Spatializing Simultaneous Speech with Application to Increased Understanding

Honours Thesis

Simon Plain

Supervisor: Professor P. Kabal

April 29, 1997

Abstract

How well we understand speech from several simultaneous speakers depends on several factors, including whether we process the speech in a serial or parallel fashion. Our ability to understand one speaker amongst many is quite strong but a much greater challenge is to understand two speakers at once, especially if we are hearing these speakers over headphones. An enhancement to the speech which will aid us in this endeavor is desired.

One proposed method of increasing this ability is to cause the listener to perceive the speakers as being separated in space. This paper will examine how this can be done using digital signal processing to allow the listener to hear the moved speech over headphones.

Our perception of a speakers' location has do with the speakers' direction relative to the listener and the environment (room, open space, concert hall, etc.) around them. The speaker's direction can be simulated by filtering the speech through a stereo finite impulse response filter called an HRTF (Head Related Transfer Function). The speaker's distance can be simulated by sending reverberation to the listener. Reverberation is composed of early and late reflections of sound off the surfaces in a room. The ratio of direct sound to reverberant sound is a strong cue to the distance of a sound source.

An algorithm was implemented to perform these transformations and tested with several subjects. The subjects were able to determine direction fairly well although the well documented front-back reversal error was often encountered. Distance is difficult to model properly in headphones due to the sound source being right beside the ear. Consequently, tests on subject's estimation of distance distance resulted in the judgement being quite a bit shorter than the designed distance. Reverberation, however, clearly helped in externalizing the sound from the head.

Spatialization of sound was then applied to the problem of parallel speech understanding, and several tests were performed. The results indicated that parallel speech was indeed easier to understand when the speakers were separated and externalized from the head. Understanding was higher for separated speakers than for one speaker in each ear (one mono-phone speech file per ear) and for both speakers superimposed at a distance external to the head.

Acknowledgements

I would like to thank my supervisor, Peter Kabal, for his direction when it was needed. Without his help I would have been lost long ago. I would also like to thank Margaret Sharp for lending her voice, and all the subjects of my tests for their help as well.

Contents

List of Figures

Section 1

Introduction

There has been a great deal of work recently in the field of three dimensional (3-D) simulations of sound for various applications[1, 2, 3, 7, 11]. Most of these applications relate to the creation of a kind of virtual reality sound field.

3-D sound involves conceptualizing the location of a virtual sound source relative to the listener and manipulating sound so it seems to be coming from that virtual source. This is done through a series of steps to:

(i) Change the direction from which the sound is coming.

(ii) Change the perceived distance of the sound.

If these two goals can be effectively accomplished through signal processing so that the sound is truly spatial over headphones, three dimensional sound can be achieved.

Other ways of implementing three dimensional sound, including the use of loudspeakers, have been explored [15], but are not practical for many applications.

One application of three dimensional sound which has not been much investigated is that of increasing people's understanding of simultaneous speech.

It has been well documented that a person in a crowd is able to focus in on the speech of one person in that crowd, despite the confusion of environmental noise (other people talking). Several researchers have demonstrated [5, 13, 19] that spatial separation of speakers is a key factor in people's ability to perform selective listening. Work has also been done on the extent to which humans

can understand simultaneous speech in parallel. Parallel understanding is a much more difficult task which, from previous research and some experiments described in this document, seems to be aided by separation of speakers. The ability to process sound so that listeners could understand two speakers at a time has potential application for situations where there are two or more speakers, such as in a plane's cockpit, and two or more are important to be understood simultaneously. Changing the directions from which two speakers are speaking to you is easy when the speakers are present in a room with you, but much more difficult when you are hearing their speech over headphones [16]. Here is where the idea of three dimensional sound processing comes in.

The endeavor described in this document is an attempt to use 3-D signal processing to spatialize two speakers and, in a preliminary fashion, determine the effectiveness of this algorithm for simultaneous recognition of speech.

Section 2

Background

2.1 Overview of Spatial Hearing

Head Shading Cues to Spatial Hearing

The first question we must ask is how is the location of a speaker detected by the ears and brain? When we are being spoken to, sound waves reach our ears from the direction of the person speaking. Let us imagine that someone is speaking to us from our left. The ear the speaker is closest to (the left) receives the information before the right ear. The time lapse between the two ears receiving the signal is called the Interaural Time Difference (ITD). The right ear also hears the speech to be slightly quieter than the left ear did for two reasons:

1. The right ear is further away, and sound intensity decreases with distance.

2. The sound has had to go around the listener's head. In other words, the head has shielded the right ear from some of the noise.

Together these two effects cause what is called the Interaural Intensity Difference (IID). Another cue which tells us where a person is speaking from is the shaping effects our ears, specifically our pinnae (external part of the ear), have on the incoming sound. This shaping effect helps in sound localization by making the ear more sensitive to sounds coming from the front of the listener then from behind the listener [12]. All of these physical effects and our brain's interpretation of them help us determine from what direction a sound originates.

