SPEECH PERCEPTION IN VIRTUAL ENVIRONMENTS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE,
FACULTY OF SCIENCE
AT THE UNIVERSITY OF CAPE TOWN
IN FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
Johan Verwey
August 2005

Supervised by
Edwin H. Blake

Abstract

Many virtual environments like interactive computer games, educational software or training simulations make use of speech to convey important information to the user. These applications typically present a combination of background music, sound effects, ambient sounds and dialog simultaneously to create a rich auditory environment. Since interactive virtual environments allow users to roam freely among different sound producing objects, sound designers do not always have exact control over what sounds a user will perceive at any given time. This dissertation investigates factors that influence the perception of speech in virtual environments under adverse listening conditions.

A virtual environment was created to study hearing performance under different audio-visual conditions. The two main areas of investigation were the contribution of "spatial unmasking" and lip animation to speech perception. Spatial unmasking refers to the hearing benefit achieved when the target sound and masking sound are presented from different locations. Both auditory and visual factors influencing speech perception were considered.

The capability of modern sound hardware to produce a spatial release from masking using real-time 3D sound spatialization was compared with the pre-computed method of creating spatialized sound. It was found that spatial unmasking could be achieved when using a modern consumer 3D sound card and either a headphone or surround sound speaker display. Surprisingly, masking was less effective when using real-time sound spatialization and subjects achieved better hearing performance than when the pre-computed method was used.

Most research on the spatial unmasking of speech has been conducted in pure auditory environments. The influence of an additional visual cue was first investigated to determine whether this provided any benefit. No difference in hearing performance was observed when visible objects were presented at the same location as the auditory stimuli.

Because of inherent limitations of display devices, the auditory and visual environments are often not perfectly aligned, causing a sound-producing object to be seen at a different location from where it is heard. The influence of audio-visual integration between the conflicting spatial information was investigated to see whether it had any influence on the spatial unmasking of speech in noise. No significant difference in speech perception was found regardless of whether visual stimuli was presented at the correct location matching the auditory position, at a spatially disparate location from the auditory source.

Lastly the influence of rudimentary lip animation on speech perception was investigated. The results showed that correct lip animations significantly contribute to speech perception. It was also found that incorrect lip animation could result in worse performance than when no lip animation is used at all.

The main conclusions from this research are: That the 3D sound capabilities of modern sound hardware can and should be used in virtual environments to present speech; Perfectly align auditory and visual environments are not very important for speech perception; Even rudimentary lip animation can enhance speech perception in virtual environments.

Acknowledgements

I would like to thank the CAVES consortium for providing equipment, payment of experimental subjects and for financing not only my bursary but also my two-month visit to Boston University, USA.

I would also like to thank John Turest-Swartz and his staff from the Collaborative African Music and Arts Directive for providing a recording studio and allowing the use of their offices while conducting my experiments.

The guidance and advice from Prof. Barbara Shinn-Cunningham, Antje Ihlefeld and Tim Streeter from Boston University while I visited the Boston Hearing Research Center is much appreciated. Your input has proved to be invaluable to the success of this research. Most importantly I would like to thank my supervisor Prof. Edwin Blake for his continual support during the year and allowing me the freedom to explore my own interests.

Contents

List of Tables

List of figures

Chapter 1
Introduction

Rich, life-like audio plays an important role in creating a feeling of immersion in virtual environments [44]. The auditory environment could consist of music, sound effects, spoken dialog, and various background noises. A multitude of sounds presented simultaneously can however be distracting when listening to someone speaking [2]. Virtual-reality applications often rely on speech to convey important information. Be it instructions in a training simulation, teaching in educational software or dialog in an interactive computer game. Since the user can move and interact freely in a virtual environment, sound designers do not always have control over the sounds the listener will perceive at any given time. Adverse listening conditions are therefore sometimes unavoidable. This differs greatly from the case with cinema where post-production sound editing ensures that the optimal listening experience is created. This dissertation investigates some factors that influence speech perception in virtual environments under adverse listening conditions.

We first considered the effect of directional auditory cues on hearing performance. Although it has been shown that directional information can be used by the auditory system to enhance speech perception [27], there are fundamental differences between the techniques investigated in the literature and those typically used in virtual environments. In virtual environments, directional cues are usually generated in real-time by consumer sound cards and presented either over stereo headphones or a surround sound speaker system. These methods were compared with the traditional research method of pre-computing stereo sound files with directional information. The influence of visual cues for sound source location has not received much attention in the literature either. We wanted to establish whether the correct visually induced expectation of a voice's position contributed to the perception of speech. Conversely, an incorrect expectation of a voice's location may have a negative impact. Lastly we wanted to determine how much the simple lip animations used in virtual environments could contribute to a clearer perception of spoken dialog.

1. Speech in virtual environments

The human perceptual system makes use of both auditory and visual information to understand speech in everyday life. Spatial cues for the talker's location are provided in both modalities. The talker's lip movement also conveys additional visual information. All this information may have to be convincingly reproduced in virtual environments in order to ensure adequate speech perception.

Research in virtual auditory environments has shown that it is possible for sounds to be presented over stereo headphones in such a way that they are perceived as coming from any position in 3D space [6]. Digitized sound data are manipulated to create a stereo sound file with the separate channels representing the sound that would be perceived at each ear. Slight changes in level, timing and spectrum at each ear will cause virtual sound sources to be perceived at different locations in the 3D space around the listener when played over stereo headphones. This is referred to as 'sound spatialization' or more commonly, '3D sound'. Chapter 2 presents more background information on the exact calculations involved to produce 3D sound.

Spatialized sound can influence speech perception in virtual environments. Multiple sounds presented simultaneously places a strain on the auditory system and make it difficult to pay attention to all the sounds presented. The human perceptual system has the remarkable ability to pay selective attention to a sound of interest in the midst of other competing sounds. This is often called the "Cocktail-party

Effect" [19, 30]. This ability allows listeners to attend to a specific voice while ignoring other voices and background noise. More information on this phenomenon is presented in Chapter 2. For now it will suffice to say that one of the contributing factors in distinguishing sound sources is their physical location [13]. A difference in the location of sound sources greatly enhances the intelligibility of speech in the midst of a masking noise or other competing voices. This is referred to as a *spatial release from masking* or *spatial unmasking* [30]. This benefit extends to virtual auditory environments where virtual sound sources are spatially separated from one another [27]. Placing sounds at different locations in a virtual environment can result in a spatial release from masking which enhances speech perception.

In most experiments involving spatial unmasking, spatialized sound stimuli are first pre-computed and then presented to listeners over stereo headphones. Creating such files is a computationally expensive process. Sound hardware employed in today's virtual environments however, has dedicated digital signal processors designed for this purpose [49]. Consumer sound cards typically found in personal computers can produce 3D sound in real-time. This technology however does have some restrictions that limit the quality of the sound spatialization [6]. Chapter 2 provides more information on these limitations.

Modern sound processors not only allow sound to be spatialized over stereo headphones but also with the use of surround sound speakers. When multiple speakers are used as an auditory display, a different technique called *amplitude panning* [69] is used to spatialize sound. This technique presents a sound from multiple speakers simultaneously but the percentage gain at each speaker depends on how well the physical speaker location matches the desired direction of the virtual sound source.

Virtual Environments (VEs) are of course not limited to audio and generally also provide visual representations of sound producing objects. This can be presented to the user through the use of a computer monitor, head-mounted-display (HMD) or projection and allows users to see where a sound is coming from in addition to hearing it. This constitutes another possible conceptual cue to the position of the sound. The importance of this cue to speech perception has not received much attention in the literature. Cocktail-party experiments indicate that humans make use of spatial information to disambiguate different sound sources and to pay selective attention to an area of interest [2]. The spatial information is derived from both visual and auditory localization cues. Auditory sound localization is considered poorer than its visual counterpart [11]. However, when both the auditory and visual modalities are presented together, sounds can be localized with the same spatial precision as vision [26]. Ebata has shown that hearing ability increases when paying attention in the direction of the sound source [29]. Research by Reisberg found that a visual localization cue helps to focus directional attention on the sound source and contributes to the spatial unmasking of speech in the presence of competing voices [73].

VEs differ from reality in that the visual and auditory environments are often not aligned. Chapter 2 explains how visual and auditory objects can sometimes be displayed in different locations when the display devices are not configured correctly. Fortunately the human perceptual system is capable of adapting to incongruent visual and auditory information [70]. When the visual representation of a sound-producing object is displaced from the auditory representation, the listener usually perceives the sound as originating from the same position as the visual representation. This is called the "Ventriloquist Effect" [20]. Research by Driver has shown that *even this illusionary change* in perceived position of the audio source can result in a spatial release from masking [26]. When listening to a target speech source, the displaced visual representation seems to draw auditory attention away from the masking noise. Chapter 2 presents a more thorough account of previous research in this area. The misalignment of auditory and visual environments results in auditory and visual objects to be presented at different spatial locations. This could either be beneficial or detrimental to speech perception depending on whether the visual representation is shifted closer to, or further from an auditory masking sound.

The ability to read a speaker's lips has a significant impact on speech perception [84]. In other forms of media like television and movies, this visual cue is readily available. For virtual environments lip animations have to be created for every character that will be speaking. Since this can be a time-consuming process, most applications provide only very rudimentary lip animations, if at all. While some lip movement certainly contributes to realism, it is uncertain whether this can contribute to speech perception. Most studies involving lip reading make use of video streams of real faces. It has been shown that video streams with frame rates as low as five frames per second can still contribute to

speech perception [34]. Carefully constructed lip animations may therefore also yield similar results if the animated lips are a close enough approximation of real lips and are properly synchronized with the audio. Since many VEs do not make use of accurate lip animation, it is also necessary to determine whether incorrect lip animation could have an effect on speech perception. Studies of "The McGurk Effect" [60] has shown that completely different words can be perceived when contradictory visual information is presented together with auditory speech [24]. The addition of a noise masker may further aggravate this effect since the stronger source is usually favoured when two sources of information, in this case visual and auditory, conflict [84]. Incorrect lip animation may therefore result in worse speech perception than when no lip animation is present.

2. Aims

This research investigates several aspects unique to virtual environments that may have an influence on speech perception under adverse listening conditions. Many previous studies investigating the influence of spatial separation and visual cues on speech perception have been conducted by using distracting speech as a masking sound to create adverse listening conditions. Competing sounds VEs are however not limited to speech sources. Adverse listening conditions can be introduced by background music, special effects or ambient sounds like wind and water. In this research we will make use of broadband white noise as a masking sound to determine whether findings in the literature are applicable to a broader range of masking sounds.

Many interactive virtual-reality applications make use of computer sound hardware for creating 3D sound [49]. Spatialized sound produced in real-time with modern hardware differs from the traditional method of pre-computing 3D stereo sound files for such experiments. Our first objective is to determined to what extent these differences influence speech perception and whether a spatial release from masking can be obtained using this technology.

Virtual environments that make use of spatialized sound provide additional auditory cues for sound source location that are not present in other forms of media like film and television. If presented correctly, the auditory localization cues will match the visual localization cues. Research by Reisberg has shown that a visual localization cue can contribute to enhance speech perception in the presence of competing voices [73]. In order to determine whether these results are relevant to VEs we need to reproduce these results in a VE using a masker that represents a broader range of sounds. Our aim is to establish whether speech from a visible object matching the auditory position could be more clearly perceived than speech from an invisible object. This would give an indication of the relative contribution of the visual localization cue under adverse listening conditions.

In the majority of virtual environment applications, the auditory and visual environments will not be aligned. The sound of a person speaking will typically not come from the same direction that is observed on the visual display. Our aim is to determine whether the misalignment of auditory and visual information has any impact on speech perception. We propose that incorrect visual directional information can draw auditory attention towards the visual representation of the speaking character. Where Driver [26] has demonstrated this when using a second voice as a distracting sound, our studies will be conducted in an immersive virtual environment using broadband noise as a masking sound.

Our final aim was to determine how simple lip animations used in VEs influenced speech perception. The visual information present in even simple animations may be enough to disambiguate unclear sounds presented in noisy conditions. If the animation does not match the actual phonemes being articulated, it is also possible that the contrary visual information could have a negative perceptual effect.

The four main aims of this research are therefore:

- To determine whether modern computer sound hardware can produce a spatial release from masking. The performance of both headphone and surround sound auditory displays using this technology needs to be compared to that observed in previous studies involving pre-computed spatialized speech over headphones.
- To determine the contribution of visual cues for sound localization on hearing performance in virtual environments.
- To establish whether incongruent auditory and visual spatial information in virtual environments has any effect on speech perception.
- To determine whether rudimentary lip animations used in virtual environments contribute to speech perception and whether incorrect lip animations have a negative impact.

The overall aim is to obtain a better understanding of the factors involving speech perception in virtual environments under adverse listening conditions.

3. 1.3 Overview of Experiments

Eleven subjects participated in perceptual experiments designed to investigate the following hypotheses:

- Modern sound hardware is capable of producing a spatial release from masking for both headphone and surround sound displays.
- The presence of a visual cue aiding auditory localization will enhance speech perception.
- Incongruent auditory and visual spatial information will contribute to improved hearing performance when the visual target shifts auditory attention away from a masking noise, but will degrade hearing performance when it shifts attention in the direction of the masker.
- Rudimentary lip animations matching the auditory speech improve hearing performance, but unmatched lip movement is detrimental to speech perception.

Four experiments were designed to test these hypotheses. In all experiments subjects were required to identify certain target words in a spoken sentence. A masking noise was simultaneously presented with the target sentence to create adverse listening conditions. Different auditory, visual and spatial configurations of the presented stimuli were used to either verify or refute each hypothesis. The design of the experiments and the number of subjects was in accordance with standard practice in audio perception research [51]. This will be discussed in Chapter 3.

Experiment 1: Three different auditory conditions were compared. The first two conditions both made use of real-time sound spatialization. Stereo headphones were used in the one condition and surround sound speakers in the other. The third condition used pre-computed spatialized sound stimuli and used stereo headphones for an auditory display. For all three auditory conditions the target and masking noise was presented either co-located or spatially separated to determine the amount of spatial unmasking obtained for each auditory condition.

Experiment 2: Where the first experiment was conducted in a pure auditory environment, the second experiment investigated the effect of audiovisual interaction on speech perception. Visual objects representing the target and masker was now presented with the auditory sound sources. Improved hearing performance in the audiovisual condition would confirm our second hypothesis.

Experiment 3: This experiment extended the previous one by additionally providing incorrect visual localization cues. The incorrectly placed visual objects were designed to either draw auditory attention away from the masking noise or to shift the attention towards the masker.

Experiment 4: The last hypothesis was investigated by presenting three different visual conditions. In the first condition correct lip animation was presented with the auditory target. The second provided incorrect lip animation while the last condition provided no lip animation at all.

4. 1.4 Outline of dissertation

Chapter 2: This chapter presents some background information on speech perception in virtual environments. The different cues that contribute to the human perception of spatial sound are discussed as well as how these cues are reproduced in virtual environments. Previous research into spatial unmasking and audiovisual interaction are also presented.

Chapter 3: An overview of the methodology used to conduct speech perception experiments is first provided in this chapter. This is followed by a detailed design for each experiment.

Chapter 4: This chapter first analyses the gathered experimental data to ensure that there were no gross differences in subjects' ability to perform the task. General trends in subject responses are also compared with previous studies that used a similar methodology. Finally the results from each experiment is presented and analysed.

Chapter 5: Finally the main conclusions from this research are summarized and the contributions of this work are highlighted. The chapter concludes with a discussion of possibilities for future research.

Appendix: A detailed repeated measures ANOVA analysis of each of the four experimental results is provided as an appendix.

Chapter 2
Background

Speech perception in virtual environments is influenced by a variety of factors. This chapter first provides some background information on how sound is localized by the human auditory system. We then show how 3D sound is produced in virtual environments while highlighting some limitations of modern sound hardware. The benefits of spatialized sound in relation to speech perception and the influence of audiovisual factors are then discussed and previous research in these fields is presented. Finally we evaluate two commonly used speech intelligibility tests.

2.1 Sound localization cues

This section presents an overview of 3D sound localization. More detailed information on this topic can be found in [11].

In real life we can perceive the direction of a sound source. We know whether a familiar sound occurred in front, behind, to our left or right. We can also estimate the distance and to a lesser extent the elevation of the sound source. The brain primarily makes use of three cues to determine the direction of a sound source: Interaural intensity difference (IID), interaural time difference (ITD) and the head related transfer function (HRTF).

IID refers to the difference in intensity of the sound wave that reaches each ear. Sound that originated from the listener's left side will be at a lower intensity at the listener's right ear due to the acoustic head shadow, which obstructs the sound. This is illustrated in Fig 2.1. Sound originating from in front of the listener has zero IID. The greater the difference in intensity, the more the perceived location will shift in the direction of the louder ear. This cue is most effective for frequencies higher than about 1.5 kHz. Lower frequency sound waves will diffract around the listener's head, thereby minimizing the intensity differences.

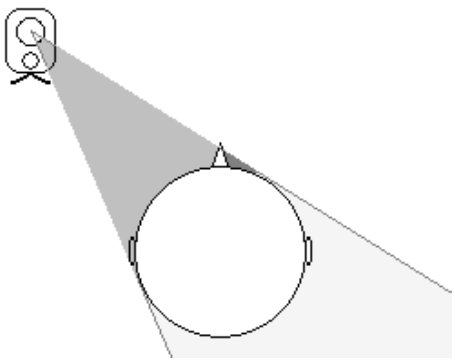

Fig 2.1: Interaural intensity difference (IID). High frequency sound waves are attenuated by the head shadow effect.

