than 25 Hz remain intact and are shown temporally—i.e. by differences between successive frames. Periodicities of more than 50 images per second are also fully displayed, although in a spatial format using shades of gray. But oscillations of between these two rates (i.e. periods of more than 20 and less than 40 milliseconds) cannot be displayed properly. Since the rate of 50 frames per second is less than the Nyquist rate (twice the repetition rate), these repetitions cannot be fully displayed in a temporal format. Depicting these periodicities in the spatial format, on the other hand, is also not possible because of the short time width of the display window (20 milliseconds) which is less than the period of the input sound so that the associated features appear to wander from frame to frame. This problem, however, fortunately has no effect on its use for speech visualisation, as the repetition rate of a speech sound is always more than 50 Hz.

Dynamic Range

The dynamic range of the intensity of the images is insufficient, compared with the huge dynamic range of the human ear in response to the intensity of sounds. Using colour instead of shades of gray in drawing the gestures might dramatically increase
6.2. Future Work

the dynamic range of the visual intensity of the displays. Also, creating a larger window, together with a proper drawing of the images, can contribute to a solution of this problem.

Image Format

The present format of the images, that is, a rectangular picture, with time depicted horizontally and frequency depicted vertically, is not necessarily the optimal format. Seeking more effective formats could be a subject for further research.
Bibliography


Bibliography


Bibliography