Distance Cues to Spatial Hearing

The other main aspect of spatial hearing is determining how distant the sound source is. This is mainly done through two cues:

1. The loudness of the sound.

2. Sound reverberating off of physical items.

The amplitude of sound decreases by $\frac{1}{r^2}$ where r is the distance from the sound source to the listener [3], so obviously the distance from a source makes a big difference in how sound is heard. In audio files, loudness is determined by many factors, including how high the volume is set on your headphones. For the situation where the final volume of the sound cannot be controlled, the listener determines the distance of a speaker by the volume of the speaker relative to other stimuli such as reverberation from walls, as described below.

Sound reverberations or "echos" off walls, floors, etc. of the area, also give you a strong sense of how far away a speaker is. When a speaker produces sound from their mouth, the sound behaves much as if it had been produced from a point source and sound waves travel in all directions. One of those directions is toward the receiver (assuming there are no objects in the path between speaker and listener), but the sound also strikes the walls, ceiling, and floor. When it strikes any surface, the sound bounces off that surface with some reduction in amplitude. The reflected sound reaches the listener, who hears the much quieter echo as coming from the direction of the wall. This reflected sound wave is called an "early reflection". These early reflections continue to bounce around the room until their magnitude becomes negligible. When the growing number of early reflections results in a dense reverberation, they are called "late reflections". The early reflections very quickly ($\approx$ 50ms) degenerate into late reflections, the waveforms of which resemble exponentially decaying noise.

The ratio of the loudness of the sound coming directly from the source to the loudness of the echoing sounds tells us how distant the emitting source is from us [3].

Figure 2.1 shows an example impulse response of a room where the early reflections are noted and most of what remains is the late reflection.

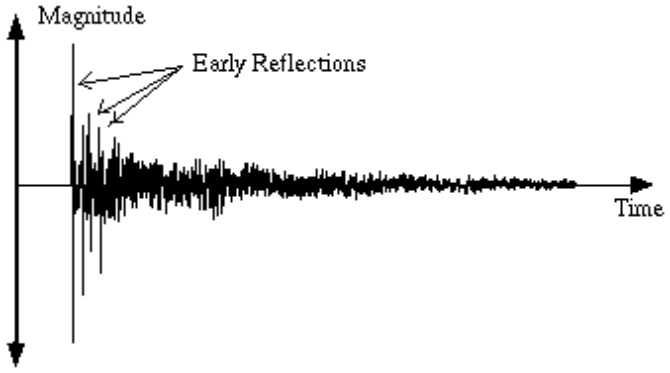

Figure 2.1: Impulse response of a room [7]

Measurement of Spatialization Position

In this document I will be giving measurements of the angle between the receiver and the emitter in the following way: from the listener's point of view, if the speaker is directly in from of them and at the same elevation, the speaker is at 0° azimuth and 0° elevation. The azimuth angle increases from 0 to 360 degrees as the speaker circles the listener in a clockwise fashion, so a speaker directly to the left is at 270° azimuth. Elevation is measured as positive degrees if the speaker is above one's elevation and negative if they are below. A speaker directly above would be at 90° elevation. Figure 2.2 gives a diagram of this measurement system.

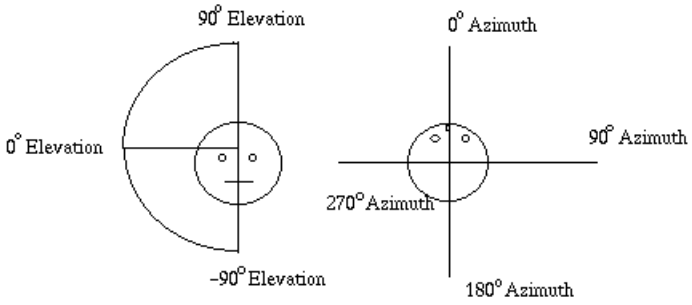

Figure 2.2: Measuring Spatialized Sound Position

2.2 Changing the Direction of a Sound Source

The first objective in spatializing speech is making it sound like it is coming from a certain direction. To do this the speech was filtered with a head-related transfer function (HRTF), usually in the form of a finite impulse response (FIR) filter. This filter is convolved with the speech signal for each ear, producing a binaural signal representing the moved speech.

Begault [3] says "the binaural HRTF can be thought of as a frequency-dependent amplitude and time-delay differences that result primarily from the complex shaping of the pinnae". The HRTFs are measured by placing microphones in either a person or a model of a person's ears and measuring the impulse response received by the subject from sound sources coming at them from many different directions. This HRTF is intended to allow sound to be filtered to model the way it would be heard from a particular direction under normal circumstances. It attempts to model the IUD and IID we would normally experience if a person were speaking from a specified direction.

One problem with using HRTFs to spatialize sound is that a particular set of HRTF data is specific to the artificial or real head it was recorded on. HRTF data measurements are somewhat involved and difficult to perform so it is not practical for each user of a spatialized sound system to obtain their own HRTF set. So, we must make do with an imperfect data set.

Another difficulty is the number of "front-back reversals" which have been observed [1]. These are situations where a sound is filtered with an HRTF designed to make it sound like the source is in front of the subject (say 20° azimuth) and the subject says they hear it from behind (160°). No good explanation has yet been found for this phenomenon except for the simple "If I can't see it, it must be behind me" [3].