The interaural time difference (ITD) refers to the difference in the time the leading wave front reaches each ear. Sound originating from the left will reach the left ear first as shown in Fig 2.2. By the time it reaches the right ear it will be slightly out of phase. The brain uses this time and phase difference to determine which direction the sound came from. When the sound comes from in front, there is no phase difference between the ears. The greater the difference in phase, the more the perceived location will shift in the direction of the ear that received the leading wave front first. This cue is most effective for sound frequencies lower than 1.5 kHz. Above this frequency the period of the wave becomes smaller than the size of the head and the ears can no longer use the phase information to determine which wave is the leading wave front. Lower frequency waves have longer periods and therefore the phase difference will be more noticeable.

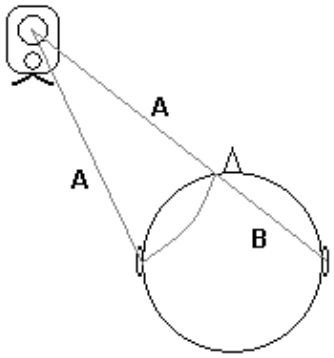

Fig 2.2: Interaural time difference (ITD). Sound needs to travel a distance of (A) to reach the left ear but a distance of (A) + (B) to reach the right ear. This results in the sound wave reaching the right ear slightly later than the left ear.

The frequency range of normal speech is approximately 200 Hz – 5 kHz [77]. Both the IID and ITD cues can therefore effectively be used to localize speech. These cues however are inherently ambiguous. For a spherical head, identical values of IID and ITD can point to sound sources located anywhere on the surface of a cone extending from the ear. This is commonly referred to as the 'cone of confusion', which is illustrated in Fig 2.3. Points (a) and (b) illustrates front/back ambiguity while (c) and (d) illustrates elevation ambiguity. If we were limited to these two cues, we would not be able to distinguish between sounds coming from anywhere along the surface of this cone.

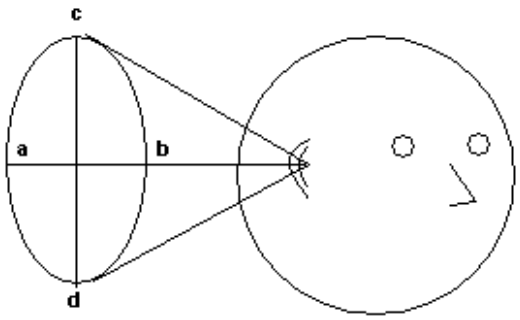

Fig 2.3: Interaural time and intensity difference cues are ambiguous. If only these cues were used, the auditory system could only place a sound on the surface of this imaginary cone

called the 'The cone of confusion.' Points (a) and (b) illustrate front/back ambiguity while points (c) and (d) illustrates elevation ambiguity.

In addition to the influence of ITD and ILD cues, the head, torso and pinnae (outer ears) also have a significant effect on the perceived sound. The spectral filtering of a sound source that occurs before the sound reaches the eardrum is referred to as the Head Related Transfer Function (HRTF). A complex sound will typically consist of a wide range of different frequency sound waves. Some of these frequencies will be filtered when the head, torso or outer ears obstruct sound waves. Sounds originating from the back of a listener will contain significantly less high frequency sound waves than the same sound that originated from the front. The shape of the ears also causes slight changes in phase of sounds, depending on the direction and elevation of the sound source. While these spectral cues by themselves are not a very strong cue in isolation, they serve to alleviate the front/back and up/down ambiguities.

Humans are comparatively poor at determining the elevation of a sound. Changing the elevation of a sound source only produces subtle changes in the perceived frequencies. Since human interaction with their environment predominantly happens in the horizontal plane, this is not really a significant drawback. Birds of prey however, are much more reliant on their ability to localize sounds in the vertical dimension. The ears of night owls are placed slightly asymmetrically in the vertical direction. This causes differences in interaural intensity providing them with an additional cue for localizing a sound.

Sound intensity is measured in decibels (dB). The dB is a logarithmic unit used to describe the ratio between the sound pressure level of the sound source and a given reference level. If P_1 and P_2 represents the sound pressure levels resulting from two sound sources, the difference in decibels between the two is defined to be:

$$10\log_{10}\left(P_2 / P_1\right) dB$$

From this equation a sound that is twice as loud as the reference sound would be 3dB louder.

Standard reference levels exist to give an absolute indication of sound pressure level. The standard reference sound pressure level (SPL) is 0.02mPa. An absolute sound level of 0 dB SPL would then have the same intensity as this reference level.

Sound pressure level reduces with distance from the sound source according to the inverse square law. This states that the sound intensity is inverse proportional to the square of the distance from the point source.

$$I\alpha(1/r^2)$$

This equates to a reduction of about 6dB for every time the distance from the source is doubled. The sound pressure level is one of the most important cues when estimating the distance of a sound source. A very loud sound is perceived as being closer than a similar soft sound. This cue in isolation is not enough to provide a correct distance estimate. Familiarity with the sound plays a significant role in distance estimation. Because we know that the sound of a diesel truck is very loud when you are standing right next to it, we can estimate the distance based on the loudness of sound that we hear. If the sound of the truck is very soft, it has to be far away. As the loudness increase, the perceived distance will decrease. Familiarity also provides us with the conceivable ranges of the sound source distance. If we can hear a bee at all, it has to be within a few meters from our ear. A very faint buzz would therefore cause us to estimate the distance at a few meters. A very loud buzz would cause us to believe the bee is very close to our head. In the same way a very faint idle of a diesel truck (at the same loudness as the bee) would place the truck at a few hundred meters from the listener.

The spectral content of a sound will also vary as a function of its distance. Molecular absorption of the air, atmospheric conditions and the curvature of the wave front modify the spectral content of the sound source. High frequency components will dissipate faster than low frequency components and therefore

sounds that have travelled further will contain less high frequency energy. When compared to loudness and familiarity, the change of spectral content is a relatively weak cue though.

2.2 3D sound reproduction

The cues that allow us to localize a sound in real life holds the key to producing 3D sound in virtual environments. If we can reproduce the exactly correct acoustic field at each ear, the sound should then be perceived as if coming from the original position. ITD and IID can easily be simulated with headphones to provide directional cues. It only involves a time delay and a scaling of the sounds received at each ear [77]. From Fig 2.3 we saw however that these cues are only useful to localize sounds to the left or right of the listener. To accurately localize a sound it is necessary to simulate the effect of the HRTF as well. This requires some basic knowledge of digital signal processing and Fourier analysis. A brief overview is provided in this section. More detailed information can be found in books on these topics [6, 56, 41].

2.2.1 Fourier analysis

A continuous mathematical function can be approximated by linear combinations of sine and cosine functions called a Fourier series.

$$s(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(nt) + \sum_{n=1}^{\infty} b_n \sin(nt)$$

For sound processing, an acoustical signal can be represented as a sum of pure tones, each with its own frequency, amplitude and phase. The Fourier coefficients a_n and b_n represent the contribution each frequency makes to the total sound wave. Fourier transforms are used to obtain a spectral analysis of a sound source, revealing the amount of energy present for different frequencies in the sound. Where an acoustical signal represents the amplitude of a waveform as a function of time, the Fourier transform represents the amplitude as a function of frequency. It effectively transforms a function in the time domain to a function in the frequency domain.

$$S(f) = \int_{-\infty}^{\infty} s(t) e^{-i2\pi ft} dt$$

The inverse transform is:

$$s(t) = \int_{-\infty}^{\infty} S(f) e^{i2\pi ft} df$$

Acoustical signals represented on a digital computer are not continuous functions. For digitized sound data a derivative called the Discreet Fourier Transform (DFT) or the more computationally efficient fast Fourier Transform (FFT) are used. Let $x(n\mathrm{T})$ represent the discrete time signal, and let $\mathrm{X}(m\mathrm{F})$ represent the discrete frequency transform function.

The Discreet Fourier Transform is:

$$X(mf) = \sum_{n} x(nT) e^{-inm2\pi FT}$$

The inverse discrete transform is:

$$x(nt) = \frac{1}{N}\sum_{m} x(mF)e^{inm2\pi FT}$$

2.2.2 Convolution

Finite impulse response (FIR) filters are used to modify a digital input signal to produce a different output signal. Discreet input data samples *x (n)* are transformed to output data *y (n)* through a process called convolution which is defined by

$$y(n) = \sum_{k=0}^{L-1} h(k)x(n-k)$$

The term *h (n)* represents sequence of *L* filter coefficients. If the input data *x (n)* consisted of *N* samples, the length of the filtered output *y (n)* is a sequence of (*N+L-1*) numbers [56].

If *h (n)* is unknown, it can be inferred from the output *y (n)* by providing an *analytic impulse* $x(n) = (1,0,0,0,0,...)$ as input. If *h (n)* does not change over time and the system is linear then the impulse response *y (n)* is equal to *h (n)*. This property is of importance when designing digital filters that will simulate the effect of the HRTF [4].

In addition to analysing the frequency content of sounds, Fourier transforms can be used to speed up the computationally expensive convolution procedure. It turns out that multiplication in the frequency domain is equivalent to convolution in the time domain. A FFT of both the input data *x (n)* and filter coefficients *h (n)* in the time domain are first computed. The resulting *X (z)* and *H (z)* sequences in the frequency domain are then multiplied to produce *Y (z)*. Finally an inverse Fourier transform is used to convert *Y (z)* back to the time domain resulting in the 'convolved' *y (n)* [41].

2.2.3 Simulating the head-related transfer function

The head-related transfer function can be thought of as a filter that causes changes in the frequency content of a sound in addition to causing interaural time and intensity differences. Digital filters that approximate the HRTF can be computed by measuring the response of an analytic impulse at both ears. The impulse is played through a loudspeaker placed at the desired spatial position relative to the listener. The impulse response *y (n)* is measured with probe microphones placed inside both ear canals. Since an analytic input signal was used, the digital filter coefficients *h (n)* that would produce the impulse response would be equal to *y (n)*. When the filters of each ear are applied to any input sound data, the resulting sound would have the same spectral modifications and time delays at each ear that would be perceived if the actual HRTF was used. Since the spatial cues present in natural hearing have been reproduced, the sound is perceived as spatialized. Note however that each simulated HRTF filter only represents a single position in 3D space. Separate measurements needs to be made for all directions surrounding the listener that are to be simulated [6].

Longer input sequences like the maximum-length sequence (MLS) [72] and the Golay sequence [92] are often used instead of the analytic impulse. The impulse response is extracted from the resulting output by cross-correlating it with the input sequence. These methods result in impulse responses with a higher signal-to-noise ratio than those obtained from using an analytic impulse.

Impulse responses are usually measured on a dummy head such as KEMAR, a standard audiological research mannequin manufactured by Knowles Electronics. Research has shown that there are

significant differences between HRTFs measured on a dummy head such as KEMAR and those measured on different individuals [90]. The differences are most evident for higher frequencies. Broadband noise refers to sounds that have a random amount of energy present at every frequency in a wide range. Localization performance of broadband noise increases when subjects are listening through their own ears or with HRTFs measured on their own ears. Speech however has relatively lower frequency content compared to broadband noise and it has been shown that these differences have no dramatic effect on subjects' ability to either localize or recognize speech [7, 27].

2.3 Modern sound hardware

Creating 3D spatialized sound can be a computationally expensive process. Digital signal processors present on modern sound hardware are capable of performing these calculations in real-time. In order to do this they do however impose certain restrictions, which may affect the quality of the result [77]. The number of filter coefficients used during convolution has an impact on the perceptual fidelity of the reproduced sound [58]. Every sound that is presented in 3D must be convolved with the HRTF filter for each ear. High quality impulse responses measured at a sample rate of 44.1 kHz could contain 512-1024 filter coefficients. The computational resources to perform spatialization for many sound sources at this quality could be prohibitive. SLAB, a software-based system for interactive spatial sound synthesis, uses 128 filter coefficients for every 3D sound [58]. A software implementation however strains the computational resources of the main processor. As the number of sounds that need to be presented 3D increase, the use of dedicated hardware becomes more important. Convolution on a digital signal processor is accomplished through a delay and gain operation for each of the filter coefficients. This delay-gain combination is referred to as a filter tap. Data reduction techniques are often used in hardware implementations to limit the amount of filter taps [6]. The Creative X-Fi processor, currently one of the most advanced consumer sound processing chips, only uses 48 filter taps for HRTF processing [21]. With these simplifications, consumer sound cards found in most personal computers are capable of producing 3D sound for up to 64 simultaneously presented sound sources [65].

To simulate 3D sound in real-time, it should be possible to spatialize a sound at any direction and distance. Because of memory restrictions however, only a limited number of HRTF filters can be used. Usually only filters for a few discreet directions at a fixed distance from the user are available. To simulate other directions and distances, the filter coefficients of adjacent directions are interpolated and the amplitude adjusted according to the inverse square law [77]. The filter coefficients for an arbitrary location are therefore less accurate for real-time processing.

Sound can also be positioned using multiple speakers surrounding the listener. Vector-based amplitude panning is used to create virtual sound sources in between physical speaker locations [64, 69]. When sound is presented with the same gain at two equidistant speakers, the sound will be perceived as if coming from the position in the exact centre between the two physical speakers. By changing the relative contribution of each source, the image can be moved between the two speaker locations. With multiple speakers surrounding the listener, such as in a 5.1 speaker setup, sounds can be presented from any direction in the horizontal plane by panning among adjacent speakers. When headphones are used as an auditory display, sounds are often perceived to be located inside the listener's head [6]. Surround sound speakers help listeners to externalize sounds to positions outside the head. It does however have the disadvantage that it is difficult to reproduce sounds close to the listener within the boundaries of the physical speakers [35].

By using amplitude panning, the costly calculations involved in producing spatialized sound over headphones are avoided. Speakers need to be carefully positioned at equal distances from the listener and the listener should sit at the centre position at all times for correct sound spatialization. It has been shown that incorrect speaker placement can cause not only incorrect spatialized sound, but also result in degraded speech comprehension [77].

2.4 Unmasking of speech

To understand how 3D sound can benefit speech perception one should consider how the human perceptual system group low-level auditory and visual information like sound and images into a higher-level perceptual stream like speech. This is best explained by a process called "Auditory scene analysis" and demonstrated by the "Cocktail-party Effect".

2.4.1 Auditory scene analysis

In an auditory environment, the brain is presented with a mixture of the acoustic energy of multiple sound events. In order to make sense of the auditory environment we find ourselves in, we need to be able to make a distinction between different sound sources. The process of separating the combined acoustic energy into different perceptual streams is referred to by Bregman [13] as auditory scene analysis. The auditory system performs this task by using a primitive process of auditory grouping as well as a higher-level process that incorporate prior knowledge of familiar sounds.

At the primitive level the auditory system will group sounds together that most likely form part of the same sound event. Various factors influence this grouping process:

- Complex tones with spectral similarity are grouped together. The auditory attribute 'brightness' is related to the proportion of energy distributed on high frequencies. In contrast, 'dullness' relates to the distribution of energy in lower frequencies. Several bright sounds are more likely to form part of the same sound event while dull sounds would form a separate event.
- Tones that have onsets with close time proximity most likely form part of a new single auditory event.
- Pure tones that are harmonics of each other most likely form part of a single complex tone.
- Tones that have a similar change in frequency over time are grouped together.
- Tones that come from different spatial directions likely originate from different sound events.

Sounds that are not likely to form part of the same event are separated. These cues help to distinguish between multiple simultaneously presented sounds. When sounds are reproduced in a virtual environment, all of these cues except the last one will naturally be present. Spatialized audio is needed to reproduce the directional cue.

2.4.2 The Cocktail-party Effect

The human brain makes extensive use of auditory scene analysis in everyday life to discern between different auditory sources. A classic example of this has been dubbed "The cocktail-party effect." [19]. This refers to the human ability to separate the voices of multiple speakers in order to pay selective attention to a single speaker. The same ability also allows us to hear a voice in the presence of a masking noise. In addition to the primitive level grouping mentioned above, higher-level processes are also used to distinguish between voices. These include:

- Lip movement and gestures that correlate with sounds.
- Voices differ in mean pitch, speed and timbre.
- Voices have different accents.
- Words in a sentence have certain transition probabilities.

Many studies have investigated the improvement of hearing performance as a result of directional separation of sound sources [27, 29, 45]. Some studies investigated this effect by studying the ILD and ITD cues in isolation [55] while others conducted experiments in the free field [32, 45] or by using a virtual auditory display [27, 9].

Pre-convolved stereo sound files were used in all studies that made use of a virtual auditory display. Studies conducted in the free field made used of physical speakers to produce directional sounds. None of these experiments made use of modern sound hardware to spatialize sounds in real-time with headphones or with the use of surround sound speakers.

2.4.3 Masking sounds

A masker can be defined as any competing sound that makes it difficult to pay attention to a certain target sound. These masking sounds can be classified into energetic or informational maskers depending on the way they achieve their goal.

