UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING AND THE ENVIRONMENT

INSTITUTE OF SOUND AND VIBRATION RESEARCH

Size Discrimination of Transient Signals

by

Niamh O'Meara

A thesis submitted in partial fulfilment for the degree of

Doctor of Philosophy

September 2012

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND THE ENVIRONMENT

INSTITUTE OF SOUND AND VIBRATION RESEARCH

Doctor of Philosophy

Size Discrimination of Transient Signals

By Niamh O'Meara

The importance of spectral cues in size discrimination of transient signals was investigated, and a model for this ability, tAIM, was created based on the biological principles of human hearing. A psychophysics experiment involving 40 participants found that the most important cue for size discrimination of transient signals, created by striking different sizes of polystyrene spheres, was similar to that of speakers listening to vowels – the relative positions of the resonances between comparison signals. It was found possible to scale the sphere signals in order to confuse listeners into believing the signal source was a different size, but two methods of scaling signals in order to sound the same size as another proved inconclusive, suggesting the possibility that transient signals cannot be scaled in a linear fashion as has been shown possible for vowels. Filtering the signals in a number of different ways found that the most important cue in size discrimination of transient signals is the difference between the most prominent resonances available in the spectra of the comparison signals. A model of the auditory system using the dynamic compressive Gammachirp filterbank, and based on the well-known AIM, was created to produce auditory images of transient signals that could be normalised for size. Transient-AIM, or tAIM used the Mellin transform to produce images that showed size normalisation was possible due to the spectral envelope similarities across the sizes of the spheres. tAIM was extended to carry out size discrimination of the spheres using the information contained within the Mellin images. There was a systematic association between Mellin phase and size of objects of various shapes, which suggests that tAIM is able to infer object size from sound recordings of objects being struck.

Table of Contents

Author's Declaration

I, Niamh O'Meara, declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

SIZE DISCRIMINATION OF TRANSIENT SIGNALS

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Either none of this work has been published before submission, or parts of this work have been published as:

   Fox, P.D., S. Bleeck, N. O'Meara, P.R. White, T.G. Leighton & V.F. Humphrey (2008). "Acoustic Network and Concepts". ISVR Contract Report No. 08/01

   Bleeck, S., Fox, P.D., White, P.R. and O'Meara, N. (2008). "Auditory models and nonlinear filter banks in underwater auralization". 2nd ASA-EAA Joint Conference, Acoustics '08, Paris, France.

   O'Meara, N. and S. Bleeck (2009). "The role of frequency in object-size perception". British Society of Audiology Short Papers Meeting on Experimental Studies of Hearing and Deafness, Southampton, UK.

Signed: ……

Date: ……

Acknowledgments

I would like to take this opportunity to extend my gratitude to all the staff of the ISVR and the university, my friends and loved ones for helping me to complete this work. Being part of the ISVR is like being part of a big family; it has been a welcoming, friendly and helpful place to work over the past few years, and I will miss it. My thanks to everyone who has been a part of my time here.

First and foremost, I would like to thank my supervisor, Dr. Stefan Bleeck, for his guidance and support through what has been a long and difficult but ultimately successful journey. I will never be able to convey the extent of my gratitude for his patience and encouragement, and many, many discussions on what road to take next. His bottomless pit of ideas and enthusiasm for his research has been a major source of my motivation. I would like to thank my second supervisor, Prof. Paul White, for all the technical support he has given me and his patience with me when it seemed like I would never finish! Thank you to Dr. Paul Fox for supplying the simulated cylinder signals used for analysis in this document, and to DSTL for funding the initial part of this project. I would also like to thank the faculty staff and the individual secretaries of the ISVR who have been so friendly and helpful to me over the years, and also the participants of the psychophysics experiment for without them much of this work would not have been possible.

My office-mates over the years, especially Delphine Nourzad and Katie Saunders, have been a continued source of fun and conversation, from the lunch breaks and tea and coffee breaks to the little words of encouragement to help me get this far. I would like to thank my proof-readers Helen Gruber and Frances Williams for their time and effort in helping me to improve the structure of my writing. Thanks also to Chris Powles and Emery Ku for lots of initial Matlab tuition, and to George Perakis for the enlightening conversations over coffee that led me to a number of light-bulb moments.

I want to give special thanks to Yasser Ibrahim, for he has dealt with the brunt of my PhD stress, and yet always managed to make me smile; I only hope I can return the favour. And finally I would like to thank my family, especially my Mam and Dad, who have never failed to encourage me, and have been tirelessly supportive of my time spent on this project. I will be eternally grateful for all they have done for me, for without their support, wisdom, kind words and prayers I would never have come this far.

List of Abbreviations

2AFC – Two alternative forced choice
2DAT – Two dimensional adaptive thresholding, an option in *aim-mat*
λ – Wavelength
ρ – Density
AIM – Auditory Image Model (Patterson et al, 1995)
aim-mat – A version of AIM implemented in MATLAB
ANOVA – Analysis of variance
b/a – ratio of inner to outer cylinder wall diameter
BM – Basilar membrane
BMM – Basilar membrane motion, a module of *aim-mat*
BSF – Band stop filter
c/o – Cut-off frequency
c_s – Speed of sound in a solid
dB SPL – Decibel Sound pressure level
dcGC – Dynamic compressive gammachirp
E – Young's modulus
ERB – Equivalent rectangular bandwidth
f – Frequency
F0 – Wideband low frequency energy in a transient signal spectrum
F1, F2 – First, second resonant peak in a spectrum
FFT – Fast Fourier transform
FM – Frequency modulated
G – Shear Modulus
GUI – Graphical user interface
HP-AF – High-pass asymmetric filter
HPF – High pass filter
IPA – International Phonetic Alphabet
JND – Just noticeable difference
L (Large) – Sphere size, 100 mm diameter
LPF – Low pass filter
µs – Microseconds
mm – Millimetres
M (Medium) – Sphere size, 90 mm diameter
MI – Mellin image
NAP – Neural activity pattern, a module of *aim-mat*
nfft – Number of FFT bins
NoF1 – F1 was filtered out
NoF1F2 – F1 and F2 were both filtered out
Pa – Pascal
PSD – Power spectral density
PSR – Playback sample rate
Revcor – Reverse correlation, a technique applied in auditory filter shape research.
RI – Reliability index

RMS – Root mean square
Roex – Rounded exponential, an early proposal of the auditory filter shape
S (Small) – Sphere size, 80 mm diameter
SAI – Stabilised auditory image
SCF – Spectral centroid frequency
SSI – Size-shape image
STI – Strobed temporal integration, a module of *aim-mat*
Stop-F1 – F1 was filtered out with a BSF
T-0 – Time Zero: a specific extraction from an auditory image
tAIM – Transient Auditory Image Model
TSP – Time separation pitch
UnF – Unfiltered
V – Poisson ratio
VTL – Vocal tract length
WithF1 – F1 was not filtered out
XL (X.Large) – Sphere size, 120 mm diameter
XL was L – Signal that now has the F1 of the XL but was originally the L sphere.
XS (X.Small) – Sphere size, 70 mm diameter

For my parents…

… go raibh míle maith agaibh.

1. Introduction

Humans and mammals are excellent at understanding their sound environment. They can identify shapes, locations, and sizes of objects from the sounds they make or echoes they reflect. Auditory perception is a well-researched area due to the many capabilities of the mammalian auditory system, and how it has evolved to be so efficient. It is capable of taking in massive amounts of sound information at any one time and separate sound sources, identify them and their locations, and ignore irrelevant information. The literature contains many studies on the perception of impact sounds, from sound and object source identification to categorical qualities about the source itself, such as size, and some of these will be discussed in chapter 2. There seems to be, however, a gap in the literature on how size discrimination is possible and very little documented research on the specific cues involved in the size discrimination of impact, or transient, sounds. In terms of periodic sounds such as the human voice, auditory models have been shown to normalise for size by averaging periodic sounds in order to extract information between resonances. This follows the theory that it is in the relationship between the resonances that size information is conveyed. Mammals and humans also have the ability of extracting size information from non-periodic sounds. In this thesis, an experiment is reported wherein the specific details in a sound that lead to size discrimination are identified and an auditory model is created in order to discriminate between transient sounds in the same way that humans can discriminate between them. To date, there has been little research conducted to prove exactly what features in sound are used in size discrimination, and there are no bio-inspired models that can extract this information from transient sounds. This project aims to fill in these research gaps.

The key results from this thesis are:

- A psychophysics experiment proves that spectral cues are crucial in a size discrimination task, and that the manipulation of these cues alone can fool the listener into believing the object has a different size.
- Size information has been shown not to be limited to one area of the frequency spectrum, and transient signals that have been heavily high-pass filtered can still be discriminated for size.
- An auditory model for the purpose of creating auditory images of transient signals, tAIM, has been created that can normalise for size of transient signals using the Mellin transform, and automatically discriminate for size using the cues found through psychophysics experiments.
- Evidence is proposed that questions the compressive quality of the dynamic compressive Gammachirp filterbank, and the effectiveness of the AIM at producing accurate images of transient signals.

This document will read as follows. Chapter 2 will discuss size information in sound starting from the evolutionary reasons behind why humans are capable of size discrimination, and how important size information is in communication for both humans and mammals such as dolphins and bats, all of which have the same basic auditory system. The extent to how this ability transfers into non-communication sounds will also be examined with regard to both humans and mammals, followed by a discussion of the literature regarding what cues might be most useful when the task is being performed.

A brief overview of the workings of the human auditory system will be presented in Chapter 3, beginning with the filtering at the outer ear, the mechanical workings of the middle ear that passes the sound from the air in the ear canal to the liquid of the cochlea, and finally explaining the spectral analysis that is performed in the inner ear by the basilar membrane that converts the pressure waveform into electrical impulses to be sent to the brain via the auditory nerve. A review of auditory modelling will follow, including a brief explanation of the different ways in which the shape of the auditory filter has been estimated, followed by a detailed description of Patterson's Auditory Image Model (AIM) that implements a dynamic

compressive Gammachirp filter, and how this model can be used to identify and then normalise for size information in periodic speech signals, i.e. vowels, using the Mellin transform. It is this model upon which a new model for transient sounds is based and created for the purposes of automatic size discrimination.