2.3 Externalizing Sound with Room Reverberation

The next step in spatialization is to make the sound seem external to the listener's head. Normally when speech is heard on headphones, the sound seems to be in the center of your head. Applying an HRTF to the sound merely makes it seem like it is coming from a certain direction in your head. Accurate externalization of sound with headphones is a relatively difficult feat to accomplish [1, 16]. The directness of the sound source (placed on one's

ears) makes the distance estimated by the listener quite a bit closer than the designer specified distance [2]. In order to effectively externalize the sound, it is necessary to create some artificial reverberation to accompany the sound. This reverberation will help us perceive the sound as distant. This is done by simulating the reflections of sound off a room's surfaces.

There are two basic methods of implementing the application of room reverberation to sound:

(i) Convolution with measured room impulse responses

(ii) Implementation of synthetic reverberation

Using the impulse response of the type of room we wish to simulate as an FIR filter has the advantage of giving a very accurate depiction of the room's sound. However, since room reverberations can last as long as 1-2 seconds, this method involves a great deal of calculation to convolute the sound with the impulse response. Synthetic reverberation can be much quicker but it can also be very tricky to make it sound natural. Since synthetic reverberation is what was eventually decided on for my implementation, that is what I will focus on.

In our particular case, since the simulation of direction as well as distance is desired, we need a binaural (stereo) reverberation method. The decision has to be made as to how to implement the early reflections and late reflections.

The two most popular systems for simulating early reflections are the ray tracing technique and the image model. The ray tracing concept is that a source emits sound "particles" in every direction. An energy distribution is executed several times for each surface: the sound wave strikes the surface, is attenuated by this collision, and then a rebounding wave bounces towards the listener. Attenuation of a wall can be modeled simply as a decrease in gain as a function of the wall material. This is called the "wall coefficient", although in reality the attenuation is not so spectrally flat. Ray tracing is very accurate but very computationally expensive.

The image model method consists of placing a virtual source within mirror image rooms all around the real room. A vector is then drawn from the sound source to the image model and from there to the listener. In this way the angle, distance, and attenuation (including that from the wall) can be calculated for a variable number of image rooms. The greater the number of images calculated, the better the model. Figure 2.3 shows an image room

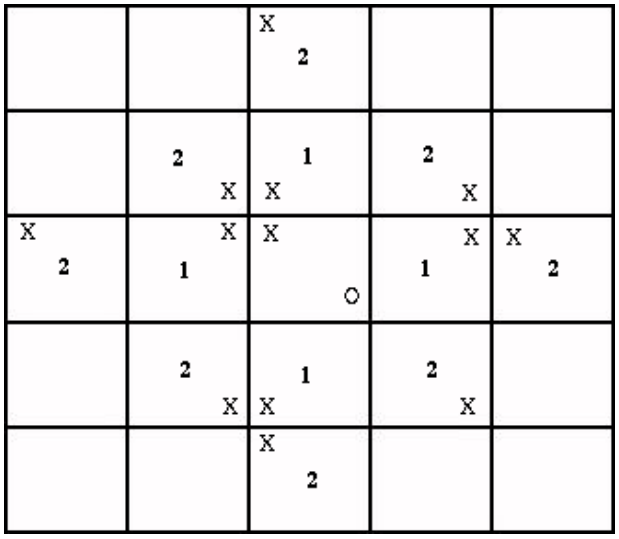

Figure 2.3: First and second order images [3]

pattern where the centre room is the real room, Xs denote sound sources and the O is the listener. The '1's and '2's denote first and second order reflections. Second order reflections have struck two walls, third order have struck three, and so on, so we can see that soon the sound amplitudes become negligibly small.

The next consideration is how to model the late reflections. Considerable work has been done on creating artificial room reverberation in the past 30 years. The now classical way of implementing it is based on a system designed by Schroeder and Login in 1961 [17] and modified by Moorer in 1979 [11]. This system involved sending the original sound through a set of parallel comb filters followed by one or more allpass filters to modify the phase for a more natural sound. The comb filters alone were found to cause a sound similar to that coming through a hollow tube.

Another method which creates very realistic late reverberation is to convolve the sound with exponentially decaying noise. This method is not as fast as the IIR approach above since it is implemented as an FIR filter with hundreds or thousands of taps.

2.4 Human Recognition of Two Simultaneous Speakers

When humans are in an environment where we are assaulted by many sounds at once, we sometimes wish to pick out a particular sound and pay attention to it exclusively. This is called the "cocktail party effect", where several people around us are speaking but we only want to listen to one of them.

Much interesting investigation has gone into our ability to focus in on one speaker [4, 5, 10]. When Cherry mixed two pieces of speech from the same speaker where both speeches were equally heard in each ear, the subjects made significant mistakes in transcription of the two sequences. When he placed one speech in the left ear and the other in the right ear, however, he found that the subjects could transcribe one of the speeches perfectly, but had very little knowledge about what was going on in the other ear. In one experiment in fact, the subject was told to transcribe the speech in the right ear, and half way through, the speech in the left ear was changed to German speech. The subject had no recollection of anything amiss. The subject could remember what gender the speaker on the left was whether it was speech or a tone, but could not reproduce any of the words spoken.