Energetic maskers
The human auditory system transforms acoustical signals into neural impulses that are interpreted by the brain and perceived as hearing [6]. Different neural channels are sensitive to different frequencies and effectively separate a complex sound into different frequency bands in a way similar to Fourier analysis [41]. Simultaneous sounds with similar frequency content causes overlapping excitation patterns in the auditory nerves [28]. This interference makes it difficult to distinguish between sounds with spectral overlap and is referred to as energetic masking. White noise is an example of an energetic masker. It contains sound energy across all frequencies in the range of human hearing and effectively masks any target sound.

Informational maskers
While informational maskers are less clearly defined, they usually refer to sounds that produce adverse hearing conditions regardless of whether the spectral energy of the target and masker overlap. It is often referred to as non-energetic masking [28]. Where energetic masking is caused by competition during primitive processing in the auditory system, informational masking occurs during higher-level processing. It interferes with the listener's ability to follow patterns in the target within a complex masker [33]. In his study of the cocktail-party effect, Cherry [19] has shown that listeners find it very difficult to attend to one talker in the presence of a second voice. One study found that an energetic masker needed to be at least 6dB louder than the target to be effective. In the same study, a voice acting as an informational masker could be up to 9dB softer as the target voice and still be distracting [16].

2.5 Audio-visual factors

The study of speech perception is not limited to auditory perception. The interaction between auditory and visual modalities has received much attention in the literature. Chen *et al*. [18] presents a review of some of this research. Studies have shown that visual cues such as lip movement and localization cues can influence speech perception [84, 26]. These cues are available in virtual environments but are not always accurately reproduced. A misalignment of the auditory and visual environments causes inconsistent auditory and visual cues for sound localization. Also, for most virtual reality applications, only rudimentary lip animations are provided, if at all. These differences may have an effect on speech perception in virtual environments.

2.5.1 Sound localization cues

Sound localization is used in the movie industry to combine auditory and visual objects in one perceptual stream. Hearing the sound of roaring engines move from left to right as an airplane flies across the cinema screen enhances the viewer's "suspension of disbelief" [46]. An exception is made for the localization of speech sounds. During movies the camera angle often changes abruptly. If the speech sources are matched to the visual position of actors, this causes noticeable jump in the auditory location whenever the scene is cut to a new camera angle. For this reason motion pictures typically use

only the centre channel for dialog even though the visual and auditory positions of the speech do not match.

Interactive computer games typically make use of spatialized sound to position speech sources. In the game 'Halo' the same problem of a sudden jump in speech source location occurred during cut-scenes when the camera angle changed [66]. During game play however the user determines the camera angle and this problem is less evident. Interactive computer games and other virtual reality applications can therefore still benefit from localized speech sources matching their visual counterparts as long as the user makes no sudden movements while listening to speech.

Previous studies have investigated whether a visual cue for sound localization influences speech perception. Reisberg demonstrated a small improvement in hearing performance as a result of visual localization cues [73]. Subjects were required to listen to the target speech presented from a speaker that was spatially separated from a second speaker. This speaker provided informational masking by presenting distracting speech sentences. When the speaker was not hidden behind a curtain, the additional visual cue for the sound's location resulted in slightly better hearing performance.

Providing accurate visual localization cues may therefore benefit speech perception in noisy virtual environments. Unfortunately auditory and visual objects can sometimes appear in different locations when the auditory and visual environments are misaligned. Fig 2.4 shows a perfectly aligned audio-visual environment.

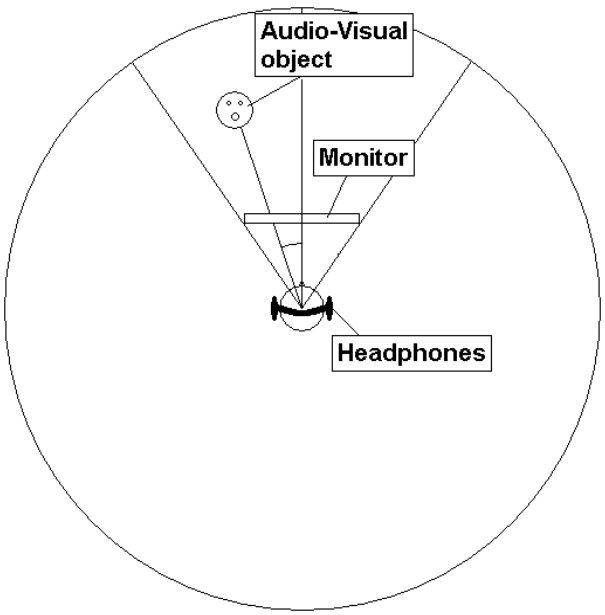

Fig 2.4: A simple two-dimensional illustration of how the user perceives the virtual environment through headphones and a computer monitor. When the auditory and visual environments are perfectly aligned, the user will perceive the sound and its visual representation in the same location.

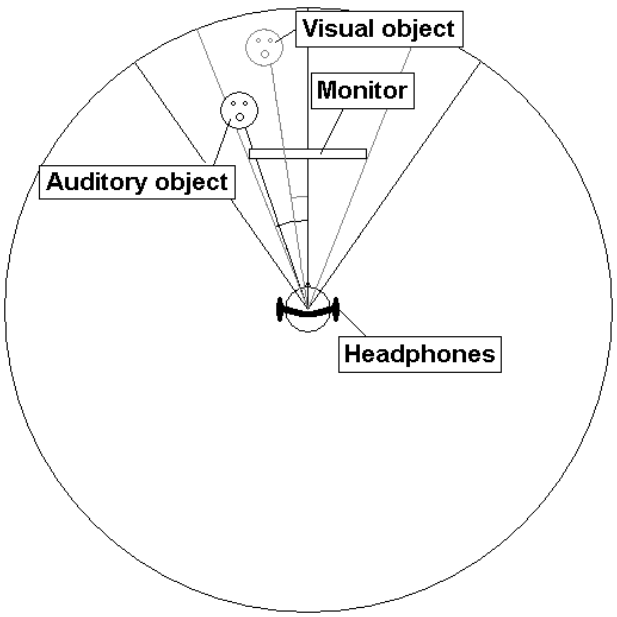

Fig 2.5: When the user is not sitting at the correct distance from the monitor, the visual representation of the object will be perceived at a different angle than the auditory object.

The auditory environment can display objects at any angle relative to the listener. The visual environment is limited by the width of the display area and can only display objects within the field of view of the user. Foley *et al* [31] presents a thorough account of how a 3D environment is mapped onto a 2D display area. For users to perceive visual objects at the correct angle, they have to sit at a precise position behind the computer monitor that is determined by the field of view being represented and also

the width of the monitor. If the user moves away from this position, the visual object is perceived at an incorrect angle as shown in Fig 2.5.

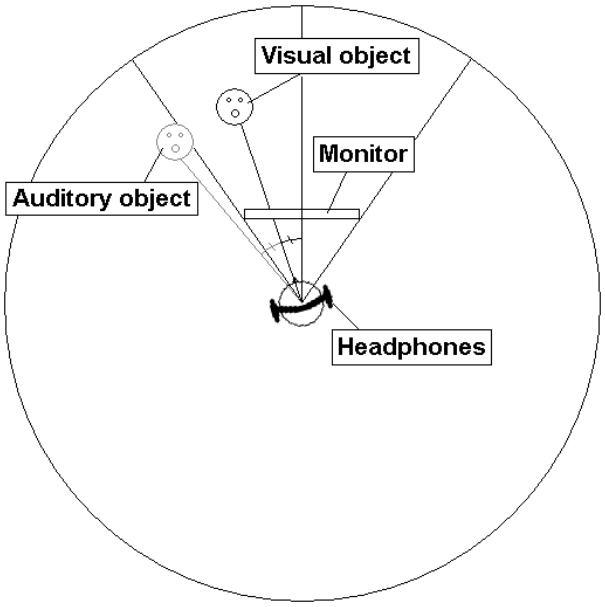

Fig 2.6 When the user's orientation changes, the auditory object stays at the same position relative to the head and therefore moves away from the visual object.

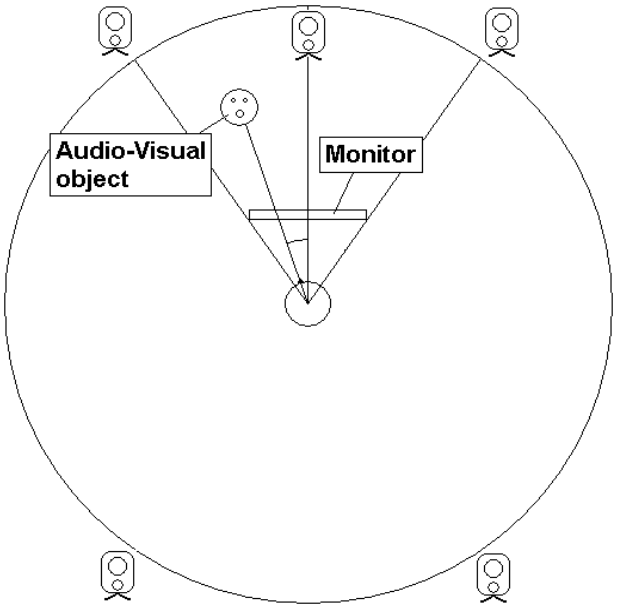

Fig 2.7 A simple two-dimensional illustration of how the user perceives the virtual environment through a surround sound speaker system and a computer monitor. When the user is not facing forward, the auditory object will still be perceived at the correct position.

If the listener is wearing headphones then the auditory environment will always stay the same relative to the user. Without head tracking, any rotation of the head would cause the auditory and visual objects to be displaced as shown in Fig 2.6 . If surround sound speakers were used as an auditory display, head orientation would not affect the position of auditory objects. The listener is however required to sit at the correct position at the centre of the surrounding speakers for the sound to be spatialized correctly. The visual display should be placed at the correct position relative to position in order to align the auditory and visual environments. This setup is illustrated in Fig 2.7. In practice, such a precise configuration is often impractical.

Misalignment can be avoided by using a head-mounted display (HMD) and headphones for visual and auditory displays. An HMD consists of two small display screens mounted inside a helmet in front of each eye [85]. Movement of the head would now have no affect on the perceived image unless a head-tracking device is used to alter this. As long as headphones are used for an auditory display, the auditory and visual environments will stay aligned. HMDs however are expensive equipment and not practical for most virtual reality applications like interactive computer games. These applications will therefore often be subject to inconsistent auditory and visual localization cues.

The Ventriloquist Effect

The human auditory system is remarkably robust in its ability to resolve spatial conflict between audio and visual inputs. This is illustrated by the illusion created by a ventriloquist that a voice comes from the mouth of his puppet. In movies, most viewers also perceive the sound of a voice to originate from the mouth of the actor even though the actual source is in the centre of the screen. According to Holman [46] professional sound engineers can notice a discrepancy of 4° between auditory and visual locations while the average layman will only notice a mismatch greater than 15°. The fusion of auditory location with the visual location is referred to as "The Ventriloquist Effect" [20]. Plasticity studies by Recanzone [70] have shown that this fusion process has after effects. Subjects were primed by presenting visual light flashes displaced from the actual audio position. When the visual cues were

removed, the perceived audio position shifted by a few degrees in the direction of the previous displacement.

Research by Driver [26] showed that the ventriloquist effect could be exploited to enhance selective listening. He presented subjects with two co-located auditory voices as illustrated in Fig 2.8. A video of the target voice was presented either co-located with the audio or separated. He showed that subjects experienced a release from masking when the visual target was separated from the auditory voices. It seems that the fusion process involved in resolving the spatial conflict, draws the auditory stream in the direction of the visual representation, resulting in a spatial separation that enhances selective attention in the same way a physically separated auditory stream would.

Though not shown by Driver's results, it seems possible that the visual target could also draw the target sound closer to a masking sound resulting in degraded speech perception. A potential experimental setup is illustrated in Fig 2.9. If the target auditory voice is presented from position (B) and the masking voice is presented from position (A), the visual target at (A) may draw attention auditory attention closer to the masking noise.

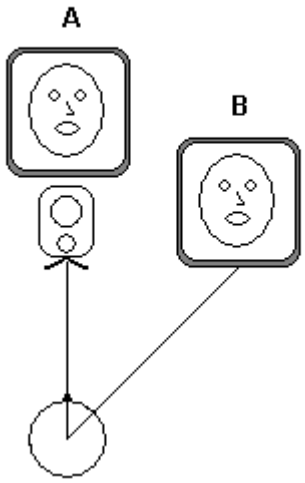

Fig 2.8: An illustration of Driver's experiment. Two simultaneous voices were presented from the speaker situated directly in front of the listener. A video stream of a talker matching the target voice was displayed on one of two television screens situated at positions (A) and (B). When the video was played on the screen at position (B), subjects were able to hear the target voice better than when it was played at position (A).

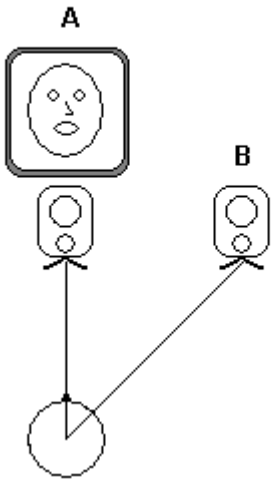

Fig 2.9: A potential experimental setup that may degrade selective attention. If the target voice was placed at position (B) and the masking voice at position (A), subjects may perform worse when the visual target is co-located with the auditory masker.

Inconsistent auditory and visual cues may influence speech perception in virtual environments. From Driver's results, speech perception may even benefit from an incorrect visual cue if it draws attention away from a masking noise. On the other hand it may be possible that the visual cue could draw attention closer to a masker, resulting in degraded speech perception. Driver's experiments made use of a second voice as a masker, which is an example of an informational masker. Masking sounds in virtual environments will however not be limited to other sources of speech. It would be of interest to see whether these experiments can be repeated in a virtual environment using an energetic masker like white noise.

2.5.2 Speech reading

The ability to read a talker's lips plays a significant role in understanding speech, especially in noisy environments [84]. Research has shown that the artificial reconstruction of lip movement can be beneficial during multimedia telephony for the hard of hearing [52]. This benefit may extend to lip animation in virtual environments. Animations need to be carefully constructed however since incongruent visual information can cause different sounds to be perceived as illustrated by the McGurk Effect [60], see below.

Lip animation
Auditory speech sounds are classified into units called phonemes. The visual counterpart for a phoneme is called a viseme [18]. A viseme represents the shape of the lips when articulating an auditory syllable. Many phonemes however have ambiguous visual representations and map to the same viseme. The Preston Blair phoneme series [10] is a popular set of visemes often used for facial animations in cartoons. In this series only 10 visemes are used to map to all possible phonemes. This can be seen in Fig 2.2.

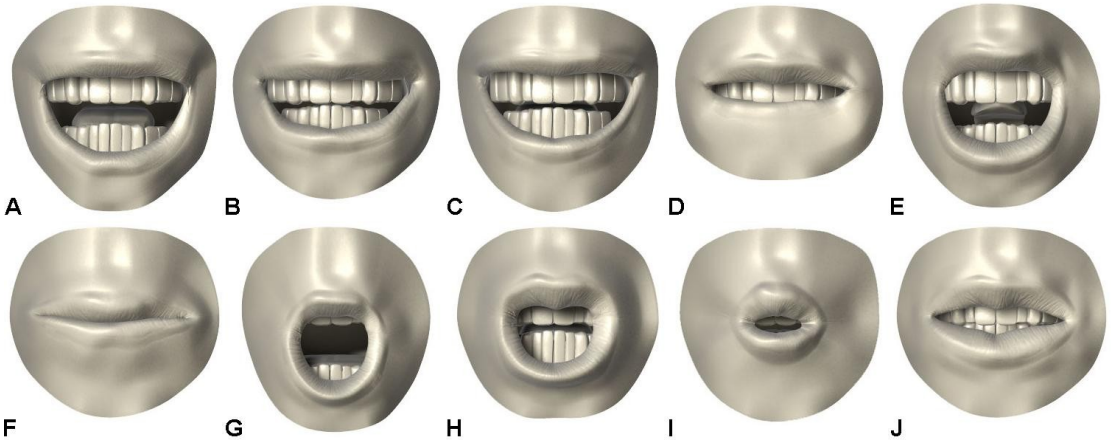

Fig 2.2: The Preston Blair phoneme series. Each visual representation (viseme) represents one or more auditory phonemes. Viseme (A) maps to phoneme A or I, (B) maps to C, D, G, K, N, R, S, TH, Y or Z, (C) maps to E, (D) to F or V, (E) to L, (F) to M, B or P, (G) to O, (H) to U, (I) to W or Q. All other phonemes map to viseme (J).

Chen *et al* [18] presents an overview of different methods of creating speech-driven facial animations and lip synchronization. Lip animations are constructed by either using a flipbook method, or by using geometry morphing. The flipbook method rapidly displays a list of consecutive visemes together with the auditory speech to create an impression of lip movement. Since there are a limited number of facial expressions, this method can result in jerky animations when no intermediate frames are drawn for the transition between different visemes. The geometry morphing method requires a 3D model of a face to be constructed. The geometry of the face can be smoothly interpolated between different facial expressions resulting in very smooth animation.