Size normalisation from AIM is based upon the ratios between any resonances in the vowels processed. This assumes that size information is solely contained within the spectral cues of a signal. To test whether or not this assumption is true for transient signals, Chapter 4 describes an experiment to test if transient signals can be scaled in a linear fashion. Recordings of five different sizes of polystyrene spheres were struck in order to create the stimuli. The experiment asked 5 participants to test if they could scale the signals by altering the playback sample-rate until the transient pitch of the sounds was the same. To differentiate it from periodic pitch, which is what people use to describe the order of musical notes or a type of voice, transient pitch is that which a person hears when a transient sounds higher or lower than another transient. Even though the definition offered by the American National Standards refers to the auditory sensation by which any group of sounds may be ordered on a scale from low to high, and depends on the frequency content of the signal (ANSI 1994), the most common understanding of pitch is that it is related to music and thus the term transient pitch is adopted here to distinguish the two. The results suggested that it might not be possible to scale transient signals in a linear fashion, unlike vowel signals.

Experiments 2, 3 and 4 are presented in Chapter 5, which explores in several ways how important transient pitch is in a size discrimination task, and where the most important frequencies of the transient pitch lie in the frequency spectrum. Forty people were recruited to participate in the experiments. The sphere recordings were processed and analysed to create a different scaling method based on the differences between the first resonances in the signals, F1. The signals were then scaled and filtered, and presented in a number of tests to participants to uncover information about if and/or how size is extracted from signals when transient pitch cues have been altered and when the spectrum contains less information. The results are presented followed by a discussion in Chapter 6 that

proposes the most important cue for size discrimination was found to be the differences between the two signals' most prominent resonance.

Chapter 7 introduces a new auditory model for transient signals, tAIM, based on the Auditory Image Model, and using the dynamic compressive gammachirp filterbank for spectral analysis of the signals. The stages of the tAIM are presented, how they differ from the AIM, and the reasons why these changes have been introduced are discussed. Control tones that emulate one period of a spoken vowel are used to illustrate how the tAIM can reproduce similar results to the AIM but for non-periodic signals. tAIM then processes more complex simulated signals and the sphere recordings in order to demonstrate how a Mellin image can be created for different types of transient signals. Simulated vowels, underwater scattered signals, and the sphere recordings from the experiments are used to test the tAIM for its ability to normalise for size with objects of different spectral content and clarity.

Chapter 8 shows how tAIM can be used for size discrimination. The size normalisation that is a result of the Mellin transform retains the size information within the phase of the Mellin image. tAIM analyses this Mellin phase information and compares it with the same from another signal in order to determine which signal comes from the bigger object. tAIM is shown to discriminate for sizes of different objects including simulated vowels from men, women and children, polystyrene spheres and other shapes, and simulated cylinder signals. There are also shown to be limitations to tAIM which brings into question the extent of the compressive quality of the dynamic compressive gammachirp filterbank. Conclusions from this work are presented in Chapter 9, and suggestions for further work are proposed.

2. Size Discrimination Abilities

2.1 Echolocation and ecological acoustics

When one considers echolocation, the first thought might be of bats and dolphins, since it is common knowledge that bats are blind, and dolphins are well known to make clicks and squeaks to navigate through unclear waters and detect prey. The echoes that are returned contain much of the information they need to carry out these tasks. This is active echolocation, a type of sonar used by some mammals to create an auditory description of their environment. The inner ear of these mammals is very similar to that of the human auditory system. The main difference lies in the range of frequencies and intensities they can hear, but the biological design is essentially the same. It may not be a surprise, then, that humans are also capable of echolocation, albeit a more passive type. Some active echolocation finds its way into everyday life; detecting the size of a room from the echoes of your footsteps, for example. For the most part, people use passive echolocation or passive listening to take information they hear and assess their surroundings and events. It is not required of them to see that a door has shut behind them, or that someone has broken a bottle in the next room, or that a flock of birds has suddenly taken flight from a tree. All of these sounds contain within them information that does not require the person to see the event in order to decipher it. In the case of both active and passive echolocation, the sounds that are heard contain information that is useful to the listener.

Echolocation, the ability to sense objects without the use of vision by hearing their echoes (Arias and Ramos, 1997), has been studied with regard to humans since the second half of the 20th century, with tests of blind and blind-folded

subjects' abilities to detect changes in distance, size and textures of objects (Kellogg, 1962); and the ability to subjects to detect the presence or absence of targets, binaurally and monaurally, and to localise the target in space (Rice, 1967). In Kellogg's study the subjects were instructed to create any sound they wished in order to create the echoes which included oral sounds, finger-snaps and clapping, but Rice limited his subjects to using just oral sounds including hisses and clicks. In each study, the case for human echolocation was positively enforced, though Rice concluded that the human ability might never rival that of bats. Other studies include the ability to navigate without the use of vision; Supa and colleagues (1994) instructed a group of blind or blindfolded subjects to detect a large board by walking towards it; and Gordon and Jarquin (2000) investigated whether it was possible for blindfolded participants to detect a large board while walking or stationary, and also to walk towards the position of the board after it had been removed. Studies like these further strengthen the case for human echolocation abilities, with the results being positive to board detection, though there were limitations on distance discrimination for the latter.

Echolocation is closely related to an area of study known as ecological acoustics, a term coined by Vanderveer (1979), and is described as sound source perception, or the perception of sound sources from their properties (Giordano and McAdams, 2006). Rice (1967) mused that the success of Kellogg's study of the discrimination of different materials was due to the intensity of the echoes returned as a direct measure of the absorption of the reflective material. There is extensive research into ecological acoustics and sound perception without the use of echolocation and the intended creation of a signal in order to create echoes. Giordano and McAdams (2006) investigated the discrimination of plates of four different types of materials and size, and found discrimination between categories of materials to be perfect, and within materials to be aided by the size of the object. The sound were created by striking the plates with a steel pendulum. For a short review of other studies on human sound perception, from identifying the gender and direction of travel of a person walking on a flight of stairs, to the sizes of bottles falling and either breaking or bouncing, and simply identifying recordings of sounds, please refer to Carello, Wagman and Turvey (2005). The research of ecological acoustics is well

documented, and it is under this heading that this study falls. The objective of this study is to investigate and model the size discrimination abilities of the human auditory system, and this ability exists due to the ability of humans to detect physical properties based on its acoustic response.

2.2 Size information and discrimination abilities

2.2.1 Size information in bio-acoustic communication

Aside from hearing sounds created by inanimate objects, or listening to echoes from surrounding objects, the hearing system plays a key role in communication. In this study, similarities are proposed between the mechanism for size discrimination of some inanimate objects and the size discrimination of humans from their speech. For this reason, a background to communication sounds is provided.

Pulse/resonance sounds are very common in bio-communication. In human speech generation, air pushed through the vocal folds produce a stream of glottal pulses and each glottal pulse produces a distinctive resonance, or impulse response, determined largely by the shape of the vocal tract at that moment. A diagram of the human vocal tract is shown in figure 1. The position of the tongue or lips can momentarily restrict or stop the airflow to produce consonants, but a free flow of air from the vocal folds through the vocal tract and out through the mouth produces the vowels. Humans hear the pulse rate as the pitch of the vowel and the resonance as the timbre, or vowel quality (/a/, /i/, /u/, …) and it is within these resonances that information regarding shape of the vocal tract and size of the speaker is contained. The important concepts are illustrated by the 'pulse/resonance' waveforms in figure 2 **Figure 2**- a man and a woman saying the vowel /a/.

Most mammals communicate with pulse/resonance sounds, including frogs, fish and insects, and mechanism used to produce the pulses and resonances scale up as the animal grows (Patterson et al., 2007). Human speech is also a type of pulse/resonance sound, where the repetitions of the pulses equate to the vibration

of the vocal folds, and the resonances give the information about the shape and size of the vocal tract. Echoes from underwater reflections are also pulse/resonance sounds. These are used by echolocating mammals to do such things as find food, understand their surroundings, and locate others of their kind (Au et al., 2000). Not unlike a glottal pulse which is created and then the resonances are formed due to the shape of the vocal tract, a click is created by the echolocating mammal and the resonances of the returned echo are as a result of the target. Thus it is reasonable to expect auditory processing is in general adapted for pulse/resonance communication and that scale normalisation is a central part of auditory analysis in most animals.

The Human Vocal Tract

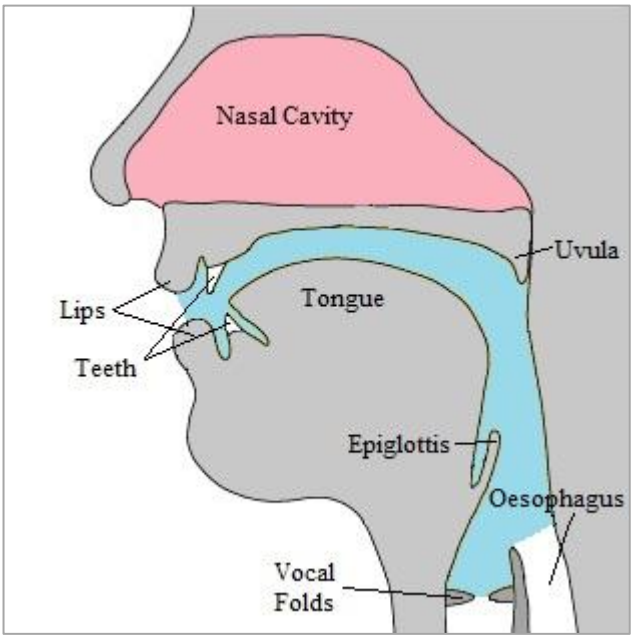

Figure 1: The human vocal tract, showing the lips and teeth, nasal cavity, tongue, and vocal folds. The space from the vocal folds to the lips (coloured in blue) is the vocal tract, and it is the length of this that varies with the size of the person or between genders. Image adapted from Gasser (2009).