E. Poulton [13] performed experiments comparing subjects' abilities to transcribe speech from one speaker and two simultaneous speakers, alternatively close together and separated. He performed these experiments with the aid of loudspeakers. The results of these experiments are summarized in Table 2.1.

Speakers Monitored	Omissions		Mishearings	
	Spk Apart	Spk Close	Spk Apart	Spk Close
1	4.9%	10.5%	3.5%	6.3%
2	13.2%	15.9%	3.5%	6.5%

Table 2.1: Speaker Separation Effects

This, along with Treisman and Fearnley's work [19] implies that for parallel understanding it is significantly better to have the speakers separated as opposed to having them coming from the same source. A striking lack of ability to process two stimuli in parallel has been associated with dichotic (one speaker per ear) listening. The hypothesis of this paper is that separat-

ing the speakers in space without the extreme of dichotic listening will allow an improvement in the parallel recognition of speech. This differs from E. Poulton's work in that the separation of the speakers will be performed by signal processing and the results heard through headphones.

Section 3

Implementation of 3-D Sound

The most difficult step in this paper was to find an effective algorithm for spatializing sound.

The procedure used to create the spatialization effect will be explained below. Succinctly, HRTF data was used to move the direct sound and its early reflections (using the image method). The late reflections were modeled by convolving this sound with exponentially decaying noise.

Of great use was a set of HRTF data published on the internet [6]. This data was a series of 512 tap FIR filters at a sampling rate of 44100 kHz, stored in big-endian format. The data needed to be converted to a format compatible with the filtering program so I used Professor P. Kabal's CopyAudio program [9] to convert the data to text (for use with his FiltAudio program). The HRTFs had been measured for sound emanating every 5 degrees azimuth and every 10 degrees elevation (-40° to 90° elevation, 0° to 355° azimuth). When the HRTFs were tested, they seemed fairly effective in moving sound around inside the head.

On first attempt, I tried simply passing the sound through these HRTF filters and using a monaural reverberation filter to spatialize the two signals and then recombine them. This gave output speech which was far less spatialized than required i.e. it still sounded like it was inside the head.

Figure 3.1 displays the revised method used for moving the sound. The first step was to create the early reflections. I chose the image model method for these because of its relative simplicity and speed. A ray tracing approach would have required a great deal more processing power, even for non-real time work such as this.

The image model was implemented by calculating the position of the im-

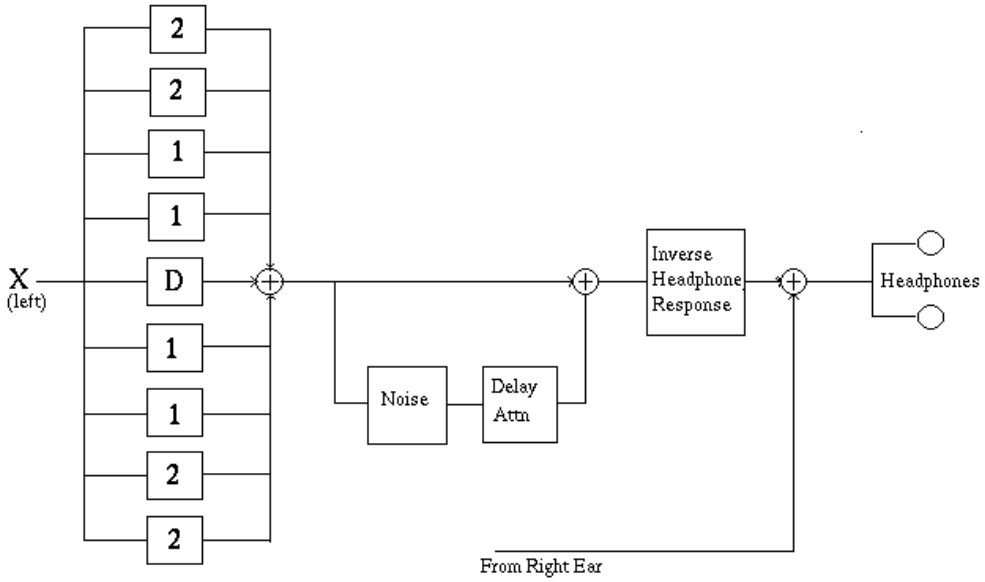

Figure 3.1: Sound Spatializing System

ages, then calculating the distance and angle from the image to the receiver. The angle was rounded to the nearest 5° HRTF model. I decided this was an adequate approximation since Begault found that people's azimuthal detection ability is nowhere near that fine [2], and interpolation would have unnecessarily added complexity. The distance from image to receiver and the wall coefficient were used to determine the attenuation of each particular reflection. A wall coefficient of 0.9 (somewhere between wood and plaster [7]) was used as I found that a lower value resulted in speech which sounded too close, similar to an anechoic chamber. Figure 3.2 illustrates how one early reflection was calculated. The solid line shows the actual path of the sound reflection and the dashed line shows the image source's sound path.

In this experiment, only two dimensional data were used for early reflections, as three dimensional calculations would have increased the complexity significantly and the final aim of this work is not to change the elevation of speakers. Three dimensional echos may, however, increase spatial localiza-

tion.