Both methods require the different visemes to be synchronized with auditory phonemes as they are spoken. Lip animations can be derived from acoustical speech input by using various computational methods. Lavagetto made use of neural networks for speech-driven facial animation in a multimedia telephone application for the hard of hearing [52]. He showed that the resulting lip animations were useful for enhancing speech perception. Much simpler methods are used for creating animations when using the flipbook method. Software tools like PAMELA [82] extract phonemes from a given text sentences and map them to visemes. The time offset for each viseme can be manually adjusted until the animation looks realistic.

The computational cost involved in creating facial animations directly from the acoustical speech data can be prohibitive for virtual environments that typically spend most processing time on graphics, physics and artificial intelligence computations. The flipbook method is more suitable for these kinds

of applications since it uses very few computational resources [18]. It has not been established however whether this kind of animation can contribute to the perception of speech.

The McGurk Effect
When creating lip animations it is important that phonemes are correctly mapped to visemes. The McGurk effect [60] illustrates how incongruent auditory and visual information can cause a different perception of the auditory stimuli. When someone hears the auditory syllable /ba/ but sees the visible syllable /ga/ (Viseme (B) followed by viseme (A) in Fig 2.2) being articulated, it is usually perceived as /da/. The perceived audio-visual syllable has the same visual representation as the presented visual syllable but differs from the presented auditory syllable. Only some combinations of auditory and visual syllables produce McGurk effects. These studies therefore typically use a limited set of stimuli that usually only consist of single syllables. It has however been shown that the McGurk effect can be obtained for normal words. If the visually and auditory presented words are picked very carefully a completely different English word can be perceived. If for example the auditory word ‘mail’ were presented together with the visual word ‘deal’, the word ‘nail’ would be perceived [24].

2.6 Speech intelligibility tests

The study of human perception requires carefully designed experiments that measure the influence of different factors on subjects’ ability to perform a perceptual task. For speech perception, subjects are required to complete speech intelligibility tests to determine their hearing performance under different conditions.

The Modified Rhyme Test (MRT) [47] is commonly used in speech intelligibility experiments. Subjects are required to identify a target word in a sentence with the form “Number (number) is (target)”. They are then presented with an ensemble of six words from which to choose the target. The MRT consists of 50 such ensembles. As a rule, words have the form consonant-vowel-consonant (CVC). For the first 25 ensembles, all six words have the same initial consonantal element while the final element is varied. For example: “bat, bad, back, bass, ban, bath”. For the last 25 ensembles the initial element is varied while the final element is the same. For example: “led, shed, red, bed, fed, wed”. The challenge in each trial is therefore to identify the correct consonant.

The Coordinate Response Measure corpus [63] is another popular set of sentences often used in speech perception studies. This corpus has a limited vocabulary with target words consisting of a call sign, a colour and a number. Sentences in the CRM corpus have the following format:

“Ready (Call sign) go to (Colour) (Number) now.”

The call sign can be ‘Arrow’, ‘Baron’, ‘Charlie’, ‘Ringo’, ‘Laker’ or ‘Tiger’. The possible colours are ‘Blue’, ‘Red’, ‘White’ or ‘Green’ while the numbers ranges from one to eight. When multiple spoken sentences are presented simultaneously, subjects are required to identify the correct colour and number combination of the sentence with the appropriate call sign. For experiments involving speech in noise, only a single sentence is spoken and the call sign can be ignored. In such experiments a single call sign will typically be used for all sentences [2].

2.7 Summary

Previous research has shown how spatialized sound can aid in speech perception through spatial unmasking. These studies have either made use of pre-convolved stereo sound files to produce virtual 3D sound sources, or were conducted in the free field using physical speakers as sound sources. The use of real-time sound spatialization capabilities of modern sound hardware for this purpose has not been investigated. Both headphone displays using real-time HRTF computations and surround sound displays using amplitude-panning techniques require further investigation.

While it has been shown that spatial discrepancy between auditory and visual sound source locations can result in a spatial release from informational maskers, this has not been established for energetic maskers. It also needs to be determined whether an incorrect visual cue can draw auditory attention towards a masking noise, resulting in degraded speech perception. It has also not been established whether the visual localization cue makes any contribution to the spatial unmasking of speech in noise.

It has been shown that speech-driven lip animation of 3D models can enhance speech perception. Deriving animations from the acoustic speech input is computationally expensive and not feasible for interactive virtual reality applications. It remains to be seen whether simple lip animations that make use of the flipbook method can contribute to speech perception.

Chapter 3
Experimental design and Methodology

Four experiments were designed to investigate the areas of interest outlined in Chapter 1. The general methodology and procedure followed for all experiments are first explained. This is followed by a detailed design for each of the four experiments.

3.1 Overview

All experiments were designed to determine the influence of different auditory, visual and spatial conditions on hearing performance. The CRM corpus described in Section 2.6 has proven to be more sensitive to intelligibility changes in very noisy environments than other speech intelligibility tests [17]. Since this research concerns the perception of speech in virtual environments under adverse hearing conditions it was decided to use this corpus for all experiments.

Four experiments were designed to investigate the following areas of interest:

- To determine whether modern sound hardware can produce a spatial release from masking in real-time. We expected spatial unmasking for both headphone and surround sound displays

- To determine the contribution of visual spatial cues to speech understanding. We expected the visual cue to enhance speech perception.

- To establish whether incongruent auditory and visual spatial information in virtual environments has any effect on speech perception. We expected increased hearing performance when the visual target shifted auditory attention away from a masking noise, but degraded hearing performance when it shifted attention in the direction of the masker.

- To determine whether rudimentary lip animations used in virtual environments contribute to speech perception. We expected matching lip animations to increase hearing performance, but unmatched lip movement to degrade speech perception.

The different auditory, visual and spatial conditions are explained later in this chapter during the description of each experiment. The task for all experiments was to determine the colour and number that was spoken in a target sentence in the midst of a competing noise masker. We believe steady state noise approximates the energetic masking inherent in the large variety of background sounds typically present in virtual environments.

3.2 Methods

A "within-subjects design" was used to compare hearing performance under different conditions. The procedure used for measuring hearing performance was the same for all experiments. During each session an adaptive method was first used to determine the subject's speech reception threshold (SRT). Experimental blocks consisting of trials presented under different conditions followed after the adaptive trials.

3.2.1 Within-subjects designs

Within-subjects designs require multiple measurements to be made on the same subjects under different experimental conditions. Between-group designs require a separate group of subjects for each condition. Within-subjects designs typically require fewer subjects than between-groups designs. These designs also reduce the error variance since individual differences between different conditions are taken into account [59].

Within-subjects designs however have some drawbacks that are not present in between-group designs. Depending on the type of experiment it is possible that exposure to one condition could influence the subject's response to a different condition. Learning effects could also cause subjects to perform better under conditions presented later in the experimental session than those presented at the beginning. In the beginning, the subject may be unfamiliar with the task and may take a while to adapt to the experimental situation. These problems can be controlled by repeating different experimental conditions and randomizing the order in which they are presented [59].

3.2.2 The transformed up-down adaptive method

The speech reception threshold (SRT) for intelligibility tests refers to the minimum target-to-noise ratio (TNR) at which subjects can reliably perform the task. Hearing performance is not linear with respect to the TNR of the stimulus. Fig 3.1 illustrates the relation between the stimulus level and the percentage correct responses. This is referred to as the *psychometric function*.

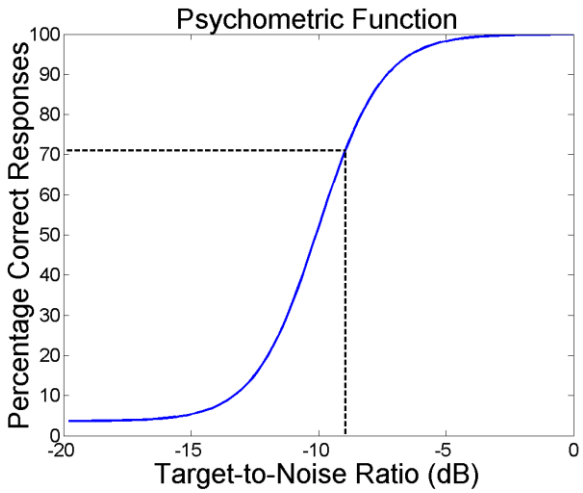

Fig 3.1 The psychometric function describes the relationship between the physical intensity of the stimulus and an observer's ability to detect or respond correctly to it. The transformed up-down procedure targets the 71% correct response threshold on the psychometric function.

Adaptive procedures are often used in psychoacoustic experiments to determine the SRT. The transformed up-down procedure [53] and the maximum likelihood estimation (MLE) procedure [67] are two of the most popular adaptive methods. One variant of the transformed-up-down method, the two-down/one-up method targets the 71% correct response threshold. In this procedure the level of difficulty is increased every time the subject is able to give two correct responses sequentially. A single incorrect response leads to a decrease in difficulty. Another variant of the same procedure, the three-down/one-up method estimates the 79% correct response threshold. The MLE procedure can target any desired threshold and is a much faster method of finding the threshold estimate. However the variance in threshold estimates for this method is significantly higher than the variance produced by the transformed-up-down methods [4]. For this reason we made use of a transformed-up-down method. We favoured the two-down/one-up method targeting the 71% threshold over the alternative method which targeted the 79% threshold. This would leave more room to show an increase or decrease in performance in subsequent trials using a fixed target-to-noise ratio.

The two-down/one-up method starts with equal target and masking noise levels yielding a target-to-noise ratio (TNR) of 0 dB. This represents the amount by which the target signal is attenuated. For normal hearing subjects it is very easy to achieve a 100% correct score in this condition. The task is then progressively made more difficult by lowering the volume of the target stimulus whenever the subject scores two correct answers in a row. As soon as the subject gives a single incorrect response the level is adjusted to make it easier again. A reversal happens when the subject either scores two consecutive trials correct after an incorrect response, or if an incorrect response directly follows two or more correct responses. The process is stopped when the number of reversals reaches a predetermined threshold.

Care should be taken when adjusting the target level. If the amount by which the volume is adjusted is too small, it will take a long time to find the final SRT. If the value were too big the final SRT would not be optimal. The step size can be adjusted after a pre-determined number of reversals. The choice of the step sizes depends largely on the prior knowledge of the experimenter. Incorrectly chosen step sizes will not change the result but may be less efficient in determining the SRT [53]. The following values was determined during pilot studies and used for all subsequent experiments.

- Until 1st reversal, adjust the volume by 5dB.
- Until 3rd reversal, adjust the volume by 3dB.
- Until 7th reversal, adjust the volume by 1dB.
- Until 13th reversal, adjust the volume by 0.5dB.

After the 13th reversal, the current volume is used as the SRT. Once the SRT has been determined for the subject, all experimental trials can be presented at the measured TNR for different conditions. The 71% correct response threshold leaves enough room to show an increase or decrease in performance when the experimental condition is changed.

3.2.3 Subject selection

The number of subjects necessary to show significant differences between experimental conditions depends on the variance between subjects. If this variance was known, one could estimate the required number of subjects. Unfortunately this information is usually not available and experimenters typically rely on similar experiments previously reported in the literature to determine the variance between subjects [59]. Speech perception studies usually make use of only a small sample size, often between four and ten subjects [17, 37, 79]. To justify the small sample size, variability between subjects needs to be small. To minimize the variance, subjects' hearing ability is often first tested to confirm that they have normal hearing ability. The normal hearing threshold is 20 dB HL in the 0.25–8 kHz frequency range [77]. This control however is often omitted and many researchers are content with subjects reporting whether they have normal hearing ability [27, 32, 40]. In such cases the experimental data needs to be analysed to ensure that there is no gross difference in subjects' ability to perform the task. This analysis will be shown in Section 4.1. In our experiments adjusting the signal-to-noise ratio relative to each subject's 71% correct hearing threshold further minimized variability, resulting in only a small between-subject standard deviation.

Eleven paid volunteers were recruited as test subjects for this research. All subjects were between the ages 20 and 30, had self-reported normal hearing and normal or corrected-to-normal vision. Subjects were not informed of the goal of the experiments. Four subjects were used in each experiment and some participated in more than one experiment.

3.2.4 Equipment

Hearing experiments are usually conducted in a sound proof room. When speech is presented in the free field, an anechoic chamber is sometimes used to minimize reverberation, which has been shown to impair speech perception [67]. An isolated room was refurbished for the purposes of this research. The walls of this room were covered with medium density fibreglass padding to minimize reflections and to help isolate the room from outside noise.

The experimental software was run on a desktop-based system with a 3000 MHz Intel Pentium processor, 512 MB RAM and a 19" monitor. The system was also equipped with a GeForce FX5900 graphics card with 128 MB onboard RAM and a Creative Labs Sound Blaster Audigy 2 sound card. A Bose 5.1 speaker system and a Rotell amplifier were used as the surround sound auditory display. For the headphone display a pair of Sennheizer HD 580 circum-aural headphones was used. A Virtual Research V6 Head-mounted display was used in all but the first experiment. This HMD supports a resolution of 640x480 and can display a 60° field-of-view. The virtual environment application was written in C++ using the Microsoft DirectX API [57].

3.2.5 Procedure

When gathering experimental data it is important that the trials presented under different conditions are equally difficult. If some words presented in one condition were easier to identify than words presented in another condition, this would create a misleading bias towards one condition. In the CRM corpus some colours and numbers are easier to recognize than others [17]. During data gathering, the same sentences were presented an equal amount of times under all experimental conditions. This ensured that all an equal number of easy and difficult sentences were presented for all conditions, removing the bias towards any one condition.

All subjects participated in five experimental sessions on five consecutive days. To minimize the effect of fatigue, all sessions were kept under one hour and subjects were given a short break between blocks of trials. For the first 3 days subjects had to complete 3 adaptive learning blocks to find a adequate TNR for each subject. Each of these blocks lasted for about 5-6 minutes. During the adaptive trials only audio was presented and the target and masker objects were invisible. The auditory masker was always presented at 0° while the target was presented at 15° to the right. An experimental block of up to 20 minutes followed after this. The average TNR measured in the 3 adaptive blocks was used as the TNR for the experimental block. On the last two days no adaptive blocks were conducted, but two experimental blocks, using the average TNR measured on the third day. During pilot testing it was observed that subjects tend to perform better towards the end of a block than at the very beginning. To account for any learning effects within a block, a few warm-up trials were first presented. These trials were not considered for data analysis. The number of conditions determined the number of trials that could be presented in subsequent experimental blocks in the time available. For experiments with 4 different conditions, 56 usable trials were gathered for each condition during a block. Only 28 trials could be gathered per condition for experiments with 6 different conditions. The last two sessions contained two experimental blocks and no adaptive blocks. A total of seven experimental blocks were conducted over the five days. For experiments with 4 different conditions this resulted in 392 trials per experimental condition. Experiments with 6 different conditions only had 196 trials per condition. This excludes any adaptive trials since the number of trials presented during each of these blocks naturally varies. To account for learning effects, the first two experimental blocks were not considered for data analysis.

3.2.6 Data Analysis

A Repeated Measures Analysis of Variance (ANOVA) was used to analyse the experimental data of each experiment. This kind of analysis is typically used for within-subject experimental designs [83] and is commonly used in speech perception studies [34, 37, 40]. Where significant differences between experimental conditions were found, a post-hoc Newman-Keuls test was used to determine which conditions contributed to the effect.

3.3 Experiment 1 – Spatial release from masking with modern sound hardware

Our first hypothesis was that modern commercial sound hardware is capable of producing a spatial release from masking for both headphone and surround sound displays. This has not been verified for either real-time HRTF processing using headphone displays or amplitude panning techniques using surround sound speakers. The purpose of this experiment was to determine whether our hypothesis held true. The Microsoft DirectX API [57] was used to produce spatialized sound with a consumer 3D sound card for both auditory displays. Previous research has demonstrated that sound spatialization achieved by pre-computing stereo files with HRTFs measured on a KEMAR dummy head can result in spatial unmasking [30]. This would serve as a base condition to compare with the performance of the other two techniques.

3.3.1 Materials

Although the CRM corpus is publicly available for research, all speakers used for the recordings had American accents. Since some subjects might find it difficult to recognize a foreign accent, especially in noisy conditions, it was decided to create a CRM corpus using a South African speaker. A native South African English-speaking female drama student was used as voice talent. Professional sound engineers were employed to record the target stimuli. The length of sentences ranged between 2.29 and 2.73 seconds with an average of 2.50 seconds. All sentences were recorded at 48 kHz. The sound files were first edited to make sure every file immediately started with the first word without any delay. The sound files were also trimmed at the end after the last word has been spoken.

Since the call sign was not important for our experiments, only call sign “Baron” was used. The number 7 was not used in any trials since it is the only two-syllable number and would be easier to recognize. This left four colours and seven numbers in the vocabulary. With 28 possible permutations of colour and number, the chance of a subject guessing both the correct colour and number is about 3.6%. In some cases subjects may be able to recognize only one of the target words. This would clearly be better than recognizing nothing at all. Since this information would be lost when using absolute scoring, it was decided to award a point for answering the correct colour and another point for the correct number. When scoring in this way the level of chance scoring correctly is increased to about 19.6%.

Since distracting sounds in virtual environments are not limited to speech sources, it was decided not to use speech spectrum or speech shaped noise for these experiments as is common in speech perception studies. White noise of the same length as the longest speech stimulus was generated for the masking

stimuli. Ten different masking files were created in this way and were randomly presented during experiments.