The vocal tract length of a human is directly related to the size of the human (Fitch & Giedd, 1999). The vocal-tract consists of the air passages from above the

larynx where the vocal folds are situated, to the lips (Borden et al., 2003), as shown in the diagram in Figure 1. The length of the vocal tract varies depending on the size of the person; for example a woman's vocal-tract is typically larger than that of a child's, and a man's vocal-tract is normally even larger again. Despite these differences in vocal-tract size, the same speech uttered by all three humans can be recognised by a listener.

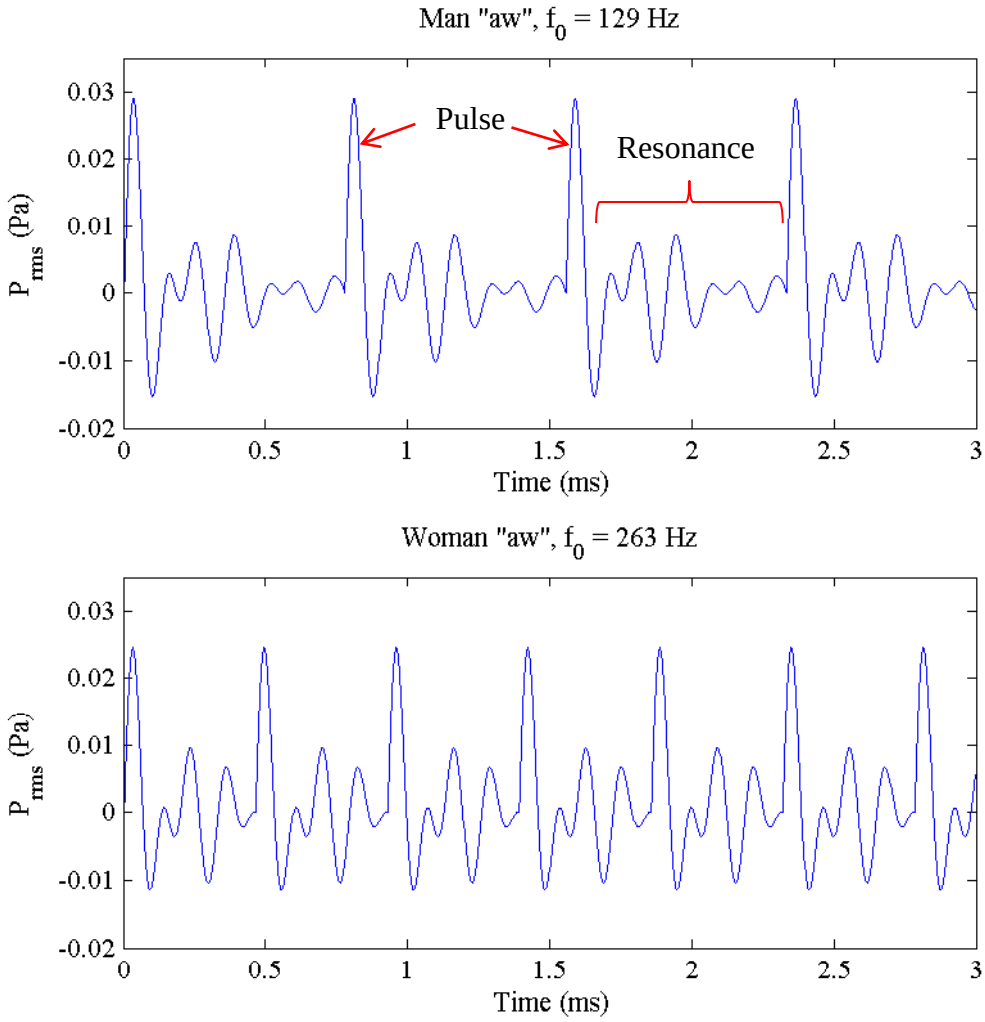

Figure 2: Natural pulse resonance communication sounds, the vowel /aw/ as uttered by a man (top panel) and woman (bottom panel). The pulse is followed by the resonances, which contain size and shape information.

The size of the vocal tract affects the fundamental frequency of the vowel being uttered as well as the frequencies of the resonances, but movement within the vocal tract changes the shape of the tract and therefore the relationship between the resonances of the tract. These movements include lowering and raising of the tongue, and rounding or opening of the lips. It is movements such as these that change the resonances and shape a sound into a specific vowel, and it is the relationship between these resonances that allow the listener to identify which vowel has been uttered (Irino and Patterson, 2002). These resonances can be higher or lower in frequency depending on the size of the person that has spoken, but the ratio between them remains comparatively similar.

It is this ability to ignore the actual frequencies of the resonances, but hear their relative positions with respect to each other that allow humans to effectively ignore *size*, and listen to *shape*. Considering this, it may be reasonable to assume that a type of normalisation procedure takes place in the auditory system that allows the listener to disregard the size information and keep the size-invariant properties of the speech (Patterson et al., 2007). Conversely, if the auditory system is able to isolate size information to ignore it, could it be possible that it is also able to isolate the size information to use it in a size discrimination task? And finally, is this limited to just speaker size discrimination of periodic sounds, or is it also used in the size discrimination of transient signals from inanimate objects? These ideas are discussed in due course with evidence from the literature.

2.2.2 Size discrimination abilities of mammals

Size discrimination refers to the ability to distinguish between objects which differ only in size. There have been many studies on the size discrimination abilities of the auditory systems of mammals. Here, one example of each is given to illustrate the extent to which bats and dolphins are able to use echolocation to perform size discrimination tasks. After this, the size discrimination abilities of the human auditory system will be discussed.

Being predominately nocturnal animals and being unable to rely on sight, bats use echolocation to navigate while flying, and in particular to find prey. They emit

trains of high-frequency sounds during flight that vary in duration from 0.3 to 300 ms, and in frequency range from 12 to 200 kHz, and are capable of hearing as much. The structure of the signal they emit can vary within and across species and location, with most using either frequency-modulated (FM) signals or a combination of constant frequency with FM components (Au, 1997). The extent of their ability to echolocate using these signals means they are capable of detecting small insects close to bushes from the returned echoes, which shows considerable skill (Grinnell, 1995). As part of an experiment in object classification by nectar-feeding bats, Von Helversen (2004) tested the ability of the bats to discriminate between hollow spheres of different diameters. After training it was shown that 90% of the time the bats could correctly identify the 36-mm diameter sphere which contained a reward of sugar-water from either an 18-, 30- or 50 mm diameter sphere which was unrewarded. Analysis of the echoes from the hollow spheres showed similar frequency spectra, but shifted up for the smaller spheres, similar to the effects of a human's size on their speech. The similarities between intensity and echo durations were found to be very similar in training pairs, and so the authors suggested the spectral composition of the echoes was the most important cue. They also noted, however, that some of the single echo spectra were too similar for this to be the only cue as the bats took longer to learn to discriminate between the smaller hemispheres. It was proposed that it was more likely that discrimination was based on a sequence of echoes rather than any single echo.

In a continuation of this experiment, von Helverson and colleagues (Simon et al., 2006) conducted a study where the hemispheres were placed on a rotating column in order to avoid position learning; one of the main cues suggested to have been used by the bats in the previous study. A larger number of hemispheres were used: five sets of seven hemispheres with diameters ranging from 12 mm to 117 mm. In general, most confusion in the task surrounded the hemispheres closest in size to those which contained the reward. Also, it was noted that for the larger hemispheres, a greater difference in absolute size difference was required for the bats to correctly discriminate between them. The very short length of the signals and the low intensity level differences compared to previous bat studies suggested that these cues were unlikely to have helped with discrimination. However, the

authors noted clear size-related differences in the frequency spectra for the hemispheres used, and proposed that for certain cases the spectral content was analysed, and in others a temporal comparison between the incident pulse from the bat and the echo was made.

Dolphins and other marine mammals use echolocation in their environment because often their vision is impaired by dark, turbulent or muddy waters. They echolocate by emitting clicks in order to navigate and find food (Au et al., 2000), but it is believed they also use passive listening to scan their sound environment due to the more efficient propagation of acoustic energy underwater (Au, 1997). Similar to bat hearing capabilities, dolphins can hear a very wide range of frequencies, considerably wider than human capabilities (see chapter 3 for an introduction to human hearing). There are two types of signals that dolphins and similar species emit. The first type is a broadband short-duration click, typically less than 100 μs in length, with a tendency for the centre frequency to increase with intensity of the signal (Au, 1997). The second type is considered to be of longer duration, greater than 125 μs, and is a narrowband signal, generally associated with smaller mammals, and the frequency of these signals does not tend to relate to intensity levels as these mammals have not been found to emit high-level signals as their larger counterparts (Au, 1997). Whichever category of signals used, these underwater mammals can demonstrate great abilities to use acoustics to understand their environment.

Au and Pawloski (1992) examined the ability of Atlantic bottlenose dolphins to discriminate between cylinders of different wall thickness. The dolphins were trained to echolocate and choose the standard target. Comparison cylinders had incremental differences in wall thickness of ±0.2, ±0.3, ±0.4, and ±0.8 mm from the standard cylinder, which had an outer diameter of 37.85 mm and a wall thickness of 6.35 mm. All the cylinders had the same outer diameter. Results showed that the dolphins could discriminate, with 75% success, between cylinders which were more than 0.23 mm thinner, and 0.27 mm thicker than the standard. This equates to a required size change of 3.6% and 4.3% respectively. They noted that the frequency spectra shifted with the cylinders of different wall thicknesses.