The 'D' in Figure 3.1 denotes the direct sound which does not need the image method, the angle and attenuation are calculated directly. The '1's on the diagram denote first order reflections and the '2's denote second order reflections, all of which involve two bounces. It was decided to use all four first order reflections but only four of the eight second order reflections since the four used were sometimes closer to the listener than the first order reflections. The unused four second order reflections were somewhat further and were less likely to have a significant effect. Also, the time it was taking to process this algorithm was already becoming unwieldy. The first and second order images can be seen in Figure 2.3 on page 8.

The attenuated signals were then sent through the appropriate left and right HRTF filters to form two signals for each early reflection.

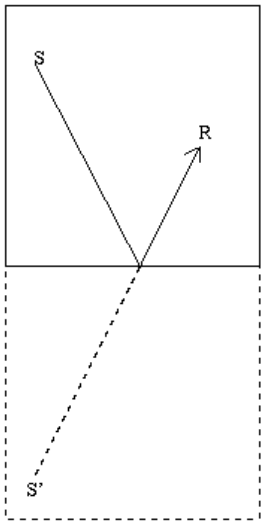

Figure 3.2: One early reflection

At this point, there were left and right signals for nine different angles and distances. All the left signals were then combined into one left speech file. This was repeated for the signals for the right ear.

The next step was modeling late reverberation. I chose not to implement the IIR filter due to the many complications this would entail (the possibility of instability, fine tuning coefficients, etc.) and the late reflection convolution would only have to be performed twice, so filtering time was not a large issue.

I therefore decided on the method of convolving the combined early reflections and direct sound with exponentially decaying noise. The resulting signal was then delayed and attenuated so that its beginning corresponded to the last early reflection.

The noise was generated using the GenNoise program[9] and then decayed and clipped (so there were no values which would clip the signal) using MATLAB. A 2048 point (46.4ms) noise signal was used and the decay was calculated by determining how attenuated the signal should be at the end of the filter using the $\frac{1}{r^2}$ rule and speed of sound. A diagram of the final noise signal is shown in Figure 3.3. In reality, sound in a room does not decay exactly exponentially, but this turned out to be a reasonable model based on the naturalness of the output sound. This is one possible area of improvement in the future. Two noise signals were generated to decorrelate the sound coming to the two ears [3].

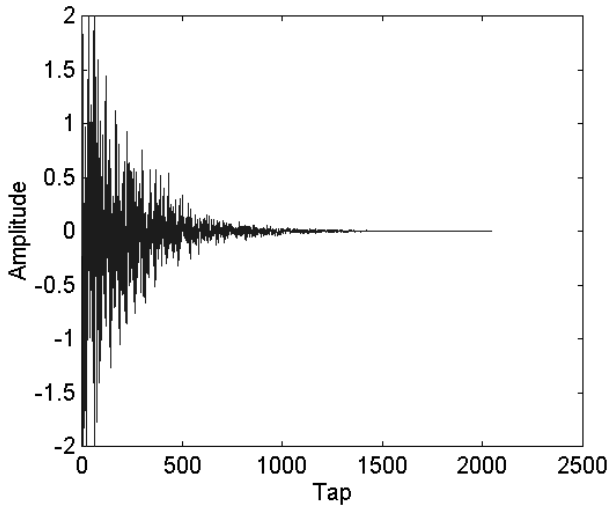

Figure 3.3: Noise used to model late reflections

The two resulting speech files were then sent through a filter modeling the inverse response of headphones. These were provided in the HRTF package and significantly improved the quality of the played back speech.

Finally, the left and right sound files were combined into one stereo file which contained all the location filtering and reverberation.

This algorithm was implemented as a C program which took the dimensions of a room, the location of the speaker and listener within it, and wall

coefficients of the room. It performed the image calculations and output a UNIX shell script to be run using CopyAudio and FiltAudio. The shell script would then perform the entire procedure. An example shell script is shown in Appendix A.

Section 4

Description of Experiment Procedures

4.1 Human Recognition of 3-D Speech

The experiments performed on people's ability to recognize the direction and distance of spatialized sound involved five cases. In each case the subject was played the speech and asked to determine the distance and direction of the speaker. The cases tested were:

1. Monophonic speech (no change in azimuth or distance).
2. Direct sound only, attenuated and HRTF filtered only.
3. Direct sound and early reflections only, no late reflections.
4. Total spatialization algorithm.
5. Same as 4 but with female voice.

The first four tests were with a male speaker. The final test was designed to see if there was a difference in perception of male and female voices. In all cases the sound was designed to be 25° to the right of the listener at a distance of four meters (0° elevation).

Some *ad hoc* experiments were also performed where other values for azimuth were chosen. In general subjects could judge the area the speech came from, although judgements for speech directly in front of the subject were very often misjudged to be directly behind them and reversal errors for

other azimuthal positions also were fairly common. See Begault [2] for more experiments of this type.

In all experiments the subjects kept their eyes open (for a realistic situation) and were permitted to listen to each speech file as many times as they wished. They were asked to describe the distance as less than four inches if it sounded like it was coming from inside their head and greater than four inches if it sounded external to their head.