The root mean square (RMS) energy of a sound file refers to the square root of the mean of the squares of the all the digitized sound sample values. In order to make sure the target-to-noise-ratio was calculated correctly, the RMS energy of the masker should be equal to that of the target sound. This was done by first scaling all target files to have data values in the (-1, 1) range. The minimum RMS energy for these files was then calculated and all files were scaled to have the same RMS energy. The masker stimuli were then scaled to have the same RMS as the normalized target stimulus. All stimuli were ramped with a cosine-squared window to remove any clicking at the beginning and end of sentences when presented. The MS DirectX API made use of these normalized single channel files to spatialize the sounds.

These sound files were convolved with the impulse responses measured on a KEMAR dummy head to create a separate set of stereo sound files. MIT Media Lab measured the impulse responses used in this study [36]. Spatialized sounds were produced by convolving the signals with KEMAR HRTFs for angles 0° and 15° in the horizontal plane. All sounds were created with a zero elevation angle. No further processing was performed on the stereo sound files during presentation.

Visual stimuli of the sound producing objects were not important for this experiment. A virtual room was modelled by mapping pictures of a carpet, wooden wall panels and a ceiling onto the surfaces of a shoebox shaped object. No other objects were visible while the trial was being presented as seen in Fig 3.2. The virtual environment was presented on a computer monitor.

Fig 3.2 The virtual room. No other objects were visible while trials were presented in Experiment 1.

3.3.2 Conditions

The three auditory conditions matched the different types of sound spatialization techniques. In one condition the sounds were spatialized using real-time techniques and using stereo headphones as an auditory display. In the second condition, amplitude-panning techniques were used to spatialize sounds

over a surround sound speaker display. The last condition made use of headphones and pre-computed stereo spatialized sounds.

Two spatial conditions were used for the auditory stimuli. In one condition both the noise masker and target sentence would be presented at 0° (straight ahead of the listener). In the second condition the masker was still located at 0°, but the target sentence was presented at 15° to the right.

A summary of the conditions is provided in Table 3.1.

Condition	Visual Target	Auditory Target	Audio-visual Masker	Auditory display
1.	Invisible	0°	0°	KEMAR / Headphones
2.	Invisible	15°	0°	KEMAR / Headphones
3.	Invisible	0°	0°	DirectX / Headphones
4.	Invisible	15°	0°	DirectX / Headphones
5.	Invisible	0°	0°	DirectX / Surround
6.	Invisible	15°	0°	DirectX / Surround

Table 3.1: Spatial, visual and auditory conditions for Experiment 1. Three different auditory displays were compared in this experiment. Sound producing objects were never visible. The masker was always presented from the front while the target sentence was presented either from the front at 0° or 15° to the right of the masker.

3.3.3 Test Environment

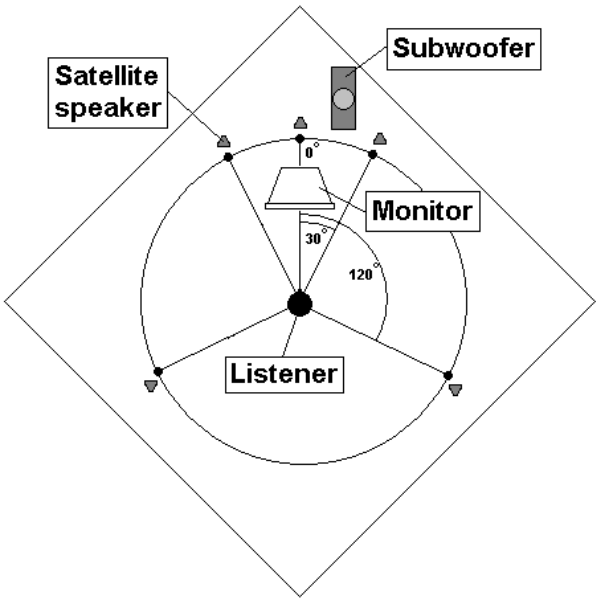

Fig 3.3: The surround sound speaker configuration. Five satellite speakers were placed in a circle around the listener. The listener was oriented to look at the front centre speaker, which was located behind but slightly elevated above a computer monitor. The front left and right speakers were placed at 30° to either side of this speaker and the rear speakers at 120° to each side. The subwoofer was placed on the floor in between the centre and front left speakers.

Experiments were conducted in a room with dimensions of 6m x 3m. Only one half of the room was used in order to create a symmetric speaker placement around the listener. The room was equipped

with a 5.1 surround speaker setup consisting of five satellite speakers positioned on speaker stands around the listener as shown in Fig 3.3. The centre speaker was placed in one corner of the room. The front speakers were separated by 30° either side of the centre speaker. The rear speakers were displaced with 120° either side. All five speakers were placed at the same height, which was slightly higher than the top of the computer monitor. The subwoofer was placed on the floor between the centre and front right speakers. A chair was fixed at the exact centre of the five surrounding speakers. This configuration agrees with the recommended room layout specified by Dolby Laboratories [25].

The experimental software was run on a desktop computer also located in the room. An optical mouse was used as an input device to allow subjects to choose the colour and number combination they heard.

3.3.4 Procedure

The task in every trial was to identify the correct colour and number combination spoken in the target sentence in the midst of a competing noise masker. Subjects were instructed to sit upright in the chair and to stare at the centre of the screen while listening to the auditory stimuli. At the end of each trial subjects were prompted to select the correct colour and number combination. The user interface can be seen in Fig 3.4. Once the subject has responded, the user interface was removed and the next trial was presented.

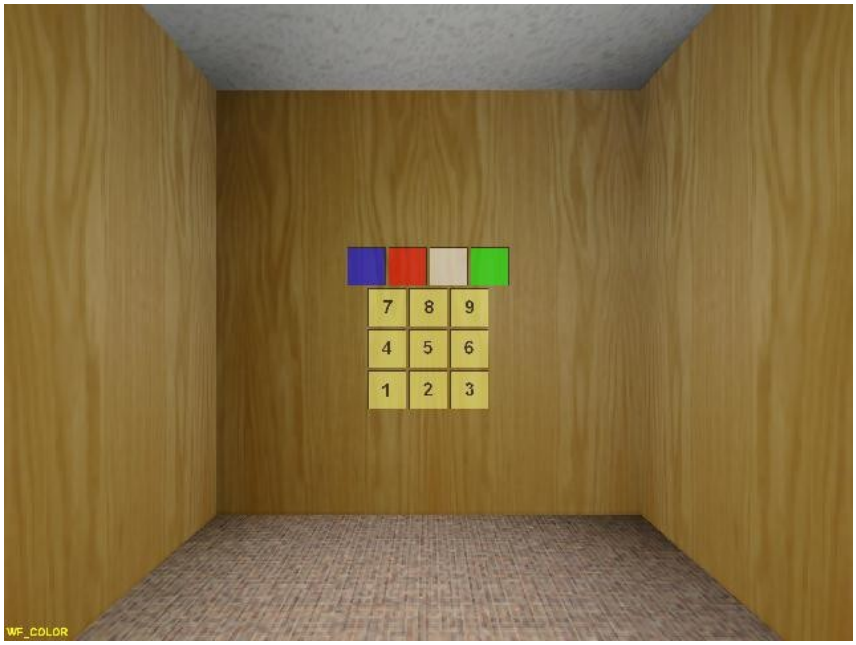

Fig 3.4: The user interface. Subjects were prompted to choose the correct colour and number combination at the end of each trial.

The adaptive method was used to determine the SRT. Pre-computed spatialized audio and headphones was used for the auditory display for all adaptive trials. The TNR at the measured threshold was used for all other auditory and spatial conditions. It would be desirable to randomize the different conditions being presented [59]. However, since it would be impractical to remove the stereo headphones every time a surround sound trial was presented, each experimental block were divided into three sub-blocks in which the different audio conditions were presented separately. The order in which the sub-blocks were presented over the 5 days of the experiment was randomized. During sub-blocks the different spatial conditions were randomly presented. 196 trials were gathered per condition for this experiment.

3.4 Experiment 2 – The influence of a visual localization cue

Our second hypothesis is that a visual cue aiding auditory localization will enhance speech perception in the presence of masking noise. This experiment investigates this hypothesis by presenting auditory stimuli with and without a visual counterpart at different spatial locations. The lips of the talker were not animated during this experiment to ensure that the visual cue conveyed only directional information. In a previous study Reisberg showed that visual localization cues are beneficial in the presence of a second talker, which acted as an informational masker [73]. We attempted to reproduce this result using an energetic noise masker.

3.4.1 Materials

The same auditory stimuli from the first experiment were used for this experiment. For visual stimuli 3D models were used to represent the sound producing objects. A television screen that displayed a snowy picture, as is common when there is bad reception, represented the masker object. A face representing the target was presented on a separate television in the virtual environment. The snowy television was animated by randomly switching between different noisy images at a constant frame rate. The face however was static for this experiment.

Fig 3.5: A screen shot of the virtual environment. Two television screens represented the sound producing objects. The masking noise was associated with the snowy picture while the speech sentence was associated with the face. Note that the actual visual environment observed by the subjects differed slightly from the image above. A head mounted display (HMD) was used to project a stereoscopic display, which had a slightly smaller field-of-view.

3.4.2 Conditions

Two visual conditions were presented. In one condition the subject could see the target and masker objects in the correct spatial positions. This can be seen in Fig 3.5 and Fig 3.6. In the other condition none of the visual objects were invisible as illustrated in Fig 3.2.

The same spatial conditions from the first experiment were used. The target speech sentence was either presented at 0° or 15°. The masker was always presented directly in front of the listener at 0°.

Headphones and pre-computed stereo spatialized sounds were used for the auditory display under all conditions. The different conditions are summarized in Table 3.2.

Condition	Visual Target	Auditory Target	Audio-visual Masker	Auditory display
1.	Invisible	0°	0°	KEMAR / Headphones
2.	Invisible	15°	0°	KEMAR / Headphones
3	0°	0°	0°	KEMAR / Headphones
4.	15°	15°	0°	KEMAR / Headphones

Table 3.2: Spatial, visual and auditory conditions for Experiment 2. The KEMAR auditory display was used in all trials. The masker was again always presented from the front. The target and masker objects could either be visible or invisible.

Fig 3.6: The visible, co-located condition. In this condition the target object could be seen but it obscured the masker object, which was always presented in the centre position.

3.4.3 Test Environment

For the purposes of this experiment it was important that the visual objects were displayed at the same spatial position as the auditory object. It was therefore decided to use a head mounted display and headphones for an auditory display. In Section 2.5.1 we saw that this configuration ensures that the auditory and visual environments remain aligned regardless of head movement.

The same user interface used in the first experiment was presented. The slightly transparent input console was superimposed on the display area and subjects did not have to remove the HMD when providing responses. This is can be seen in Fig 3.4. They did however have to keep their hands on the mouse during the experiment.

3.4.4 Procedure

The procedure for the second experiment was similar to the first. Because there were only four different conditions (the combination of the two visual and two spatial conditions), more trials could fit into each experimental block. Subjects were allowed to rest for a few minutes in between blocks.

Since surround speakers were no longer used, subjects did not have to sit in any specific position. They were instructed to sit in any way comfortable to them. The use of a single auditory display also allowed the different visual and spatial conditions to be randomly presented during each block. 392 trials were gathered per condition over the course of 5 days.

3.53.5 Experiment 3 – The effect of incongruent localization cues

Our third hypothesis states that incongruent auditory and visual spatial information will contribute to improved speech understanding when the visual target shifts the localization of the audio-visual event away from a masking noise, but will degrade hearing performance when it shifts localization in the direction of the masker. The purpose of the third experiment was therefore to determine whether incorrect visual cues for sound source location, could have an influence on the spatial unmasking of speech. It was expected that by moving only the visual representation of the target further from the auditory noise masker, speech understanding could be improved. Conversely it was expected that moving the visual target closer to the noise masker would degrade speech understanding.

3.5.1 Materials

The same auditory stimuli from the second experiment were used. Where the visual cues in the previous experiment contributed only directional information, the face of the talker was now animated to increase *visual capture*. This refers to the phenomena where the apparent direction of an auditory stimulus is dependant on a corresponding visual stimulus [4]. It was thought that lip movement matching the auditory speech would increase visual capture and have a greater probability of drawing auditory attention towards the visual image.

Fig 3.7. Target speech animation frames.

Illustrations of the Preston-Blair phoneme series [10] were used for animating the face of the character. An animation file containing the relative timing offsets of different frames was created with the help of a lip synchronization utility called Pamela [82]. This tool can determine the correct phonemes to use for any given English sentence. While Pamela cannot create the correct timing offsets for each frame from the speech file, it does allow the user to adjust the timing offsets until the animation looks correct. The animation frames are illustrated in Fig 3.7.

3.5.2 Conditions

Two different auditory spatial conditions were again presented in the same way as in the previous experiment. The target visual conditions were either correctly positioned or at the opposite spatial condition from the auditory position resulting in four different conditions for this experiment as shown in Table 3.3.

Condition	Visual Target	Auditory Target	Audio-visual Masker	Auditory display
1.	0°	0°	0°	KEMAR / Headphones
2.	0°	15°	0°	KEMAR / Headphones
3	15°	0°	0°	KEMAR / Headphones
4.	15°	15°	0°	KEMAR / Headphones

Table 3.3: Spatial, visual and auditory conditions for experiment 3. The KEMAR auditory display was used in all trials. The masker was always presented from the front. The target and masker objects were always visible and animated but the visual representations were presented either at the correct locations or displaced by 15°.

If subjects performed better in condition 3 than in condition 1, we could conclude that the visual target drew auditory attention away from the masker, resulting in spatial unmasking. If subjects performed worse in condition 2 than in condition 4, the visual target drew auditory attention in the direction of the masker resulting in a smaller release from masking.

3.5.3 Test Environment

The same test environment from the second experiment was used.

3.5.4 Procedure

The same procedure from the previous experiment was followed.

3.63.6 Experiment 4 – The influence of lip animation

Our final hypothesis is that rudimentary lip animations matching the auditory speech improve hearing performance, but unmatched lip movement is detrimental to speech perception We investigated this hypothesis by presenting correct, incorrect and no lip animation at different spatial positions in the virtual environment.

3.6.1 Materials

The same auditory and visual stimuli from the third experiment were used during this experiment. For incorrectly animated conditions, the facial animations of a different target sentence were randomly selected. For example, if the auditory stimulus was the sentence "Ready Baron go to blue, one now", the visual animation for the sentence "Ready Baron go to green, five now" could be presented.

3.6.2 Conditions

Pre-computed stereo spatialized sounds with headphones were again used as an auditory display. The same spatial conditions from previous experiments were used. Three visual conditions were presented: Animated, non-animated and incorrectly animated. The different conditions are shown in Table 3.4.

Condition	Audio-visual Target	Audio-visual Masker	Animation	Auditory display
1.	0°	0°	Animated	KEMAR / Headphones
2.	0°	0°	Non-animated	KEMAR / Headphones
3	0°	0°	Incorrectly animated	KEMAR / Headphones
4.	15°	0°	Animated	KEMAR / Headphones
5.	15°	0°	Non-animated	KEMAR / Headphones
6.	15°	0°	Incorrectly animated	KEMAR / Headphones

Table 3.4: Spatial, visual and auditory conditions for experiment 4. The KEMAR auditory display was used in all trials. The masker was always presented from the front. The visual target was either correctly animated, non-animated or incorrectly animated. The audio-visual target was presented either in front of the masker or to the right of the masker.

The reason for using two spatial positions in this experiment was to investigate the influence of lip-animation at different levels of hearing difficulty. Because of spatial unmasking, the target sentence would be easier to understand when presented at 15° than in the co-located condition.

3.6.3 Test Environment

The same test environment as the previous experiment was used.

3.6.4 Procedure

The same procedure was followed as in the previous experiments. The visual conditions were randomly presented so subjects were never sure whether the animation was correct or incorrect. Since six conditions had to be tested in the session less data could be gathered than in the previous two experiments. 196 trials were gathered per condition over the course of 5 days.

3.7 Summary

The methods and design employed for this research is not novel in any way. Many previous studies have made use of the CRM corpus for investigating speech perception. Within-subject designs with a small number of subjects are commonly found. The methods of data analysis are also typical of perceptual studies. The experimental design and methodology discussed in this chapter is therefore consistent with previous research found in the literature. However, these experiments are unique in examining how different rendering schemes affect spatial unmasking in a virtual environment and whether audio-visual interactions can both improve and degrade speech understanding in a virtual environment.

Chapter 4
Data Analysis and Results

In this chapter we first provide an analysis of the data gathered during the adaptive learning trials. We show that the eleven subjects all had comparable ability to perform the experimental task and that no learning effects were observed after the second day of experiments. We then evaluate the speech stimuli to show the relative difficulty of recognizing different colours and numbers presented in the CRM corpus. Our native CRM recordings yielded very similar results to those observed in other studies that make use of this corpus. Finally we present a detailed analysis of the results found in each experiment.

4.1 Subject variance

An adaptive method was used to determine the hearing performance of all subjects before every experiment. Data were gathered in 9 experimental blocks over the first three days of each experiment. Fig 4.1 shows the variance between subjects as well as how their performance changed over this time. It is clear that by the third day (blocks 7, 8 and 9) subjects were comfortable with the task and all subjects had comparable performance.