There are many such studies of dolphin target discrimination, including size as above, shape, target distance and material composition of structures. Another study on the thickness of metallic plates was conducted by Evans and Powell (1967) and found a performance level of 75% and above, with only two exceptions for the detection of 3 cm diameter 0.22 cm thick copper plate against nine comparison plates of varying thicknesses and materials. Nakahara and colleagues (1997) conducted a size discrimination ability study on the finless porpoise to test their ability to detect a standard cylinder of 15 mm diameter from comparison cylinders of 12-, 14-, 18-, and 20-mm diameters. All results showed performance levels above 87% for all comparison pairs, except for the 14 mm diameter target, where performance was down to 68%. In their discussion, the authors suggest either temporal or spectral cues as those used for discrimination, or a combination of both, due to the possibility that frequency differences within the target spectra were not large enough for detection (Nakahara et al., 1997).

These, and many other studies, go far to showcase the abilities of mammals to echolocate and differentiate between materials, target sizes, shapes and thicknesses, with many suggestions as to what are the exact cues used for discrimination. The echoes provided by the targets show spectral differences, but in some cases the signals are too short or the differences are too small to be detected by the auditory system of the mammals in question. Analysis of the signals can suggest correlation between cues and the results, but so far there is no definitive answer to the role of spectral cues, and mammals are unable to describe what they hear. This project aims to understand the role of frequency in size discrimination by humans, and so the next section will discuss relevant studies of human abilities in the area of ecological acoustics. In some cases the signals heard are recordings of echoes created by mammals, and in other cases the signals are due to impact on an object, but studies on humans using active echolocation are not mentioned here. The benefit of carrying out studies like these is that humans can describe what cues they listen to in order to discriminate for size.

2.2.3 Size discrimination abilities of humans

A good starting point for investigating the ability of humans to discriminate for size is to understand how it plays a part in speech. Section 2.1.1 mentioned that when listening to speech a person can hear the same speech produced by a male speaker and a child speaker despite large differences in the spectral content of the sound. It has been suggested that there may be a sort of size normalisation that occurs in the human auditory system that allows for this ability to ignore certain aspects of spectral information pertaining to size in order to extract shape information that allows the vowel to be identified. A study on the human-size perception from speech sounds was performed to study the auditory scaling mechanism (Smith et al., 2005). Size discrimination and vowel perception was tested, the former by changing the vocal tract length (VTL), and the latter using both normal vowels and vowels scaled to represent people with vocal tracts longer and shorter than usual. In both cases, the scaled signals were created using a vocoder called STRAIGHT (Kawahara et al., 1999). Their results showed that the discrimination of speaker size with different VTL was relatively easy for the subjects, with performance remaining above chance for a wide variety of different VTLs. Equally, performance in vowel recognition remained above chance. The authors suggest that the listener learns the spectral relationship between the formants in the vowels and that in speakers of different sizes the frequency spectra of the vowels shift up and down in frequency according to size, supporting the hypothesis that some form of scaling transform is applied to the sounds to remove any vowel recognition problems that may arise from dealing with speakers of very different sizes (Smith et al., 2005).

So it is known that humans can discriminate between speakers, but it is believed that humans can also normalise for speaker size. Putting aside any scaling mechanism, how do humans fare with sounds that radiate from inanimate objects? From studies of ecological acoustics and echolocation, size discrimination of inanimate objects is known to be possible, but the mechanism remains unclear. But is it known how? Consider a simple example such as a plastic sphere. The sphere is struck and the impulse response of the sphere may be heard. Subsequently, another

sphere of a different size but the same material is struck. The listener compares this impulse response with the first and a difference between the two sounds is heard. These impulse responses contain the resonances within each sphere that holds information about the material, shape and size of the spheres. One or more of these cues are used to help discriminate between the spheres. A discussion on the cues that could possibly be used in size discrimination will follow shortly, but first a review of human size discrimination of inanimate objects will be presented from selected examples in the literature. There a number of such studies, but the main points relevant to this project are presented below.

In an effort to understand whether it was possible to perceive precise sizes of objects from their sounds, Carello and colleagues (1998) conducted a psychoacoustic experiment on size perception. From behind an occlusion screen wooden rods of lengths ranging from 30 cm to 120 cm were dropped onto a linoleum floor and the participants were asked to indicate how long they thought each rod was. The percentage size difference decreased with increasing rod length from 33% to 12.5%. Although the perceived lengths were underestimated, the more important result was the fact that the listeners were able to order rods of different lengths into a meaningful scale without any standard of comparison (Carello et al., 1998). While this might not show effective size perception skills, it shows size discrimination abilities. A second experiment with a similar setup was employed, this time using rods ranging in length from 10 cm to 40 cm in 5 cm increments. The participants were again able to order the rods according to size, but they found length discrimination for the smaller rods much easier. This raises the question, is there a limit to how large or small an object has to be before we are no longer able to detect its size? Although this was a task for size perception, size discrimination did play its part. The acoustic structure of the sound sources, more specifically signal duration, amplitude and frequency, were analysed for their relationship to the perception results, but regression analysis did not show these to be as successful as using the actual length of the rods.

Due to the presentation method and the type of sound source used, little can be drawn from the results as to what cues the listeners used to aid their perception.

The presentation of the sound source to the listeners did not allow for the authors to control the intensity of the rods as they fell, so it could be assumed that loudness was a cue, along with the features of the bounce of each rod at it hit the floor, though this has many variables. In the second experiment the rods fell onto an elevated plywood surface. The sound emitted from the plywood itself along with the rods, in comparison to the rods falling onto a linoleum floor from the first experiment could have a large effect on perception, as was found to be the case for Grassi (2005) when he investigated the effect of the size of a plate on the perception of wooden balls. Nonetheless size discrimination abilities were evident to a precision of 12.5% of difference in length for the largest spheres.

In another size perception and discrimination task, Houben and colleagues (2004) constructed an experiment where the participants listened to wooden balls rolling over a wooden plate. They also tested the discrimination of the sources while rolling at different speeds, and the interaction between size and speed, but these are not of interest here. Recordings of the wooden balls of seven different diameters ranging from 22 mm to 83 mm were presented in pairs. The sound was caused by the wooden balls rolling across the plate at a mean speed of 0.75 m/s. For all pairs of wooden balls the participants performed significantly above chance at identifying the larger ball, except for those with the smallest diameter. The authors noted that this was due to the percentage size difference between these two wooden balls was the smallest of all the pairs, only 14%, and too small to be perceived by the participants. The size difference was smaller for the rods in the previously experiment by Carello and colleagues (1998), but the rods were comparatively larger than the small wooden balls used here. The authors carried out measures to remove any cues that were created during the recording process that could contain size information except for temporal and spectral cues. These included the noise of the table upon which the wooden plate was placed, the amplitude modulation produced by the rolling of less than perfectly spherical balls, the length of the rolling time and intensity cues. After analysis, they suggest that the differences in spectral cues play a larger part in the size discrimination abilities of their participants than temporal cues (Houben et al., 2004).

Since there are apparent similarities in human auditory system and that of the dolphin, a study to understand further what features are used in dolphin echolocation was carried out by DeLong and colleagues (2007) in which they presented dolphin echolocation sounds to human participants. Part of the study was a discrimination task involving hollow cylinders with varying wall thicknesses, the same used by Au and Pawloski (1992). The stimuli were recordings of returned echoes of a dolphin click of 7 μs long from the cylinders which were scaled down by time-stretching the signal by a factor of 167 so to place the frequencies in the human auditory range. This is because the frequency range used by dolphins for echolocation is much wider than human hearing (Au et al., 2000). Since this study was intended to understand dolphin echolocation, the echoes were grouped together in sets of six, and presented at a rate of 60 /s to the participants. The subjects were asked to indicate whether they heard a difference between a standard target and a comparison target, and report what echo features they heard that were different. Their performance showed correct identification of the standard cylinder for all comparisons apart from -0.2 and -0.3 mm. The participants reported their primary cue used was pitch, where they indicated the pitch of the cylinders increased as the thickness increased. The second was duration, where in some cases the participants noted shorter durations than the standard for cylinders with thicker walls. There was also the mention of time separation pitch (TSP) as a cue, which is the pitch perceived when pulses are repeated in close succession, and is related to the reciprocal of the space between each pulse (McClellan and Small, 1966). In this case, the TSP would be close to 60Hz, varying as a result of the length of each individual pulse itself. For the implications of dolphin echolocation, this could be a possible cue, but for human echolocation, TSP is not a relevant cue since the environmental sounds humans hear are not usually presented as strings of pulses.

So far, the literature has produced a large body of research documenting the echolocation abilities of mammals, including humans. Since it is already clear that humans are capable of discriminating for size, the purpose of this study is to understand more about size discrimination cues, and more notably spectral cues. The literature presented above has already shown discussion of spectral and temporal cues as being the most important. Temporal cues are more obvious with

objects that are very different in size and have been shown to be variable depending on the presentation method (Grassi, 2005; Houben et al., 2004; Carello et al., 1998). Intensity cues have already been addressed in some studies (Houben et al., 2004), but others deemed it necessary for perception. It would seem that to minimise the interference of cues other than spectral, a method of signal creation that eliminated temporal and level differences is required. If possible, this would require the listener to listen only to spectral cues, and the cues within spectral content that could allow the listener to discriminate for size.