4.2 Parallel Speech Recognition

The next set of experiments determined the ability of subjects to retain information from two speakers simultaneously. Three tests were performed on each subject. Two pieces of speech, both of the same female voice, were played simultaneously where the two voices were in the following configurations:

1. One mono-phone speech file in each ear, no spatialization.
2. Both speakers directly in front, superimposed (no separation), 4 meters away.
3. Speakers separated at 25° and 335° azimuth.

The subjects were asked to write down how much they could recall of both speeches. They were then asked to listen to each test a second and third time, each time filling in any words they missed after the first listening. It seemed likely that the subjects would have trouble remembering both speeches after just one listen but after three listens should be likely to transcribe both messages in their entirety. This part of the experiment was performed to assess just how difficult it is to recognize parallel speech in these three cases.

Section 5

Experiment Results and Discussion

5.1 Spatialization

This experiment was expected to result in the subjects being able to locate, at least in general terms, the azimuthal position of the designed sound source. With respect to distance estimates however, it was expected there would be some variability in people's judgements, based on Begault's findings [2]. He also found that reverberation causes a decrease in people's ability to determine the azimuthal location of sound.

The monophonic speech signal was for most cases estimated to be inside the head, at the very center. One subject described it as a "fuzzball" inside her head. This result was expected; this is the familiar sound of monophonic speech with headphones.

For the direct only sound, it was expected that the direction estimates would be good while the distance of the speech from the listener would still seem very small. The result was that most subjects still described the sound as in their heads, but behind them to their right. This corresponds to a reversal error (from 25° to 155°).

The next sound file included early reflections but no late reflections. This was expected to be slightly more externalized and perhaps more difficult for detecting its azimuth. The results, interestingly, were that people had a great deal of difficulty in locating the direction of the sound. There were answers in many different directions. This was perhaps due to the possi-

ble unnaturalness of the sound of early reflections from a room with no late reflections. Direction estimates improved with the addition of late reverberation as shown in the next part of the experiment. This experiment was the first time distance estimates started being greater than 4 inches (external to the head). Most distance estimates were of about 6 inches.

The fourth speech had the greatest amount of reverberant sound and it was expected that distance estimates would increase because of the addition of late reverberation. Distance was indeed estimated to be greater than before, though not by a great degree. Most estimates increased to about 10 inches (although one subject determined the distance to be 4 meters where the designed distance was 4.7 meters). One subject described this speech as being "like God talking". The direction judgements improved, as noted above, and most subjects determined the direction to be "behind to the right", again indicating the reversal phenomenon.

The last experiment in this group was the same as the fourth but with a female speaker speaking the same text. With the female speech, distance estimates remained the same but reversals appeared to be less of a problem with several of the subjects revising their estimates to having the speaker in front of them. This outcome is possibly because the voice of the male speaker used in the previous experiments may have had a slightly "directionless" quality to it, or perhaps the female's higher pitch aided in direction estimate. More research in this area would be necessary for definite conclusions.

All in all, it seems that it was possible to move the sound perceptually as well as theoretically, although the difficulty of spatialization using headphones has been demonstrated.

5.2 Simultaneous Listening

The major predictions for the simultaneous listening experiments were that:

1. Dichotic speech would make it easy to understand two speakers in series but result in very poor parallel recognition

2. Superimposed speakers would result in poor recognition for both parallel and serial speech

3. Spatialized, separated speech would allow increased ability to understand two speakers in parallel.

Cherry's work [4] indicated that a listener could focus in on a speaker in one ear while ignoring the other. I therefore expected that after one listen in the first experiment, the subject would recognize only one utterance or else be confused if the two dichotic speakers were overwhelming. If this were the case, probably very little would be retained on the first listen. An improvement would likely be seen in subsequent listens.

When the subjects listened to this first configuration (one speaker per ear) the trend seemed to be as follows. On the first listen, most of one speech and a fraction of the other one were remembered. On subsequent listens, the rest of the two speeches were mostly filled out, although in most cases there were still some omissions and juxtapositions between the speeches.

It seems that since the subjects were told they would be hearing two speeches and should try to understand both of them in parallel, they did not attempt to understand one on the first try and the other on the second try. Instead they seemed to be attempting parallel understanding and having difficulty with this speech configuration.

The second case was the superposition of two speakers placed directly ahead, 4 meters away. Here a great deal of confusion between the two speeches was expected. This was supported by the results, where most of the listeners could not recognize more than 50% of the words, let alone the sentence structure, in all three listens. Only one listener was able to put together the sentences after three tries.

Finally, the subjects listened to two speakers spatialized and separated from each other. The expectation for this experiment was an increase in the ability of subjects to understand the two speeches in parallel, following Poulton's work in this area [13].

As expected, this experiment seemed to show the best results of the three for parallel understanding. After one listen, subjects had generally understood approximately 70% of the two speeches and both speech segments were understood to high accuracy by the second listen.

Section 6

Conclusions and Future Work

Three dimensional audio spatialization has many interesting applications today. Any simulated situation where we wish to increase our sense of realism requires it. It has applications outside of virtual reality however, and this paper has explored one of those areas: increasing people's ability to understand simultaneous speech.