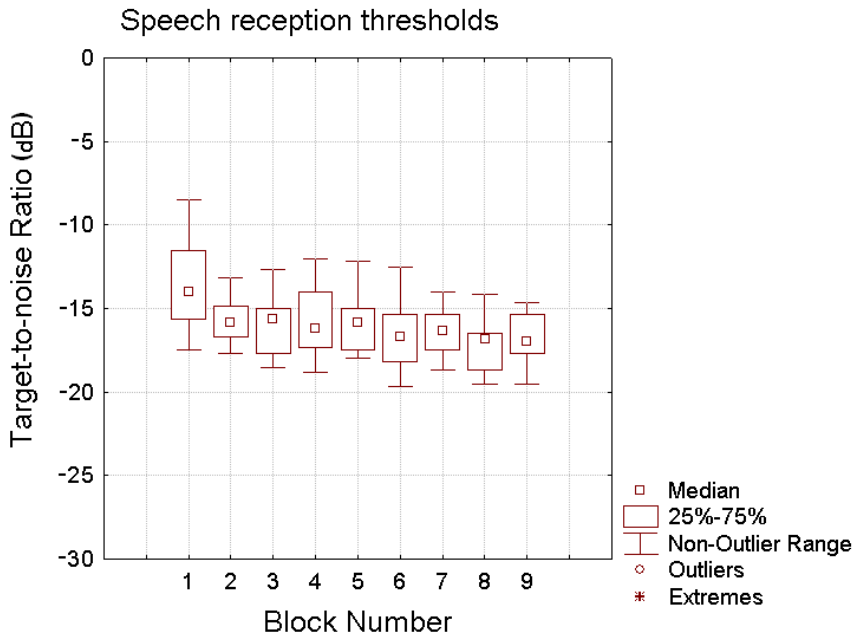

Fig 4.1: Speech reception thresholds (SRT) of all subjects as measured over the first three days of experimental trials. Lower target-to-noise ratios indicate better speech reception thresholds. It is clear that subjects very quickly adapted to the experimental method used. There was no significant improvement after the first block. The variability did seem to

decrease over time. There were no outliers or extremes in the data set and no gross differences were found between subjects' ability to perform the task.

4.2 Evaluation of the speech stimuli

As mentioned in Chapter 3, independent recordings of CRM sentences spoken by a South African native English speaking person were used as auditory stimuli for all experiments. We evaluated our speech stimuli in order to verify that they had the same characteristics observed in other studies. Brungart has documented the recognition performance for the different colours and numbers in his evaluation of the CRM corpus [17]. A similar comparison can be seen in Fig 4.2 and Fig 4.3.

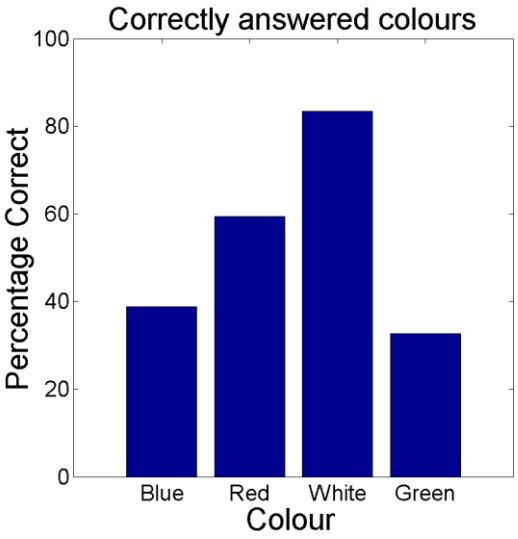

Fig 4.2: Relative recognition performance for different colours in the CRM corpus. Subjects found the colour 'White' the easiest to identify and had the greatest difficulty with the colour 'Green.'

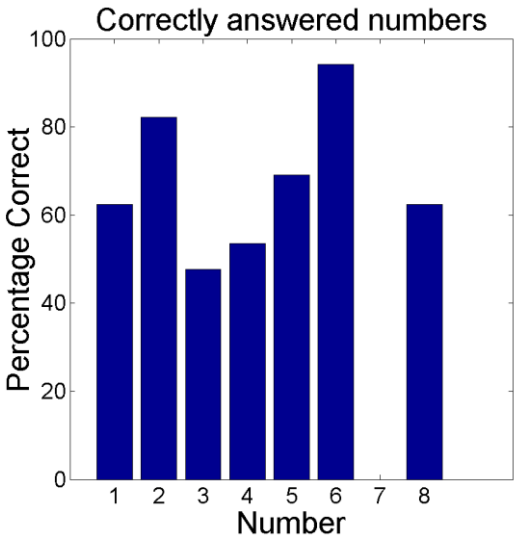

Fig 4.3: Relative recognition performance for different numbers in the CRM corpus. The numbers 'Two' and 'Six' was the easiest to identify while the number 'Three' was the most difficult.

Both the colour and number recognition performance are similar to Brungart's findings. The easiest numbers to recognize during his study was also 'Six', 'Five' and 'Two'. He also found the colours 'Red' and 'White' to be the easiest. He did however find the colour 'Blue' and the number 'Eight' were the most difficult to recognize which is contrary to the current results. Overall the percentage correct number identifications were consistently higher than the correct colour identifications, which agree with Brungart's findings.

4.3 Experiment 1 – Spatial release from masking with modern sound hardware

The purpose of this experiment was to determine whether modern sound hardware is capable of producing a spatial release from masking. The experimental design for this experiment was discussed in Section 3.3. The results of subject performance under the different experimental conditions can be seen in Fig 4.4.

4.3.1 Results

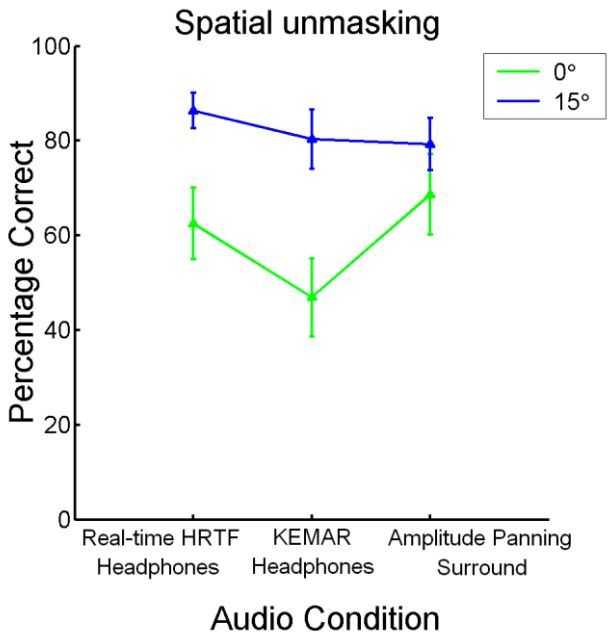

Fig 4.4: Subject performance using three different auditory displays. The dark line represents the spatially separated condition where the target sound was located at 15º to the right. The lighter line represents the co-located condition where both the target and masker was located at 0º. A spatial release from masking was observed for all three auditory displays. The headphone display using HRTFs measured on KEMAR resulted in the worst performance.

We measured the amount of spatial unmasking achieved for three different auditory displays during this experiment. The first method made use of real-time HRTF processing using stereo headphones. The second display used pre-computed spatialized stereo sound files which were created using HRTFs measured on KEMAR. The last method made use of amplitude panning to spatialize sound over a surround sound speaker system. The target stimuli were presented from two different spatial positions for while the masker was always positioned in the centre. A repeated measures ANOVA across the two spatial conditions revealed a statistically significant difference [$F(1,3) = 66.183$, $p < 0.01$]. A Neuman-Keuls post-hoc comparison revealed that the difference was significant for all auditory conditions indicating that all auditory displays resulted in a spatial release from masking. See Table A.1.

The difference between the auditory conditions was also found to be statistically significant [$F(2, 6) = 26.23$, $p < 0.01$]. Post-hoc comparisons revealed that for the co-located condition the KEMAR display differed significantly from both other auditory displays. No difference was found between the DirectX and Surround displays in the co-located condition. In the separated condition no significant differences were found between the different auditory displays. See Table A.2.

4.3.2 Analysis

For the experiment involving different audio conditions we expected modern sound hardware to produce a spatial release from masking comparable to that observed when using pre-computed methods of sound spatialization. Surprisingly, real-time sound spatialization techniques resulted in better hearing performance in the co-located condition when compared with the pre-computed method using KEMAR HRTFs, which has been a standard in audio research.

The surround sound display made use of amplitude panning techniques to present sound in 3D. It may be that presenting a sound in the free field in this way has advantages for discriminating between the target and masker stimuli. Hawley *et al* has shown that free field presentation of an actual sound source has no benefit for speech perception over a virtual auditory display using headphones [42]. When using amplitude panning however, sound is presented from more than one speaker simultaneously in the free field. The gain at each speaker will depend on the position of the sound source. All three front speakers are used to present a sound at 0º in the horizontal plane. Although the perceived direction is still 0º, where the centre speaker is located, it may be that the additional auditory information presented from the front left and front right speaker could help the listener to distinguish the target. This may explain the increase in hearing performance when comparing the co-located conditions of the KEMAR and surround sound displays.

The headphone display using real-time HRTF also resulted in better performance than the KEMAR HRTFs in the co-located condition. The HRTFs used by consumer sound hardware is very different from those used when pre-computing a spatialized sound file. The different filter characteristics could have contributed to different levels of masking in the co-located condition. Unfortunately hardware manufacturers do not specify the exact processing performed by the card. The results however show that modern consumer sound hardware can alleviate the energetic masking of speech.

All auditory displays resulted in spatial unmasking, confirming our hypothesis. Since the psychometric function is not linear, one should be careful when making direct comparisons between the amounts of masking achieved in each condition. Ceiling effects may have prohibited the Real-time HRTF and amplitude panning techniques from showing a greater release from masking.

4.4 Experiment 2 – The influence of a visual localization cue

The purpose of the second experiment was to determine whether a visual cue for sound source location contributed to a spatial release from masking. Two visual conditions were presented. In the one condition sound producing objects were visible and in the other condition they were invisible. The auditory and visual environments were aligned for this experiment and both the auditory and visual stimuli were presented at the same location in the virtual environment.

4.4.1 Results

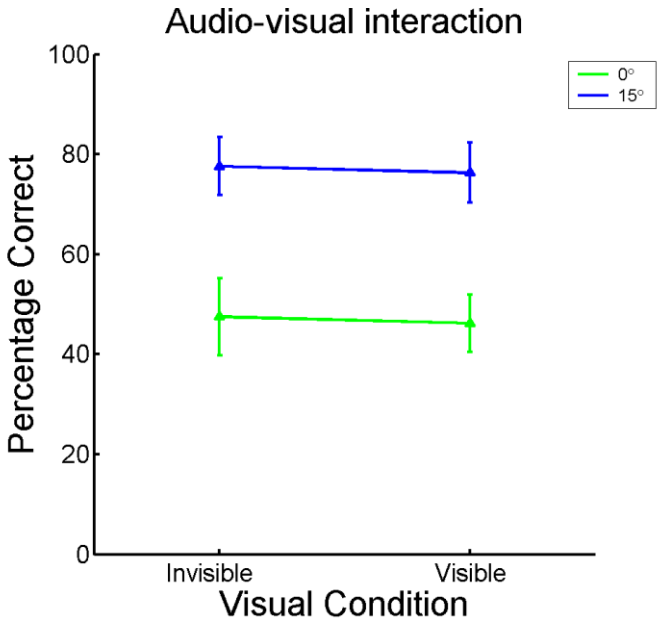

Fig 4.5: Subject performance under different visual and spatial conditions. The dark line represents the spatially separated condition where the target sound was located at 15º to the right. The lighter line represents the co-located condition where both the target and masker was located at 0º. A spatial release from masking was observed for both visual conditions. The addition of a visual cue for sound source location did not have any significant effect on subject performance.

In Fig 4.5 subject performance for the visible visual condition can be seen on the right and for the invisible condition on the left. From this figure one can see a notable difference in performance between the co-located and separated condition. All subjects experienced a spatial release from masking and performed significantly better in the separated condition. A repeated measures analysis of variance showed a statistically significant difference between the two spatial conditions [$F(1, 3) = 87.36$, $p < 0.01$]. Further post-hoc analysis shows that the difference was significant for both the visible and invisible conditions. See Table A.3. No difference were found between the visible and invisible conditions [$F(1, 3) = 0.16$, $p = 0.713$].

4.4.2 Analysis

For the second experiment we expected visual cues to contribute to spatial unmasking. No significant difference was however found between the visible and invisible conditions. This result differs from Reisberg's findings in a similar study [73]. He found a small improvement when subjects could accurately localize the sound sources.

Since the lips of the talker did not move, it is possible that this resulted in a lack of visual capture. Adding lip animation in this experiment would however have added additional visual information that would not be present in the invisible condition. Subjects may not have associated the motionless face with the speaking voice and therefore the visible condition did not show any improvement over the invisible condition. It is possible than one might find it easier to believe that a sound is coming from a physical speaker, as in Reisberg's study, than from a motionless face.

The lack of a significant difference may also be attributed to the nature of the speech stimuli. The sentences were around 2.5 seconds in length and always started with the primer "Ready Baron go to…" It may very well be that the time in which the primer was spoken was enough to allow subjects to localize the sound in the invisible condition. If shorter sentences were use or if only the colour and number was called out, subjects may have a harder time to localize the sound and may perform worse than in the visible condition. However, most virtual environments would make use of even longer sentences when conveying dialog.

The nature of the masking stimuli could also have had an influence on the results. The way in which informational maskers and energetic noise maskers achieve their goals differs substantially. Energetic maskers with similar frequency content as the target cause overlapping excitation patterns in the auditory nerves [28]. Some auditory information is essentially lost during this process. Informational masking occurs as a result of higher-level processes. All the auditory information is still present but the auditory system has trouble making sense of it all. Any hearing improvement as a result of visual cues would have to be achieved through a process of audio-visual integration. If some auditory information has been lost, this process may be less effective. Where Reisberg has observed a small benefit of visual localization information to speech perception when using an informational masker. Previous studies have shown that lip movement, another visual cue, is much more beneficial in the presence of a speech masker than a noise masker [48]. Our results seem to indicate that visual localization cues are also less effective in overcoming energetic masking. This will however have to be confirmed in a follow-up study that compares the effectiveness of different maskers. See Chapter 5 – Future work.

We have again observed a significant difference between the separated and co-located conditions. This agrees with findings in the literature and indicates that subjects did experience a spatial release from masking as expected. These results show that the inability to see a talker does not significantly affect a user's ability to recognize speech in virtual environments.

4.5 Experiment 3 – The effect of incongruent localization cues

The third experiment attempted to show that incorrect visual cues for sound source location could influence the spatial unmasking of speech in noise. The experimental setup for this experiment was explained in Section 3.5. The results found for the different conditions are shown in Fig 4.6.

4.5.1 Results

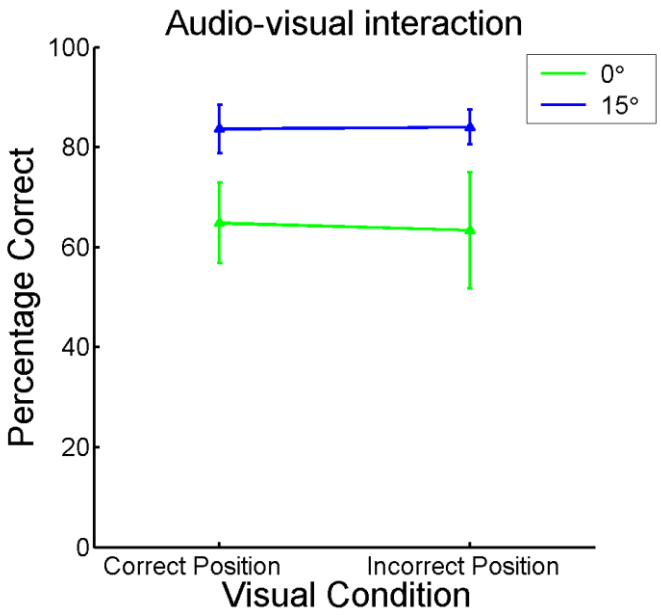

Fig 4.6: Subject performance under different visual and spatial conditions. The dark line represents the spatially separated condition where the target sound was located at 15º to the right. The lighter line represents the co-located condition where both the auditory target and masker was located at 0º. A spatial release from masking was observed for both visual conditions. The incorrect positioning of the visual cue did not have any significant effect on subject performance.

The speaking character's face was animated during this experiment, resulting in better overall speech understanding than observed in the previous experiment. The influence of lip animation will be covered in greater detail in Experiment 4. Fig 4.6 shows the mean and standard deviations of subjects' performance. The visual target was animated in both conditions but was incorrectly positioned in the one. Data for the correct condition is shown on the left while data for the incorrectly positioned target can be seen on the right. An ANOVA of the spatial conditions revealed a statistically significant difference $F(1,3) = 21.62$, $p = 0.019$. Further analysis showed that the difference is significant for both visual conditions. See Table A.4. There was no significant difference between the correctly positioned and incorrectly positioned conditions [$F(1, 3) = 0.1$, $p = 0.774$].