2.3 What do we listen to in SD tasks?

Some of the cues that could be used for size discrimination have already been mentioned, but their roles in human hearing are unclear. An overview of the biological workings of the human auditory system is presented in Chapter 3; for now the limits of certain aspects of sound perception will be summarised in order to show how much of the cues can be extracted from a sound for effective size discrimination. The cues suggested above are mainly spectral and time cues, there is also mention of intensity from the study by Carello and colleagues (1998), and timbre in the case of the discrimination of object material carried out by DeLong and colleagues (2006). Timbre can be defined as the harmonic structure of one sound that causes it to be perceived as different to another sound that has the same fundamental frequency (Darwin, 2005), just how a plucked G-string on a violin sounds different to the G-string on a guitar. For this reason, timbre will be included under the heading of spectral cues. For all both spectral and time cues, but also intensity cues, there are limitations within human perception for each. These will be discussed within the confines of size discrimination of the signals presented in the literature above.

The frequency range of healthy human hearing extends from approximately 20 Hz to 20 kHz, but the upper limit of this range begins to decrease with age from the teenage years. Below the lower limit, frequencies can be felt more than heard, and sounds at the limits of this range need to be at a very high intensity in order to be sensed. The physical make-up of the hearing organ means that the frequency resolution, the ability to tell tones of different frequencies apart, decreases with

increasing frequency. So for pure tones at low frequencies below 1 kHz, differences of 1 Hz can be heard which is a 0.001 % change in frequency; for mid-frequencies of 4 kHz this increases to 0.004 % change; and for high frequencies of about 8 kHz tones would have to differ by about 0.008 % in order to be perceived as being different (Weir et al., 1977).

Frequency discrimination depends on the intensity of the signals. The range of pressures audible to humans is from 20 μPa to 20 Pa, more easily expressed on the logarithmic decibel scale as from 0 – 120 dB SPL. The upper limit broaches onto what is known as the Threshold of Pain, and prolonged periods of exposure to sounds of 100 dB SPL and above can do irreversible damage to a person's hearing. Speech at a comfortable level is in the range of 60 - 65 dB (Moore, 2003). In tasks pertaining to signal discrimination, in general the louder the tones the easier they are to differentiate, and the frequency differences mentioned in the previous paragraph were for tones at 40 dB SPL, when tones were 5dB the differences needed to be much larger to be heard. At the limits of the hearing range, tones need to be nearly 60 dB SPL louder than at most sensitive part of the hearing range, 2-5 kHz, although absolute sensitivity is maintained between the ranges of 100 Hz and 10 kHz (Gelfand, 2007). Moore (2003) summarises the research conducted on the minimum detectable intensity changes of which the ear is capable of discerning, and for wideband and narrowband noise, the detectable change in intensity is a constant fraction of the intensity of the signal itself, i.e. Weber's law. For pure tones, Riesz (1928) noted this not to be the same, and that for 1 kHz tones, the higher the intensity of the signal, the smaller the detectable intensity difference, up to 100 dB SPL (Viemeister and Bacon, 1988). For example, Reisz (1928) found that for a 1 kHz tone at 40 dB SPL, a difference of 0.3 dB can be detected.

The third cue mentioned, the temporal cue, refers to the duration of a signal. This affects the perception of both the frequencies and the intensity of the signal. The temporal processing ability of the human auditory system is extensive, but for signals under 10 ms it has the effect of spreading energy across the entire frequency spectrum (Wright, 1964). Wright noted that the longer the signal, the more tonal its quality becomes, for the shortest durations only a click is heard, and for slightly

longer durations some tonal qualities shine through, this is what he calls 'click pitch', similar to the term transient pitch that is referred to by this study. Once a signal is long enough for several periods to be heard, it is perceived as a tone and no difference is observed in discrimination by lengthening the signal further. Finally, thresholds for tone detection are also affected by duration. For a 1 kHz tone of 200 ms duration to still be detectable when its duration drops to 20 ms, the intensity needs to increase by 10 dB (Gelfand, 2007).

Different vowels are identified by the relationship between the resonances that are contained within each pulse (Peterson, 1952), and are dependent on the shape of the vocal tract; different sized vocal tracts affect the frequency values of these resonances in order to give information about the size of the speaker. A vowel is a continuous complex signal composed of a number of frequencies of different intensities known as formants. The frequencies with the largest intensities are the most integral to vowel and speaker-size identification (Borden et al., 2003). The transient signals used in the size perception studies above are also complex signals. The signals are made up a range of frequencies of different intensities, but are very short in duration. Therefore, as Wright (1964) noted, the energy dispersion in the frequency spectrum will be much wider than if it were of a longer duration. In all the studies above, spectral content was suggested as the clearest indicator of size after analysis of the frequency spectrum. The energy was spread across the full spectrum for those that were transient but spectral notches were present that shifted according to size. This was shown in the bat echoes in the study by Simon and colleagues (2006), and Carello (1998) mentioned that the centre of gravity of the rod signals was heavily related to physical length. Although signals which are not pulse-resonance, the rolling wooden balls used in Houben's study also exhibited spectral notches that shifted with size. The echoes used by DeLong and colleagues (2006) were just over 1 ms long, but their perception as single events in time was affected by the time-separation pitch that most likely aided the perception. Further, Carello (1998) noted that on the perception of rod length, the effect of amplitude was most likely the cause of discrimination ability.

It would appear that although there is plenty of suggestion as to the importance of spectral content as a cue for size discrimination, there needs to be more research to definitively show this. Considering the wide range of frequencies, intensities and temporal processing abilities of the auditory system, along with the resolutions of each of these, the following should have been enough for successful size discrimination: a single instance of one of the echoes from DeLong's study instead of a number of them in short succession which caused TSP cues; or normalised recordings of Carello's rods instead of live sounds which gave intensity cues; or one bounce of the wooden balls by Houben and colleagues instead of listening to it roll, which provided temporal cues from balls which were not perfect spheres. A simpler sound source and presentation method is required to address this, as well as careful attention paid to controlling temporal and intensity cues, and more analysis on the spectral cues is necessary. Grassi (2002) already attempted altering the recordings of the wooden balls by equating RMS, by either high-pass filtering with a cut-on of 5 kHz, and by low-pass filtering with a cut-off of 5 kHz. He found that equating the RMS of each signal had the greatest effect of reducing discrimination ability, but that filtering the signals also affected performance. However, this alteration of the spectral domain in this study does not effectively answer where in the spectrum the most important information lies. It is known that for speech, the first three formants contain the most information (Peterson, 1952) and these occur below 5 kHz. It has not been found if this is the same case for transient non-vocal sounds.

The mention of the relationship between the centre of gravity of the signal and the size of the rods used by Carello (1998) is of interest with regard to spectral cues. It is believed that the centre of gravity, or Spectral Centroid Frequency (SCF) of a musical tone is highly correlated with the timbre of the tone (Schubert et al., 2004), and is a measure where the centre of mass is in a frequency spectrum. The more energy there is in the higher frequencies of a signal, the brighter it is in timbre, and the higher the SCF value. SCF is calculated from the FFT of a signal using the following equation:

$$\frac{\sum_{n=1}^{N/2} f_n x_n}{\sum_{n=1}^{N/2} x_n}$$

Equation 1

where x_n is the magnitude of the bin in the FFT, and f_n is the centre frequency of that bin (Schubert et al., 2004). Its relevance to size perception by Carello suggests that for transient signals the timbre and so the SCF could be a possible size cue. If a pulse-resonance has notches in its frequency spectrum that shift according to size, the SCF will be affected by these in the same way. Multi-dimensional scaling techniques have allowed for the measurement of timbre on a scale of high to low, or soft to loud, and have shown that for periodic signals pitch and timbre are independent of each other (Plomp, 1970). A conclusive link between SCF and size discrimination of transient signals has yet to be determined.

So far, the cues that have been discussed here are spectral cues, including SCF, temporal cues, including TSP, and loudness/intensity cues. In the studies, there has been a lot of conjecture about the importance of spectral cues in size perception tasks, but in each case the spectral cues have been subjected to interference by other cues. The studies up to now have not isolated spectral cues in pulse-resonance signals. The sounds presented have been temporally-altered dolphin echoes repeated at a high rate which resulted in TSP interference; the rods dropped from a height to a flat surface were of different intensities; and the wooden balls of normalised loudness recorded rolling along a plate were not transient signals. In Chapters 4 and 5 experiments are described where the spectral cues in transient signals are analysed and which part of the spectrum that contains the most important information and how robust this spectral information may be is ascertained, as has been done for speech sounds. Chapter 3 presents an overview of the anatomy of the human ear and a brief history of models of the auditory system. One such model, AIM, analyses vowel sounds and is capable of normalising for size by using the fact that vowel formant frequencies retain their ratios in order to be distinguished, but the spectral envelope shifts with the size of the speaker. If the importance of the spectral envelope is proven for transient sounds, as it has been for speech, this study aims to create an auditory model of size normalisation and size discrimination for transient signals, inspired by AIM.

3. The Ear and Auditory Modelling

3.1 Anatomy of the Ear

The human auditory system is a remarkable piece of biology. The hearing organ, the cochlea, is about the same size as a pea and is capable of complex sound analysis, being able to extract information from sounds with wavelengths between 17 metres and just 1.7 cm long, and pressure levels from 20 μPa to 20 Pa. The auditory system includes the auditory periphery that provides cues for localisation, amplifies, and performs spectral analysis of a sound before it is converted into electrical pulses that are sent down the auditory nerve to the central auditory pathways of the brain.

The auditory periphery consists of three parts; the outer, middle and inner ear. The outer ear is composed of the pinna (the visible part of the ear) and the auditory canal, which are respectively used for helping with the localisation of sound sources and boosting the high frequencies that are especially important in speech perception: 2-5 kHz. The sound travels down the auditory canal and causes the tympanic membrane, or eardrum, at the end of the canal to vibrate. The other side of the eardrum is the air-filled cavity of the middle ear and it contains three bones of the ossicles: the malleus, incus and stapes. The malleus transfers the vibrations from the eardrum through the incus to the stapes which is attached to the oval window of the fluid-filled cochlea in the inner ear. The relative positioning of these bones acts as an impedance-matching device for the effective transmission of sound from the air into the fluid, and also minimises the amount of sound reflected back out of the auditory canal. A diagram of the outer, middle and inner ear is shown in Figure 3.