The first step towards this goal was determining an effective way to spatialize speech over headphones. Headphones were chosen since using audio speakers is unrealistic in many situations. To this end, an algorithm was assembled and fine tuned until a satisfactorily spatialized effect was achieved. The algorithm was essentially as follows: speech was processed to create early reflections and these reflections, as well as the direct sound, were individually passed through HRTF filters to make them seem like they were coming from the proper direction. This was then convolved with exponentially decaying Gaussian noise to model late reflections, delayed, and added to the previous signal.

As the experiments indicate, the performance of the spatialization system was still imperfect: most distance judgements were closer than the designed distance, and front-back reversals were common, especially with the male speaker. This is consistent, however, with the results of other researchers in the field[2]. However, It was demonstrated that 3-D sound processing is quite possible, and it seems that very good simulation may be soon achieved.

It also seems that this speech processing does help parallel speech recognition for humans. These preliminary experiments seem to indicate that spatialized, separated, simultaneous speakers are easier to comprehend concurrently. The performance of people listening to dichotic (one speaker in

each ear) speech for parallel understanding was better than that of two speakers mixed, directly in front, but still not as good as separated, spatialized speakers.

Research has demonstrated [4, 19] that dichotic listening is a superior method for serial speech understanding, but this does not involve comfort issues. Subjects tended to find the dichotic speech somewhat "overwhelming" due to having both speakers so close (inside the head). This may advocate the use of spatial processing for serial understanding as well.

There are several possible areas of further research in the areas covered by this paper. Quite a bit of work can still and is going on in improving 3-D spatialization. There are a number of paths on which this beginning algorithm could be taken. For example, in the determination of early reverberation, the addition of three dimensional image modeling would probably help in externalization, as would using more images (all the second order images, perhaps some third order). On the other hand, a ray tracing algorithm may give superior results, notwithstanding the higher processing time.

To improve late reverberation response, it may be helpful to use the binaural impulse response of an actual room. This could also solve the problem of how to implement the early reflections. If the algorithm were to be used in a real-time fashion however, the IIR method of comb filters and allpass filters may need to be implemented, despite the added complexity.

As stated previously, the experiments performed here were largely preliminary and to effectively test the success of parallel speech understanding, some more extensive tests should be performed. Some other areas of further research for enhancing parallel comprehension include the following. Different locations for the separated speakers (different azimuthal angles and perhaps different elevations as well) could be attempted. The speech segments used were played at a relatively low volume, so different volumes could be attempted. Different speakers (both male, one male one female, etc.) could be used to see which are the easiest to understand. The two speakers could also be placed at different distances away, which initially seems like it would hinder parallel understanding but may allow the listener to better separate the speakers.

These extensions all fall under the category of psychoacoustic research and the aid of psychology researchers could perhaps be enlisted.

All of this research will hopefully lead to a better understanding of how humans process speech, and how we can process signals to make it easier for us to perform this processing.