4.5.2 Analysis

In this experiment we expected to reproduce the results from Driver [26] by showing an increase in speech understanding when the auditory objects were co-located and only the visual target was spatially separated from the auditory target and masker. According to Driver the incorrectly positioned visual target can influence the localization of the audio-visual event in the direction of the visual target. The auditory information is in some way integrated with the visual directional information, resulting in a release from masking. We also expected a decrease in speech understanding when the auditory objects were separated and the visual objects co-located. This may cause the visual target to draw audio-visual event in the direction of the masking noise, resulting in greater masking and degraded speech perception.

As in the previous experiments, the results again show that subjects experienced a spatial release from masking. The difference between the two spatial positions was found to be significant for both visual conditions. The results however showed no statistically significant difference between the correct and incorrect visual conditions. This suggests that the incorrect visual cues had no effect on the unmasking of speech. This is contrary to Driver's results. The main difference between the current experiment and that of Driver is the type of masking noise. An energetic noise masker was used for this experiment while Driver made use of an informational voice masker. It seems that the release from masking obtained for informational maskers does not carry over to energetic maskers. This result agrees with our previous experiment that indicated that visual localization cues are not beneficial in overcoming energetic masking effects. Further research is needed to investigate the influence of greater spatial separation between the auditory and visual cues and to compare the effectiveness of different maskers. See Chapter 5 – Future work.

In all other experiments, care was taken that the auditory and visual positions of sounds correlate. The head-mounted display ensured that the auditory and visual virtual environments were always aligned no matter how the user moved their head. Most virtual environments are however viewed on a computer monitor. Users are normally not perfectly positioned and this will cause misalignment as explained in Section 2.5.1. These results show that having perfectly aligned auditory and visual space may not be very important for speech perception in noisy virtual environments.

4.6 Experiment 4 – The influence of lip animation

The last experiment investigated the influence of both correct and incorrect lip-animation on speech perception in virtual environments. The results from this experiment can be seen in Fig 4.7.

4.6.1 Results

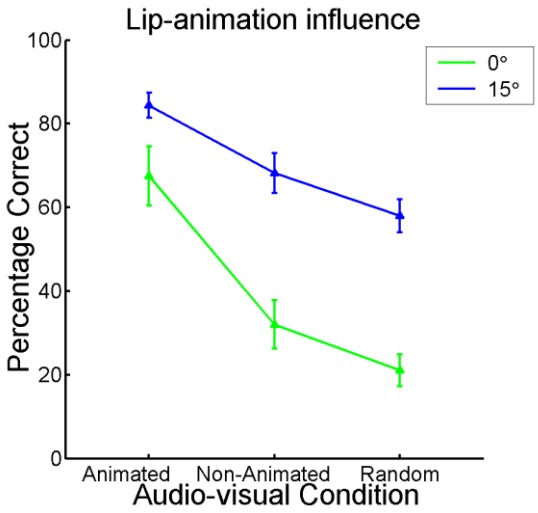

Fig 4.7: Subject performance under different visual and spatial conditions. Correctly animated, non-animated and randomly animated visual stimuli were presented. The dark line represents the spatially separated condition where the target sound was located at 15º to the right. The lighter line represents the co-located condition where both the target and masker was located at 0º. A spatial release from masking was observed for all three visual conditions. Subjects performed best for correct lip animations and worst when incorrect animations were used.

Fig 4.7 shows subject performance for different visual conditions. From left to right the conditions were: correctly animated, non-animated and randomly animated. An ANOVA between the co-located and separated conditions again showed a significant difference between the two spatial conditions [$F(1,3) = 664.97$, $(p<0.001)$]. Further analysis showed that the difference is significant for all visual conditions. See Table A.5. An ANOVA across the different visual conditions also revealed a statistically significant difference between the three visual conditions [$F(2, 6) = 28.2$, $p < 0.001$]. Further comparisons revealed that all visual conditions differ significantly for both spatial conditions. See Table A.6.

Fig 4.8 compares the hearing benefit between no-lip animation and correct lip animation for the two different spatial conditions. Correct lip animation provided a greater benefit in the co-located condition than in the separated condition.

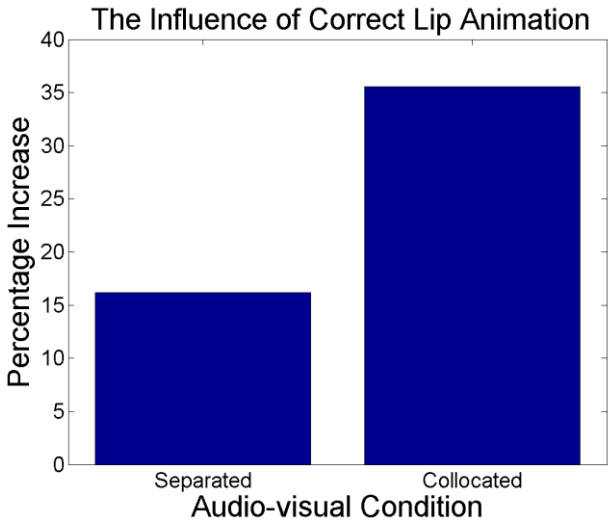

Fig 4.8 The benefit of correct lip animation over no lip animation for the different spatial conditions. In the collated condition both the target and masker objects were presented at 0º. In the separated condition the target was presented at 15º and the masker at 0º. Lip animation had a greater influence in the co-located condition where the lack of directional auditory cues resulted in very challenging listening conditions.

From Fig 4.7 we saw that performance for the incorrectly animated condition was worse than the correctly animated and non-animated visual conditions. In this condition, animations from different colour and number combinations were used as visual stimuli. The question arises whether this incorrect visual information is merely distracting or whether it created a perceptual bias in favour of the visually presented words.

One could use an alternative scoring to determine how well the subject would have performed if we used the visually presented colour and number as the correct response instead of the auditory. If subjects consistently picked the colours and numbers they saw, one could conclude that subjects relied more strongly on the visual than the auditory cues.

From Fig 4.9 it is clear that when scoring in this way there is a dramatic difference in the results. On the left the responses are scored according to the auditory presented stimuli. On the right, subject responses are scored against the visually presented stimuli. The dark line again represents the spatially separated condition and the lighter line represents the co-located condition. For the co-located condition, subjects performed better when using the alternative scoring method. The visual score was significantly higher than the auditory score, which is almost the same as chance (19.6%).

This implies that subjects tended to answer according to the visually presented stimuli, that is, the visemes, in the co-located condition. In the separated condition, where spatial unmasking resulted in better auditory information, subjects tended to answer according to the auditory presented stimuli, ignoring incongruent visual information. In this condition the visual score was slightly above chance indicating that the incorrect animation still had some impact.

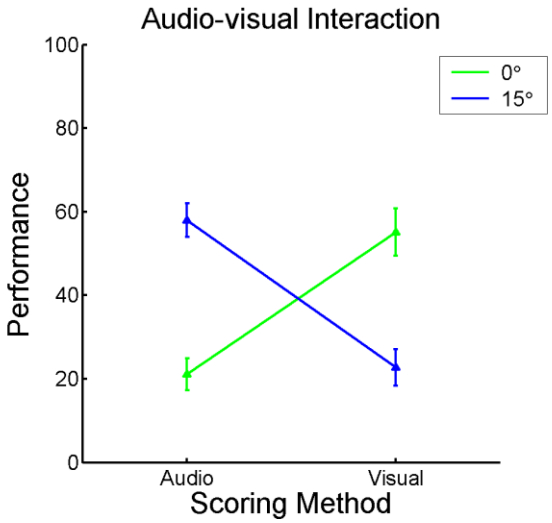

Fig 4.9.: Subject performance when using two different scoring methods for the randomly animated condition. In this condition colours and numbers that were presented visually in terms of the animated visemes did not match the auditory stimuli. On the left the responses are scored according to the auditory presented stimuli. On the right, subject responses are scored against the visually presented stimuli. The dark line again represents the spatially separated condition and the lighter line represents the co-located condition. For the co-located condition subjects perform better when using the visual scoring method. This implies that subjects tended to answer according to the visually presented stimuli in this condition. In the separated condition, where spatial unmasking resulted in better acoustic cues, subjects tended to answer according to the auditory presented stimuli, ignoring incongruent visual information.

4.6.2 Analysis

In the last experiment we expected animation to aid in speech recognition. The results confirmed this and show that correct lip animation significantly contributes to speech understanding. A spatial release from masking was observed for all visual conditions. This resulted in better speech perception when the target and masker was spatially separated than when their positions coincided. Fig 4.8 shows the benefit of lip animation to speech perception is greater in the co-located condition than in the separated condition. This suggests that visual cues become more important when the auditory cues are weak. This is consistent with findings in a similar study by Helfer *et a*l [43]. Their experiments were conducted with the use of a video stream instead of animations.

As expected, the incorrect animated condition did result in worse performance than the non-animated condition. It may be that the interaction between visual and auditory information caused subjects to hear something completely different as is found in experiments involving the McGurk effect. However McGurk experiments are generally very carefully constructed. Only some combinations of strong visual cues with opposing weak auditory cues produce this effect. It is unlikely that the vocabulary of the CRM corpus would result in any McGurk effects when presenting random combinations of the auditory and visual stimuli. The massive increase in performance between the non-animated and correctly animated case in the co-located condition suggests that subjects are able to lip-read well. It therefore seems likely that the visual cue had a big influence during the incorrectly animated condition. From Fig 4.9 we can see that subjects indeed scored higher for the visually presented visemes than for the auditory phonemes. This suggests that at least for the co-located condition the visual cue was favoured. In the separated condition subjects performed reasonably well and the auditory score was better than the visual score. These results are consistent with findings in the literature that suggests that the stronger cue will usually be favoured when two sources of information conflict [24]. In the co-located condition the auditory cues are very poor while the visual cues are comparatively stronger. In the separated condition the auditory cues are stronger and they are favoured above the visual cues.

Overall these results suggest that adding lip animation to characters in virtual environments will significantly improve speech understanding if done correctly. What makes these results even more interesting is that the animations used were extremely basic. Other studies suggest that 5 unique frames per second is the bare minimum for animation to contribute to speech recognition [34]. Those results were obtained with the use of a video stream and the 5 unique frames did not necessarily include the visemes linked to each phoneme. By constructing the animation in such a way that all visemes are included, a significant improvement in speech understanding can still be obtained with minimal effort.

Note that these conclusions are only relevant under conditions where it is very difficult to hear. Under normal listening conditions the strong auditory cues will usually be enough to disambiguate an incongruent visual cue. These results do however show that users rely heavily on visual cues in adverse hearing conditions.

4.7 Summary

In this chapter it was first shown that all subjects participating in this research achieved comparable performance in the adaptive trials. Small inter-subject variability is important in experiments where only a few subjects are used. The auditory stimuli used were then shown to be comparable to those used in previous studies. Finally the results of for each experiment were presented. It was shown that modern hardware could produce a spatial release from masking. Correct or incorrect visual cues for sound source location did not have a significant effect on hearing performance. Lastly it was shown that even rudimentary lip animation could contribute to speech perception but that incorrect lip animation was detrimental to the perception of speech.

Chapter 5
Conclusion

In this research we have shown that the recreation of perceptual cues in virtual environments has great benefits to the perception of speech. We first showed that even the simplified methods of producing 3D sound with consumer sound hardware have a positive influence on speech understanding. We then showed that having accurate knowledge of where a speaking character is located is not important for speech discrimination in noisy environments. We also showed that the inadequacies inherent in the audio-visual display devices do not have a negative influence on the perception of speech in virtual environments. Finally we demonstrated that even rudimentary lip animations could have a significant impact on the perception of speech in noisy virtual environments.

A surprising result of our first experiment was that in some conditions subjects performed better when using modern sound hardware than when using accepted pre-computed sound spatialization. When the target and masker sounds were presented from the same location, where masking effects would be the greatest, subjects performed better when using real-time sound spatialization techniques than when using the pre-computed method. Our results show that at least for energetic noise maskers, masking is less effective when using real-time sound spatialization with either headphones or surround sound speaker displays, resulting in better hearing performance. We also showed that a spatial release from masking could be obtained for all three methods of sound spatialization. This result is significant since previous accounts in the literature have only demonstrated this effect using pre-computed methods. No previous studies have made use of modern consumer sound cards' ability to spatialize speech in real-time. Our results show that virtual environments that make use of this technology are capable of producing a spatial release from masking for both headphone and surround sound displays and can enhance hearing performance by presenting speech in 3D. The hearing benefit observed in cocktail-party experiments therefore naturally carries over to virtual environments that employ this technology. The clarity of dialog in movies could greatly be improved by panning different voices between speakers as actors move across the screen. However, to preserve auditory spatial continuity when cutting between camera angles, movies usually only present dialog only from a single location in a theatre [46]. Since the user is usually in control of the camera in interactive virtual environments, sudden changes in a voice's location would normally not occur. Dialog presented in virtual environments can therefore greatly benefit from making use of 3D sound spatialization hardware.

In our investigation of visual cues for sound source localization, we expected more accurate knowledge of a talker's location to benefit speech perception in virtual environments. We observed no difference in hearing performance when subjects were able to see the talking character than when only an auditory voice was presented from the same location. The results from our experiment, which made use of a noise masker, are contrary to Reisberg's observations in a similar experiment using a speech masker [73]. He showed slightly better hearing performance when subjects could see the speech source (with no lip cues) than when the speech source was hidden behind a curtain. It seems that the process of audio-visual integration is more effective in overcoming informational masking than energetic masking effects. We believe that a noise masker is a more realistic representation of the conditions that obtain in virtual environments.

This idea was reinforced by our third experiment. Although correct localization cues used in the previous experiment did not benefit hearing performance, we were also interested to see whether misleading visual information could influence speech perception. Driver has previously demonstrated

that an incorrect visual localization cue could draw auditory attention in the direction of the visual representation, resulting in a spatial release from masking [26]. Where Driver used a second voice as an informational masker, we expected to show a similar improvement using a noise masker. Furthermore, we expected that it could also be possible for visual cues to have a detrimental effect on speech perception when it draws auditory attention in the direction of the masking sound. Our results however showed that incorrect visual localization cues have no influence on speech perception in the midst of a noise masker. This result has promising implications for virtual reality applications that present multiple distracting background sounds such as special effects, ambient noise and music together with dialog. The nature of the auditory and visual displays typically used in these applications causes the auditory and visual objects to be displayed in different positions. The results from this research show that the spatial discrepancy will not influence the perception of speech in the midst of other distracting sounds, providing that these sounds only produce energetic masking.

Our final experiment demonstrated that even very basic lip animations could benefit speech perception if done correctly. Conversely, we showed that incorrect animations could actually be detrimental to speech perception. The benefit of lip animations was greater under adverse listening conditions than under conditions where the auditory target was clearer. We showed that a much greater reliance is placed on visual cues when the auditory conditions are weak. This was evident when using correct lip animations and when using incorrect lip animations. Under adverse listening conditions subjects tend to answer according to the visually presented words even when different auditory words were presented. This resulted in worse performance when incorrect lip animations were used than when no animations were used at all. This could have implications for virtual environments with dialog in different languages. Creators of virtual environments do not have exact control over what the user will hear at any given time. Adverse listening conditions may sometimes be unavoidable. Having separate animations for different languages therefore becomes more important for virtual environments than for other forms of media like animated films where there is more control over the final audio track.

The results of this research have two major implications on the design and authoring of virtual environments. It was shown that 3D positional sound produced by modern sound hardware could be employed to enhance speech perception in virtual environments. Since this technology is found on most hardware platforms, VE authors should invest the time to integrate spatialized sound into their applications. It was also shown that even simple lip animations could significantly enhance speech perception. Any time spent on lip animations will therefore be well worth the effort.

Future work

In our first experiment we compared the spatial release from masking obtained using different methods of sound spatialization. Although we showed a significant benefit for all three auditory displays, we did not measure the actual target-to-noise ratios produced by the different methods. Such measurements would show to what extent differences in TNR at each ear contributed to the release from masking. In a follow-up study one could measure the TNR. This would require placing inner-ear microphones inside the head of a KEMAR mannequin while presenting sounds using different auditory displays. If notable differences were found, one would also expect a change in the perceived position of the sound sources [7]. The study could further be extended by making use of head-tracking technology to determine the localization accuracy for each of the different auditory displays.

Experiments 2 and 3 investigated the effect of visual localization cues on speech perception in noisy environments. Although these experiments were very similar to that of Reisberg[73] and Driver[26], our masking stimuli differed from those used in their respective studies. Both authors made use of informational speech maskers where our experiments made use of steady state noise, which acted as an energetic masker. Experiment 2 attempted to reproduce Reisberg results, showing that audio-visual presentation of speech even without lip movement cues is still better than auditory only presentation. Experiment 3 attempted to reproduce results found by Driver, showing that the illusionary displacement of a speech source could enhance hearing performance. The current study could not reproduce either of these results when using a noise masker. The influence of masking characteristics

on audio-visual speech perception has not received much attention in the literature. Hygge *et al.* [48] have shown that visual lip movement cues have a greater benefit when using a speech masker rather than a noise masker. In a follow-up study on the current research one could compare the influence of visual localization cues on speech perception for energetic and informational masking sounds.

The results from Experiment 2 however could also be attributed to the nature of the target stimuli. Target sentences were relatively long and it is quite possible that the auditory system had enough time to accurately localize the sound. Visual information for target location would then be largely redundant in the audio-visual condition and provide no benefit over the pure auditory condition. The small number of possible target locations may also have contributed to faster auditory localization. One could repeat this experiment with shorter target stimuli and presented them at more locations. When there is more uncertainty over the sound source location in the auditory modality, it may be still possible to show an increase in the audiovisual condition, even when using a noise masker. However we do not believe this very realistic in practice in virtual environments.