The inner ear contains the cochlea, the outside of which is a bony construct that resembles the shell of a snail. Its walls are rigid and it contains almost incompressible fluids. The movement of the stapes at the oval window causes a travelling wave to be sent through the fluid of the cochlea. The basilar membrane and Reissner's membrane divide the cochlea along its length into three chambers; scala vestibuli, scala tympani and scala media. As the travelling wave passes over the basilar membrane, its motion is translated into neural activity that is sent to the brain for us to perceive as sound. The shape and flexibility of the basilar membrane affects how it responds to different frequencies within a sound. Higher frequencies produce maximum displacement at the base where it is wide and flexible, and lower frequencies closer to the apex, where it is narrow and stiff. The wave travels all the way along the basilar membrane to find its resonant peak before reaching the apex. In this way, the basilar membrane acts as a filterbank, performing a spectral analysis on the incoming sound, because each frequency optimally excites a certain point.

The Auditory Periphery

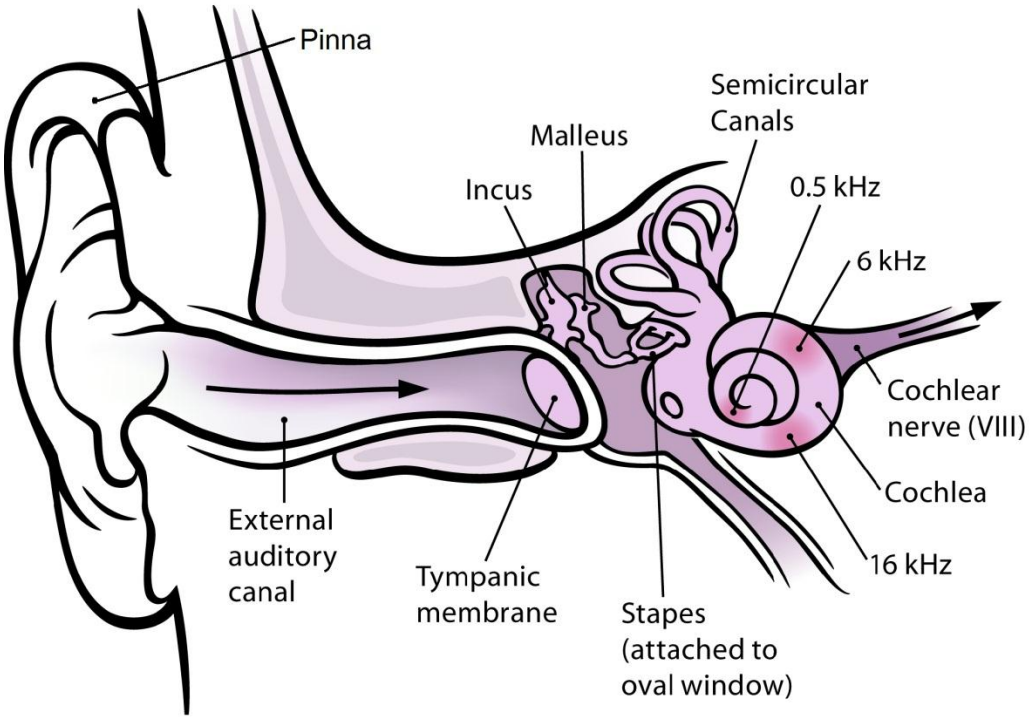

Figure 3: Schematic overview of the auditory periphery including the pinna, auditory canal, the bones of the ossicles and the cochlea (Chittka & Brockmann, 2005, open-source).

The organ of corti lies on the basilar membrane and its function is to finely tune the movement of the resonating point. It does this using 15,000-20,000 nerve receptors or hair cells, split into three or four rows of outer hair cells and one row of inner hair cells. Each hair cell has tiny hairs, or stereocilia that move back and forth when the point at which they are positioned on the basilar membrane is excited. The three rows of outer hair cells work as pre-amplifier and have an active influence on the mechanics of the cochlea, improving sensitivity and providing sharp tuning. The inner hair cells then translate the movement of the stereocilia into a chemical potential that is then translated into nerve action potentials: the movement of the inner hair cells in one direction opens ion channels and allows entry of ions into the hair cells and the excitation of the auditory nerve, and this allows the information to be sent to the auditory cortex in the brain. In short, the position of the incoming nerve pulses identifies the frequency of the sound, and the speed of the movement of the stereocilia provides amplitude information for that frequency, effectively performing a spectral analysis of a sound. For more detail please refer to Moore (2003). Figure 4 shows a cross-section of a mammalian cochlea.

The Human Cochlea

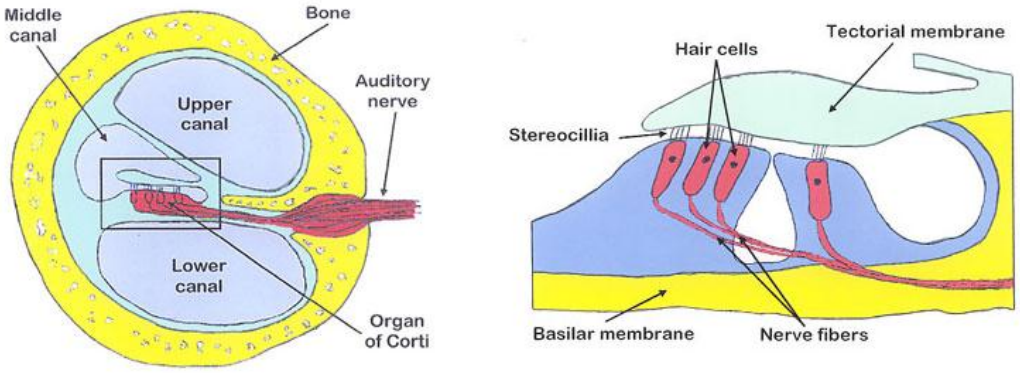

Figure 4: The left panel above shows a cross-section of the human cochlea, and the panel on the right shows a zoomed-in image of the stereocilia which move in response to the motion of the basilar membrane below. (Russell, R. 2008, reproduced with permission from author.)

3.2 The Auditory Filter

The ability of the human auditory system to process sound is complex. For example, in a restaurant full of people, humans can extract the information they want from all the sounds present and not get confused. We can listen to the music playing in one corner of the room, or have a conversation with the person at the other end of the table – this is known as the cocktail party effect, the ability to focus attention on one sound amidst lots of other acoustic information. We can hear that someone has dropped a plate even though we can't see it happening. We can pick out one instrument playing along with several others. We can tell the size of a room just by walking into it by the echoes from our footsteps (McGrath et al., 1999). We can understand the same words from a child and from a man, but also know that one is from a man and the other from a child (Irino and Patterson, 2002). For more on sound perception abilities and a more detailed account of extent to which humans can analyse their acoustic environment please refer to Bergman's Auditory Scene Analysis (1994) and Lufti's Human Sound Source Identification (1998). We may take these abilities for granted, but over the last century scientists have been working hard designing and carrying out experiments to understand how refined these abilities are, and more recently models of the auditory system have been proposed in order to further explain how the complex sound environments that surround us are perceived, to predict results, and to create machine hearing systems (Lyon, 2010). In this, the age of technology, speech recognition systems are constantly being created and improved upon, and given that the human auditory system is the most efficient speech recognition system there is, basing a machine hearing system on human hearing is an obvious choice (Lyon, 2010). The added benefit to this is the further understanding of human sound perception, and more relevant to this study, object size discrimination.

3.2.1 Auditory Filter Shape

There have been a number of attempts at creating models of the human auditory system which will be briefly introduced here, one of which will be discussed in detail, Patterson's Auditory Image Model (AIM). While AIM was created specifically for speech sounds, AIM is the inspiration behind this study where a new

model based on the ideas used in AIM will be created for the analysis of transient sounds. These auditory models are mostly based on the theories of how the basilar membrane performs spectral analysis on sound. As early as 1940, Fletcher suggested spectral analysis was carried out by a number of filters whose width depended on frequency and were spaced evenly along the basilar membrane (BM). Section 3.1 discussed how the shape and flexibility of the BM determines how spectral analysis is performed, with high frequencies resonating on the BM at the base where it is narrow and stiff, and low frequencies resonating near the apex where it is wide and flexible. Fletcher suggested there are a number of auditory filters which overlap and proposed that the filters have a simple square-topped shape and they have bandwidths that depend on frequency. The bandwidths have a critical limit, 'critical bandwidth'; where when two tones are presented, the ability of the auditory system to resolve the two tones will only occur if the tones are more than one critical bandwidth apart. This also refers to auditory masking where one tone will contribute to the masking of the other only if they are within the same critical bandwidth. (Moore, 2003)

Notch-noise masking methods (Schafer et al., 1950; Patterson 1974, 1976) and rippled noise methods (Houtgast, 1977) further researched the shape of the auditory filter, and how the shape is affected by frequency and intensity of a tone. The notch-noise masking method involves finding the threshold of a sinusoidal tone as a function of the width of a spectral notch in a noise masker. The rippled-noise method asks listeners to hear a pulsed sinusoid from a background of rippled noise, which was a masker that had a power spectrum shaped by a sinusoidal function. From these experiments, it is now known that the width of the auditory filter depends on its position on the basilar membrane, that the width of this band influences frequency selectivity, with the narrower the bandwidth leading to a higher frequency resolution. Each auditory filter has a 3 dB critical bandwidth, and the presence of a masker within that bandwidth negatively affects the ability for a tone at the centre frequency of that filter to be sensed. Patterson's study (1976) also shows that below the 3dB pass band, the skirts of the auditory filter became shallower. If a sinusoid is presented in noise, only the frequencies of the noise which occur within the critical band cause the sinusoid to be masked. Notch-noise

masking experiments where the tone was not centred in the noise have concluded that the filter shapes are asymmetric (Patterson and Nimmo-Smith, 1980), so that the slope of the low frequency side of the auditory filter is shallower. Lufti and Patterson (1984) also investigated the asymmetry of the filter with respect to stimulus intensity, and found non-linearities. The auditory filter is roughly symmetric at moderate levels, but as stimulus level increases, the skirt of the auditory filter broadens below its centre frequency, and the skirt sharpens a little above its centre frequency (Lufti and Patterson, 1984; Patterson and Moore, 1986)