Appendix A

Example Shell Script

```
#!/bin/sh

# Early reflection processing

CopyAudio -D integer16 -g 0.036000 -l -648:93185 input/fspeech1.au
          temp/delat035.au
FiltAudio -f cofhold/left035.cof temp/delat035.au temp/mv0lat035.au
FiltAudio -f cofhold/righ035.cof temp/delat035.au temp/mv0rat035.au

CopyAudio -D integer16 -g 0.013846 -l -1045:93185 input/fspeech1.au
          temp/delat300.au
FiltAudio -f cofhold/left300.cof temp/delat300.au temp/mv1lat300.au
FiltAudio -f cofhold/righ300.cof temp/delat300.au temp/mv1rat300.au

CopyAudio -D integer16 -g 0.031034 -l -698:93185 input/fspeech1.au
          temp/delat020.au
FiltAudio -f cofhold/left020.cof temp/delat020.au temp/mv2lat020.au
FiltAudio -f cofhold/righ020.cof temp/delat020.au temp/mv2rat020.au

CopyAudio -D integer16 -g 0.031034 -l -698:93185 input/fspeech1.au
          temp/delat160.au
FiltAudio -f cofhold/left160.cof temp/delat160.au temp/mv3lat160.au
FiltAudio -f cofhold/righ160.cof temp/delat160.au temp/mv3rat160.au
```

```
CopyAudio -D integer16 -g 0.023824 -l -756:93185 input/fspeech1.au
          temp/delat030.au
FiltAudio -f cofhold/left030.cof temp/delat030.au temp/mv4lat030.au
FiltAudio -f cofhold/righ030.cof temp/delat030.au temp/mv4rat030.au

CopyAudio -D integer16 -g 0.023824 -l -756:93185 input/fspeech1.au
          temp/delat150.au
FiltAudio -f cofhold/left150.cof temp/delat150.au temp/mv5lat150.au
FiltAudio -f cofhold/righ150.cof temp/delat150.au temp/mv5rat150.au

CopyAudio -D integer16 -g 0.010946 -l -1115:93185 input/fspeech1.au
          temp/delat235.au
FiltAudio -f cofhold/left235.cof temp/delat235.au temp/mv6lat235.au
FiltAudio -f cofhold/righ235.cof temp/delat235.au temp/mv6rat235.au

CopyAudio -D integer16 -g 0.010946 -l -1115:93185 input/fspeech1.au
          temp/delat305.au
FiltAudio -f cofhold/left305.cof temp/delat305.au temp/mv7lat305.au
FiltAudio -f cofhold/righ305.cof temp/delat305.au temp/mv7rat305.au

# Direct sound

CopyAudio -D integer16 -g 0.050000 -l -580:93185 input/fspeech1.au
          temp/delat025.au
FiltAudio -f cofhold/left025.cof temp/delat025.au temp/mv8lat025.au
FiltAudio -f cofhold/righ025.cof temp/delat025.au temp/mv8rat025.au

# Combining direct sound and all the early reflections

CopyAudio -cA "A + B + C + D + E + F + G + H + I" temp/mv0lat035.au
          temp/mv1lat300.au temp/mv2lat020.au temp/mv3lat160.au
          temp/mv4lat030.au temp/mv5lat150.au temp/mv6lat235.au
          temp/mv7lat305.au temp/mv8lat025.au temp/mvleft.au
CopyAudio -cA "A + B + C + D + E + F + G + H + I" temp/mv0rat035.au
          temp/mv1rat300.au temp/mv2rat020.au temp/mv3rat160.au
          temp/mv4rat030.au temp/mv5rat150.au temp/mv6rat235.au
          temp/mv7rat305.au temp/mv8rat025.au temp/mvright.au
```

```
# Convolving the combined signals with noise

FiltAudio -f cofhold/noise1b.cof temp/mvleft.au temp/leftnoise.au
FiltAudio -f cofhold/noise2b.cof temp/mvright.au temp/rightnoise.au

CopyAudio -l -1115:93185 -g 0.010946 temp/rightnoise.au
          temp/delayrnoise.au
CopyAudio -l -1115:93185 -g 0.010946 temp/leftnoise.au
          temp/delaylnoise.au

# Combining the delayed late reverberation with the original signal

CopyAudio -cA "A + B" temp/mvright.au temp/delayrnoise.au
          temp/rightnohd.au
CopyAudio -cA "A + B" temp/mvleft.au temp/delaylnoise.au
          temp/leftnohd.au

# Filtering with inverse headphone responses

FiltAudio -f cofhold/leftsenn.cof temp/leftnohd.au temp/mvlefthd.au
FiltAudio -f cofhold/righsenn.cof temp/rightnohd.au temp/mvrighthd.au

# Combining left and right signals

CopyAudio temp/mvrighthd.au temp/mvlefthd.au output/final.au

# Clean up

rm -f temp/*
```

Bibliography

[1] D. Begault. Challenges to the successful implementation of 3-D sound. *Journal of the Audio Engineering Society*, 39(11):864–870, 1991.

[2] D. Begault. Perceptual effects of synthetic reverberation on three-dimensional audio systems. *Journal of the Audio Engineering Society*, 40:895–904, 1992.

[3] D. Begault. *3-D Sound for Virtual Reality and Multimedia*. AP Professional, 1994.

[4] E. Cherry. Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, 25(5):975–979, 1953.

[5] J. Egan, E. Carterette, and E. Thwing. Some factors affecting multi-channel listening. *Journal of the Acoustical Society of America*, 26(5):774–782, 1954.

[6] B. Gardner and K. Martin. HRTF measurements of a KEMAR dummy-head microphone. MIT Media Lab Perceptual Computing #280, Massachusetts Institute of Technology, May 1994. http://sound.media.mit.edu/KEMAR.html.

[7] W. Gardner. The virtual acoustic room. Master's thesis, Massachusetts Institute of Technology, 1992.

[8] M. Halikia. *The Perceptual Segregation of Simultaneous Sounds*. PhD thesis, McGill University, 1985.

[9] P. Kabal. Audio file i/o routines. ftp.tsp.ee.mcgill.ca, October 1996.

[10] O. Mitchell, C. Ross, and G. Yates. Signal processing for a cocktail party effect. *Journal of the Acoustical Society of America*, 50(2):656–660, 1971.

[11] J. Moorer. About this reverberation business. *Computer Music Journal*, 3(2):13–18, 1979.

[12] D. O'Shaughnessy. *Speech Communication.* Addison-Wesley, 1987.

[13] E. Poulton. Two-channel listening. *Journal of Experimental Psychology*, 46:91–96, 1953.

[14] J. Proakis and D. Manolakis. *Digital Signal Processing.* Prentice Hall, 1996.

[15] B. Rakerd and W. Hartmann. Localization of sound in rooms, II: The effects of a single reflecting surface. *Journal of the Acoustical Society of America*, 78(2):524–533, 1985.

[16] N. Sakamoto, T. Gotoh, and Y. Kimura. On 'Out-of-Head Localization' in headphone listening. *Journal of the Audio Engineering Society*, 24:710–716, 1976.

[17] M. Schroeder and B. Logan. 'Colorless' artificial reverberation. *Journal of the Audio Engineering Society*, 9:192–197, 1961.

[18] K. Steiglitz. *A DSP Primer with Applications to Digital Audio and Computer Music.* Addison-Wesley, 1996.

[19] A. Treisman and S. Fearnley. Can simultaneous speech stimuli be classified in parallel? *Perception & Psychophysics*, 10(1):1–7, 1971.