In a follow-up of Experiment 3 one could also attempt to use larger discrepancies between auditory and visual angles to reproduce Driver's results for noise maskers.

Our final experiment indicated that incorrect visual speech cues could be detrimental to speech perception. In order to achieve a greater level of immersion in virtual environments, many applications attempt to provide accurate lip movement for virtual characters using lip animation software like Face Robot [80]. However, even if the facial expressions are correct, it is still possible for the voice and lip movement to be unsynchronized. Our final study can be extended by investigating how the synchronization or lack of synchronization between audio and more accurate visual stimuli influences speech perception in virtual environments.

References

1. ANSI S3.6. (1989). American national standard specification for audiometers. *American National Standards Institute, New York.*

2. Arbogast T.L., Mason C.R, Kidd G. (2002). The effect of spatial separation on informational and energetic masking of speech. *Journal of the Acoustical Society of America, 112, 2086-2098*.

3. Arons B. (1992). A review of the cocktail party effect. *Journal of the American Voice I/O Society, 12, 35–50.*

4. Baker R.J., Rosen S. (1999). Minimizing boredom by maximizing likelihood. *Speech, Hearing and Language: work in progress Vol. 11, 187-200.*

5. Banks. M.S. (2004). NeuroScience: What You See and Hear What You Get. *Current Biology, Vol. 14, 236-238.*

6. Begault R.D. (1994). *3-D Sound for Virtual Reality and Multimedia. Cambridge, MA: Academic Press Professional.*

7. Begault R.D. (2001). Direct comparison of the impact of head tracking, reverberation and individualized head-related transfer functions on the spatial perception of a virtual speech source, *Journal of the Audio Engineering Society, 49, 904-916*.

8. Bertelson P., Vroomen J., De Gelder B., Driver J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception & Psychophysics, 62,* 321-332.

9. Best V., Van Schaik A., Carlile S. (2003). Two-point discrimination in auditory displays. *Proceedings of the 2003 International Conference on Auditory Display, 9, 17 - 20.*

10. Blair P. (2005). Cartoon Animation. *Available online at* http://www.freetoon.com.

11. Blauert, J. (1983). *Spatial Hearing. Cambridge, MA: MIT Press*

12. Braida L.D. (1991). Crossmodal Integration in the Identification of Consonant Segments. *Journal of Experimental Psychology, 43A, 647-677*.

13. Bregman A. (1990). Auditory scene analysis. . *Cambridge, MA: MIT Press*

14. Bronkhorst A.W., Plomp R. (1986). The effect of head-induced interaural time and level differences on speech intelligibility in noise. *Journal of the Acoustical Society of America, 83, 1508-1515*

15. Bronkhorst A.W., Plomp R. (1992). The effect of multiple speech like maskers on binaural speech recognition in normal and impaired hearing. *Journal of the Acoustical Society of America, 92, 3132-3139*

16. Brungart D. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *Journal of the Acoustical Society of America, 109, 1101-1109*

17. Brungart D. (2001). Evaluation of speech intelligibility with the coordinate response measure. *Journal of the Acoustical Society of America, 109, 2276-2279*

18. Chen T., Rao R.R. (1998). Audio-visual integration in multimodal communication. *Special Issue on Multimedia Signal Processing, 86, 837-852*.

19. Cherry E.C. (1953). Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustic Society of America, 25, 975–979*.

20. Choe C.S., Welch R.B., Gilford R.M. (1975). The "ventriloquist effect": Visual dominance or response bias? *Perception & Psychophysics 18, 55-60*

21. Creative Technology Ltd. (2005). X-Fi Performance And Technology Specs. *Available online at: http://www.soundblaster.com/products/x-fi/technology/architecture/specs.asp*

22. Culling J., Hodder K., Toh C.Y. (2003). Effects of reverberation on perceptual segregation of competing voices. *Journal of the Acoustic Society of America, 114, 2871-2876*

23. Darwin C.J, Hukin R.W. (2000). Effectiveness of spatial cues, prosody and talker characteristics in selective attention. *Journal of the Acoustic Society of America, 107, 970-977*.

24. Dekle D.J., Fowler C.A, Funnell M.G. (1992). Audiovisual integration in the perception of real words. *Perception & Psychophysics 51, 355-362*.

25. Dolby Laboratories. (1998). *Dolby Surround Mixing Manual S98/11932/12280, Dolby Laboratories Inc.*

26. Driver J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading *Nature 381, 66-68*

27. Drullman R., Bronkhorst A.W. (2000). Multichannel speech intelligibility and talker recognition using monaural, binaural and three dimensional auditory presentation. *Journal of the Acoustic Society of America, 107, 2224-2235*

28. Durlach N.I., Mason C.R., Kidd G, Arbogast T.L., Colburn S., Shinn-Cunningham B.G. (2003). Note on informational masking. *Journal of the Acoustic Society of America, 113, 2984-2987*

29. Ebata M., Sone T., Nimura T. (1968). Improvement of hearing ability by Directional Information. *Journal of the Acoustic Society of America, 43, 289-297*.

30. Ebata M. (2003). Spatial unmasking and attention related to the cocktail party problem. *Acoustics, Science and Technology, 24, 208-219*.

31. Foley J., Van Dam A, Feiner S., Hughes J. (1990). Computer Graphics, Principles and Practise. *Addison-Wesley*

32. Freyman R.L., Helfer K.S., McCall D.D., Clifton R.K. (1999). The role of perceived spatial separation in the unmasking of speech. *Journal of the Acoustic Society of America, 106, 3578-3588*

33. Freyman R.L., Balakrishnan U., Helfer K.S. (2001). Spatial release from informational masking in speech recognition. *Journal of the Acoustic Society of America, 109, 2112-2122*

34. Frowein H.W., Smoorenberg G.F., Pyters L., Schinkel D. (1991). Improved Speech Recognition Through Videotelephony: Experiments with the Hard of Hearing. *IEEE Journal on Selected Areas in Communication, 9, 611-616*

35. Funkhouser T., Jot J., Tsingos N. (2002). "Sounds good to me!" Computational Sound for Graphics, Virtual Reality, and Interactive Systems. *SIGGRAPH 2002 course notes*.

36. Gardner B., Martin K. (1994). HRTF Measurements of a KEMAR Dummy-Head Microphone. *Available online at http://sound.media.mit.edu/KEMAR.html*

37. Grant K.W., Seitz P.F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustic Society of America, 108, 1197-1207*

38. Grant K.W., Walden B.E., Seitz P.F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition and auditory-visual integration. *Journal of the Acoustic Society of America, 103, 2677-2690*

39. Grant K.W., Seitz P.F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *Journal of the Acoustic Society of America, 104, 2438-2450*

40. Green K.P., Miller J.L. (1984). On the role of visual rate information in phonetic perception. *Perception & Psychophysics, 38, 269-276*

41. Hartman, W. (1998). *Signals, Sound, and Sensation. AIP Press.*

42. Hawley M.L., Litovsky R.Y. Colburn H.S. (1999). Speech intelligibility and localization in a multi-source environment. *Journal of the Acoustic Society of America, Vol. 105, 3436-3447*

43. Helfer K.S, Freyman R.L. (2005). The role of visual speech cues in reducing energetic and informational masking. *Journal of the Acoustic Society of America, 117, 842-849*

44. Hendrix C., Barfield W. (1996). The sense of presence in auditory virtual environments. Presence 5, 290-301

45. Hirsh I.J. (1950). The Relation between Localization and Intelligibility. *Journal of the Acoustic Society of America, 22, 196-200*

46. Holman T. (2002). Sound for film and television 2nd Edition*, Focal Press*

47. House A., Williams C., Hecker M., Kryter K. (1965). Articulation testing methods: Consonant differentiation with a closed response set. *Journal of the Acoustic Society of America, 37, 158-166*

48. Hygge S., Ronnberg J. (1992). Normal-Hearing And Hearing-Impaired Subjects' Ability To Just Follow Conversation In Competing Speech, Reversed Speech, And Noise Backgrounds. *Journal of Speech & Hearing Research, Vol. 35, 208-216*

49. IASIG. (1998). 3D Audio Rendering and Evaluation Guidelines*, MIDI Manufacturers Association Inc.*

50. Kryter K. (1962) Methods for the calculation and use of the articulation index. *Journal of the Acoustic Society of America, 34, 1689-1697*

51. Koehnke J., Besing J.M. (1996). A Procedure for Testing Speech Intelligibility in a Virtual Environment. *Ear & Hearing. 17, 211-217*

52. Lavagetto F. (1995). Converting speech into lip movements: A multimedia telephone for hard of hearing people. *IEEE Transactions on Rehabilitation Engineering, 3, 1-14*

53. Levitt H. (1971). Transformed Up-Down Methods in Psychoacoustics. *Journal of the Acoustic Society of America, 49*, 467-476

54. Levitt H., Rabiner L.R. (1967). Binaural Release From Masking for Speech and Gain in Intelligibility. *Journal of the Acoustic Society of America*, *42, 601-608*

55. Licklider J.C.R. (1948). The influence of Interaural Phase Relations upon the Masking of Speech by White Noise. *Journal of the Acoustic Society of America*, *20, 150-159*

56. Madisetti V.K., Williams D.B. (1998). *The digital signal-processing handbook. CRC Press*

57. Microsoft Corp. (2004). *MS DirectX Programming Guide. Available at http://www.msdn.microsoft.com/directx/*

58. Miller J.D., Wenzel E.M. (2002). Recent developments in SLAB: Software-based system for interactive spatial sound synthesis. *Proceedings of the 2002 International Conference on Auditory Display, 1-6*

59. Mcguigan F.J. (1968). Experimental Psychology. A Methodological Approach. *Prentice Hall Inc.*

60. McGurk H, McDonald J. (1976). Hearing lips and seeing voices. *Nature, 264, 746-748*

61. McGurk H, McDonald J. (1978). Visual influences on speech perception processes. *Perception & Psychophysics, 24 , 253-257*

62. McGrath M., Summerfield Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *Journal of the Acoustic Society of America, 77, 678-685*

63. Moore, T. (1981). Voice communication jamming research. *AGARD Conference Proceedings: Aural Communication in Aviation, 331, 2:1–2:6*

64. Naef M., Staadt O., Gross M. (2002). Spatialized audio rendering for immersive virtual environments. Proceedings of the ACM Symposium on Virtual Reality Software and Technologies 2002, 65- 72

65. NVIDIA Corp. (2002). Technical Brief: The Audio Processing Unit (APU). *Available at http://www.nvidia.com/page/nf2_tech.html*

66. O' Donnell M. (2002). Producing Audio for 'Halo', *Gamasutra., May 20, 2002, URL:http://www.gamasutra.com*

67. Pentland A. (1980). Maximum likelihood estimation: The best PEST. *Perception & Psychophysics, 28, 377-379*

68. Plomp R. (1976). Binaural and Monaural Speech Intelligibility of Connected Discourse in Reverberation as a Function of Azimuth of a Single Competing Sound Source(Speech or Noise) *Acustica., 35, 201-211*

69. Pulkki V. (1997). Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society, 45, 456-466*

70. Recanzone G.H. (1998). Rapidly induced auditory plasticity: The ventriloquism after effect. *Proceedings of the National Academy of Sciences, 95, 869-875*

71. Redeau M., Bertelson P. (1976). The effect of a textured visual field on modality dominance in a ventriloquism situation. *Perception and Psychophysics, 20, 227-235.*

72. Rife D.D., Vanderkooy J. (1989). Transfer-function measurements with maximum-length sequences. *Journal of the Audio Engineering Society, 37, 419-444*

73. Reisberg, D. (1978). Looking Where You Listen: Visual Cues and Auditory Attention. *Acta Psychologica, 42, 331-341.*

74. Repp B.H., Penel A. (2000). Auditory Dominance in temporal processing: New Evidence From Synchronization With Simultaneous Visual and Auditory Sequences. *Journal of Experimental Psychology, 28, 1085-1099*

75. Rose J. (2003). Reality (sound) bytes: Audio tricks from the film and TV studio. *Proceedings of the International Conference on Auditory Display, 9, 33 - 37*

76. Sensaura (*2003*). *GameCODA Concepts Guide Issue 1.5. Technical document CODA/005/0803/1. Sensaura Ltd.*

77. Shilling R., Shinn-Cunningham B.G. (2002). Virtual Auditory Displays Chapter 4. *Handbook of Virtual Environments Technology, Lawrence Erlbaum Associates Inc.*

78. Shinn-Cunningham B.G., Schickler J., Kopco N., Litovsky R. (2001). Spatial unmasking of nearby speech sources in a simulated anechoic environment. *Journal of the Acoustic Society of America, 110, 1118-1129*

79. Shinn-Cunningham B.G., Ihlefeld A. (2004). Selective and divided attention: Extracting information form simultaneous sound sources. *Proceedings of the 2004 International Conference on Auditory Display, 10, 51-59*

80. Softimage Co. (2006). SOFTIMAGE|FACE ROBOT. *Available at http://www.softimage.com/products/face_robot/default.aspx*

81. Spence C., Ranson J., Driver J. (*2000*). Cross-modal selective attention: On the difficulty of ignoring sounds at the locus of visual attention. *Perception & Psychophysics, 62, 410-424*

82. Strous M. (2005). Pamela – Lipsynch utility for Moho. *Available online at http://www.personal.monash.edu.au/~myless/catnap/pamela/index.html*

83. StatSoft. (2004). *Statistica Electronic Manual.*

84. Sumby W.H., Pollack I. (1954). Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustic Society of America, 26, 212-215*

85. Sutherland I. (1998). A Head-mounted three-dimensional display. Seminal Graphics: pioneering efforts that shaped the field. New York, NY, ACM Press.

86. Vroomen J., Bertelson P., De Gelder B. (2001). The ventriloquist effect does not depend on the direction of automatic visual attention. *Perception & Psychophysics, 63, 651-659*

87. Vroomen J., De Gelder B. (2000). Sound enhances visual perception: Cross-Modal Effects of Auditory Organization on Vision. *Journal of Experimental Psychology, 26, 1583-1590*

88. Warren D.H., Welch R.B., McCarthy T.J. (1981). The role of visual-auditory "compellingness" in the ventriloquism effect: Implications for transitivity among the spatial senses. *Perception & Psychophysics, 30, 557-564*

89. Watanabe K., Shimojo S. (2001). Effects of auditory grouping on visual motion perception. *Psychological Science, 12, 109-116*

90. Wenzel E.M., Arruda M., Kistler D.J., Wightman F.L. (1993) Localizing using non-individualized head-related transfer functions. *Journal of the Acoustic Society of America, 94, 111-123*

91. Wenzel E.M., Foster S.H. (1993) Perceptual Consequences of Interpolating Head-Related Transfer Functions During Spatial Synthesis. *Proceedings of the 1993 Workshop on the Applications of Signal Processing to Audio and Acoustics, 102-105*

92. Zhou B., Green D.M., Middlebrooks J.C. (1992) Characterization of external ear impulse responses using Golay codes. *Journal of the Acoustic Society of America, 92, 1169-1171*

Appendix – ANOVA Analysis

This appendix provides Post-hoc p-levels for the Newman-Keuls test performed for each repeated measures ANOVA analysis. This test was performed where the ANOVA indicated a significant difference between conditions to determine which conditions contributed to the effect.

	p-level
DirectX 0º / 15º	**0.000908**
KEMAR 0º / 15º	**0.000298**
Surround 0º / 15º	**0.008289**

Table A.1: Significance of performance differences between spatial conditions of Experiment 1. The performance differences between spatial positions were found to be significant for all auditory displays.

	Collocated	Separated
DirectX / KEMAR	**0.001447**	0.071305
Surround / KEMAR	**0.000689**	0.708670
DirectX / Surround	0.065183	0.092140

Table A.2: Significance of performance differences between auditory displays in the co-located and separated conditions of Experiment 1. Significant p-values are indicated in bold. The DirectX and Surround displays differed significantly from the KEMAR display but not from each other.

	p-level
Visible 0º / 15º	**0.000550**
Invisible 0º / 15º	**0.000678**

Table A.3: Significance of performance differences between spatial conditions of Experiment 2. The performance differences were found to be significant for both visual conditions.

	p-level
Correct 0º / 15º	**0.000945**
Incorrect 0º / 15º	**0.001298**

Table A.4: Significance of performance differences between spatial conditions of Experiment 3. The performance differences were found to be significant for both visual conditions.

		p-level
Animated	0º / 15º	**0.000721**
Non-animated	0º / 15º	**0.000246**
Random animation	0º / 15º	**0.000227**

Table A.5: Significance of performance differences between spatial conditions of Experiment 4. The performance differences between spatial positions were found to be significant for all visual conditions.

	Collocated	Separated
Animated / Non-animated	**0.000228**	**0.000503**
Animated / Random animation	**0.000245**	**0.000279**
Non-animated / Random animation	**0.002440**	**0.007343**

Table A.6: Significance of performance differences between visual conditions in the co-located and separated spatial conditions of Experiment 4. Significant p-values are indicated in bold. All visual conditions differed significantly from another for both the co-located and separated conditions.