3.2.2 Impulse response of the Auditory Filter

De Boer and Kyuper (1968) introduced a reverse-correlation (revcor) technique to characterise the firings of the auditory nerve of a cat with the input. They cross-correlated the output of the nerve with the input, which was white noise, to find the impulse response of the filter. From there an estimation of the filter shape was found, and has since become known as the gammatone function (Johannesma, 1972). Patterson,Nimmo-Smith et al's (1987) study fitted exponential curves to previous masking data, and proposed the shape of the auditory filter to be a rounded-exponential (Roex) shape – two negative exponential curves placed back-to-back with parameters to round the top of the curve and make the tails shallower. Further investigations using more controlled notch-noise methods by Glasberg and Moore (1990) assumed the Roex auditory filter shape and reviewed the Equivalent Rectangular Bandwidth, ERB, approximations. The ERB scale was first proposed by Moore and Glasberg (1983) using the Roex auditory filter shape, of which the magnitude response for each side of the filter is:

$$W(g) = (1 - r)(1 + pg)\exp(-pg) + r,$$

Equation 2

where g is the deviation from the centre frequency divided by the centre frequency; p is a free parameter that determines the shape of the slopes of the filter; and r is a free parameter that flattens the filter at frequencies far from the centre frequency and thereby places a dynamic range limitation on the filter. p values were different for the upper and lower parts of the filter skirts, according to the findings of the

shallower slope below the 3 dB pass band. The ERB equation derived using data collected through notch-noise masking and the assumption that the filter shape was the Roex as described above is:

$$ERB = 24.7(4.37F + 1),$$

Equation 3

where *F* is frequency in kHz. This equation is based on Greenwood's suggestion (1961) that each critical band takes up the same amount of distance along the basilar membrane.

3.2.3 Gammatone Filter

The modelling the basilar membrane and creation of filterbanks based on the impulse response of the basilar membrane has been one focus of research. After being introduced by de Boer (1977) and Johannesma (1972) as a way of modelling the revcor data, Schofield (1985) showed that the gammatone function could also explain Patterson's(1976) masking data. Patterson, Nimmo-Smith et al (1987) expanded upon Schofield's work and demonstrated that the impulse response of the gammatone function has a similar magnitude response to the Roex function. The equation for the gammatone filter is:

$$g_t(t) = at^{n-1}e^{-2\pi bERB(f_c)t}\cos(2\pi f_c t + \emptyset)$$

Equation 4

where *n* is the order of the filter and determines the slope of the skirts, *b* is the bandwidth parameter, and f_c is the centre frequency of the filter. The impulse response of the filter is shown in figure 5. The name gammatone was adopted by de Boer and Jongh (1977) due to the expression of the impulse response consisting of a gamma distribution and a cosine term, and also drawing from the fact that the revcor data showed the waveform of the impulse response of the filter looked like a cosine wave with the frequency of the centre frequency of the filter shaped by a gamma envelope. The gammatone function used along with the ERB scale created

simulations of the basilar membrane and thus perform spectral analysis on sounds based on the auditory system.

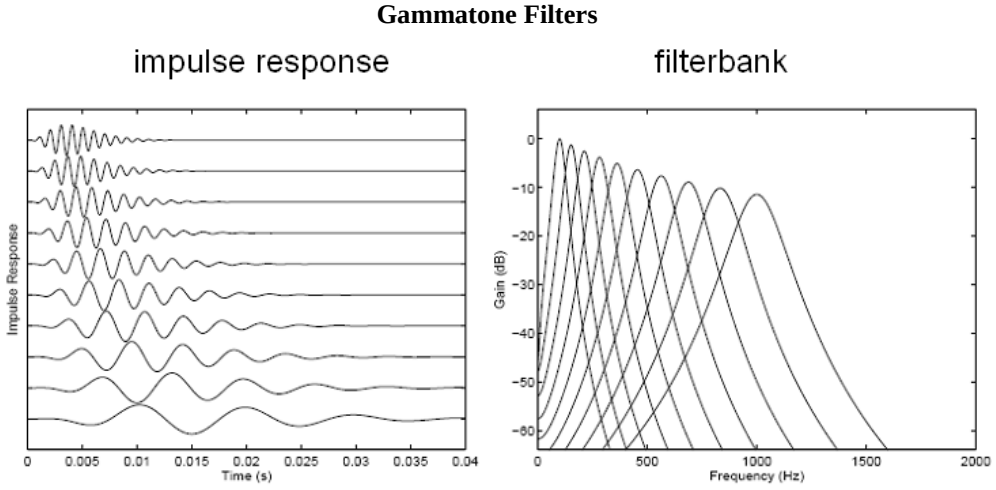

Figure 5: Impulse response and magnitude response of a number of gammatone filters

The gammatone filter is a linear time-domain representation of the auditory filter which is symmetric in the frequency domain. However, research by de Boer and Nutall (1997) found that the instantaneous frequency of an auditory filter was lower at first and then increased in frequency like a chirp. This, and the fact that intensity also affected the shape of the filter as found by Lufti and Patterson (1984) showed that gammatone function was not an appropriate method of modelling the non-linearities, more specifically, the level-dependent asymmetries of the auditory filter. The gammatone filter assumes the centre frequency does not change with time, and there is no explicit parameter for amplitude.

3.2.4 Gammachirp Filter

Irino (1995) proposed a new filter that was built upon the gammatone but had an added parameter that could demonstrate the increase in frequency with time, the

chirp, of the impulse response (Irino, 1995). This was called the gammachirp, and its impulse response is:

$$g_c(t) = at^{n-1}e^{-2\pi b\mathrm{ERB}(f_r)t}\cos(2\pi f_r t + c\ln t + \phi), \text{ where } t>0.$$

Equation 5

The inclusion of the term *c ln t* allows the modelling of the filter in the time domain and is the only difference between this and the impulse response of the gammatone filter. Instead of a cosine carrier, the gammachirp filter has a monotonically frequency-modulated carrier (a chirp) with an envelope that is a gamma function, and so models increase in frequency in a filter with time. The impulse response is shown in figure 6. When c=0 this reduces to the equation for the gammatone filter.

The gammachirp filter solves the problems of asymmetries in time and frequency, but not the level dependencies of the filter shape. Recently, Irino and Patterson extended the gammachirp filterbank to include a high-pass asymmetric function which provides a more realistic auditory filterbank (Irino and Patterson, 2006). The resulting dynamic compressive gammachirp filter (dcGC) was fitted to a large body of simultaneous masking data obtained psychophysically. The dcGC consists of a passive gammachirp filter and an asymmetric function which shifts in frequency with stimulus level as dictated by data on the compression of BM motion.

This dynamic, compressive gammachirp filter has many advantages over the Roex and Gammatone filters (Unoki and Irino, 2006). First of all, the Roex filter, on which the gammatone is based, does not accurately represent the asymmetries of the auditory filter. It can only filter stationary sounds in the spectral domain and can therefore not be used to filter a complex sounds, or even a waveform. There are also no time-domain versions of the Roex filter. Altering parameters to account for these limitations has proven to be unsuccessful, as the Roex becomes physically implausible.

The dynamic compressive gammachirp filter has a well-defined impulse response, it addresses the change of frequency in the filter with time, and it has

variable asymmetry to explain the level-dependent asymmetry in the auditory filter (Irino and Patterson, 2006). Unoki and colleagues mention that one of the major advantages of this filter is its robustness in speech recognition, where the instantaneous compression acts to compress the glottal pulses while the frequency resolution is maintained in order for clear analysis of vocal-tract resonances.

Gammachirp filterbank

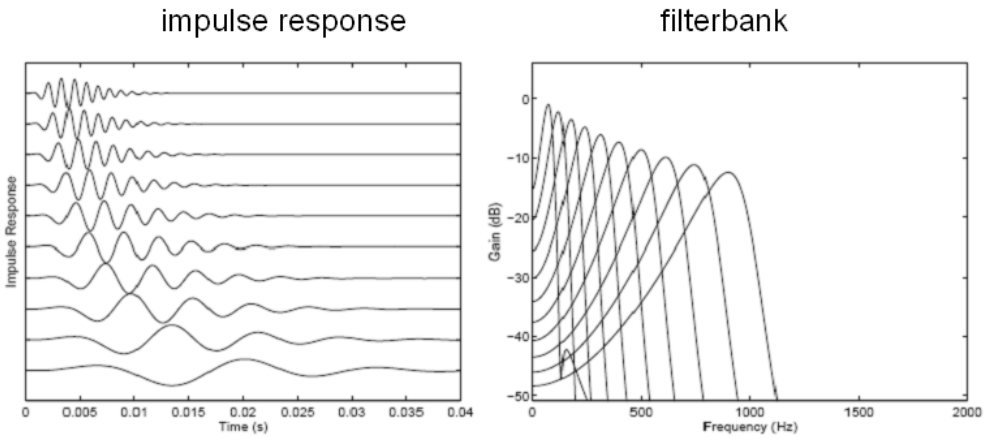

Figure 6: Impulse response and filter shape of the gammachirp filter

3.3 Auditory Modelling

A number of models have been created to represent the human auditory system. The purpose of the models is often to further research into sound perception and predict responses at untested parameters, but also to contribute to technology in the area of machine hearing. Lyon's model (Lyon and Mead, 1988) models the fluid-dynamic wave that travelled through the cochlea by using a cascade of filters based on the properties of the fluid. A set of automatic gain controls was used to simulate dynamic compression of the intensity range of the basilar membrane, a role that is carried out by the outer hair cells. A half-wave rectifier unit modelled the inner hair cells and the neural spike generation that is created when the

stereocilia are bent in only one direction. The output of the model was the probability of nerve firings against time; a cochleagram.

Another such model was Meddis's Inner Hair Cell model (Meddis, 1986), based on the physiology of the inner hair cell, and that the instantaneous amplitude of any one point on the basilar membrane influences the amount of neural impulses released into the auditory nerve. It also took into account the fact that spikes cannot occur more often than 1/ms. While these models of the auditory system are useful in their respective capabilities to predict auditory responses, there is one model that is capable of normalising for size in the way that has been shown is possible for the size of speakers. Patterson's Auditory Image Model, which includes a module containing the Inner Hair Cell model, will now be discussed in detail, as it will be used as inspiration for this study.

3.4 Simulation of the Cochlea and AIM-MAT

The auditory image model (AIM; Patterson et al., 1995) was designed for the analysis of periodic sounds such as speech. It is a time-domain model that uses signal processing techniques to simulate the neural representation of a sound (Patterson et al., 1995), and output this as a time-frequency image. The input signal is split into a multi-channel activity, where a periodicity-sensitive temporal integration then converts this pattern of the neural activity into a dynamic auditory image (Patterson et al., 1992). *aim-mat* (Bleeck et al., 2004) is the MATLAB implementation of the auditory image model in which the user can make use of the graphical interface to analyse the human response to a sound. The modules included in *aim-mat* are pre-cochlear processing, basilar membrane motion, neural activity pattern, strobed temporal integration and stabilised auditory image, which will be discussed here for the purposes of understanding the model's structure. For an in-depth understanding of the processes in *aim-mat*, please refer to the paper by Bleeck and colleagues (2004).

Pre-cochlear processing: The first module in *aim-mat* is a filtering process to simulate the action of the auditory system from when the sound enters the area of the pinna, down through the ear canal, through the bones of the ossicles to the

point of the oval window. In its simplest form this is a band-pass filter following equal-loudness contours, but the parameters can be altered depending on whether the signal would be presented over head-phones or a loudspeaker. There is also an option to choose the loudness model contour described by Glasberg and Moore (Bleeck et al., 2004).

Basilar membrane motion (BMM): This performs a multi-channel spectral analysis on the signal to simulate the role of the basilar membrane. The user can change parameters that control the type of filtering, number of channels, frequency range of analysis and the frequency scale. Linear gammatone filtering is the default setting, but dynamic gammachirp filtering is also an option. An example of this using gammatone filtering with 75 channels logarithmically spaced between 100 Hz and 6 kHz is displayed below in figure 7. The top panel shows the time-series waveform of a click pulse, with the basilar membrane motion just beneath.

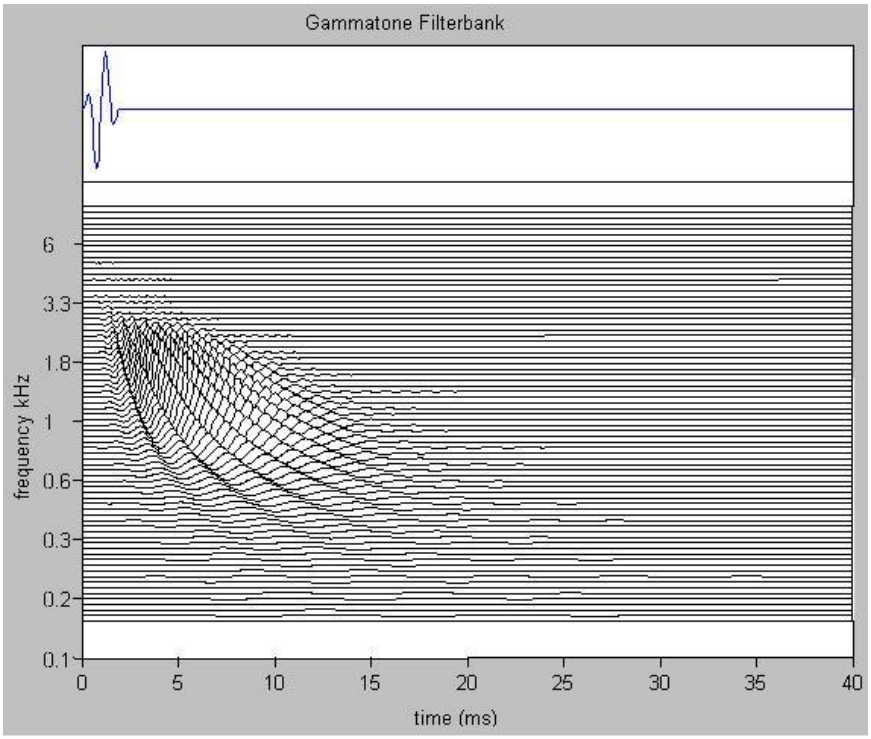

Figure 7: Spectral analysis using a gammatone filterbank of a click pulse from *aim-mat*

Neural activity pattern (NAP): The next module employs three processes in sequence to sharpen the signal, as performed by the outer and inner hair cells. In

the inner ear, the inner hair cells change the motion of the membrane into a neural transmitter, but not before the outer hair cells influence the sensitivity to the input signal. Half-wave rectification emulates the uni-polar response of the hair cell, where the nerves fire in only one direction in phase with the stimulating waveform. Non-linear compression simulates the level dependence of the outer hair cells and reduces the slope of the input/output function. Low-pass filtering is intended to simulate the progressive loss of phase-locking with increasing frequencies above ~1200 Hz. However, this simulation of the NAP lacks adaptation and therefore is not realistic. A more sophisticated solution is the two-dimensional adaptive thresholding module, known as 2-DAT, and is also available on *aim-mat*. This process adds adaptation and suppression to the half-wave rectification, compression and low-pass filtering of the previous solution. This is explained in detail in Bleeck et al. (2004). Figure 8 shows the NAP of the same stimulus as in figure 4. The neural activity follows exactly the movement of the basilar membrane. Two changes from the BMM to this image can be seen: firstly the amplitude of the NAP is always positive due to the half-wave rectification; all the negative amplitude values have been set to zero. Secondly, the activity is compressed due to the logarithmic compression; the high peaks in the mid-frequency range have been lowered, and the low amplitude activity in the lower frequency bands has been amplified. Although it is not clear from this example, the activity is less well resolved at higher frequencies as a result of the low-pass filtering at 1200 Hz, similar to the loss of phase-locking that begins approximately at this frequency.

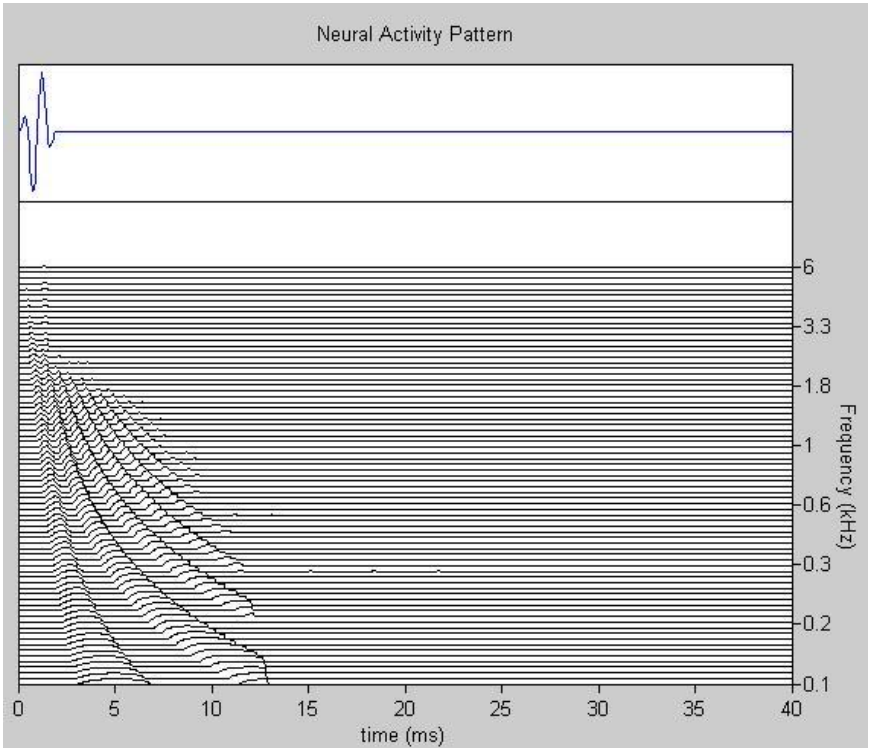

Figure 8: Neural Activity Pattern of a click pulse taken from *aim-mat*

Strobed temporal integration: Repeated stimuli produce a static perception for us. Even when the waveform of a signal is rapidly changing thousands of time per second, when it is repeated with a fixed rate, it appears stable to us. For example, a single wave with a period of two milliseconds will sound like a click, but if a number of them are presented back-to-back they will not be perceived as separate clicks, but as a static tone at a frequency of 500 Hz. Temporal integration preserves the fine temporal structure of periodic sounds from the NAP by a) finding peaks in the neural activity as it flows from the cochlea, b) measuring time intervals from these strobe points to smaller peaks, and c) forming a histogram of the time-intervals, one for each channel of the filterbank (Bleeck et al., 2004). The strobe points mark specific important points in the NAP, which are selected by an adaptive threshold which increases slightly following a strobe and then falls with a decay time set by the centre frequency of the specific channel. Strobe points are shown for the example stimulus in figure 9. From the strobe points the auditory image can be created, which is a stabilised and aligned version of the repeating neural representations of the sound coming from the cochlea.

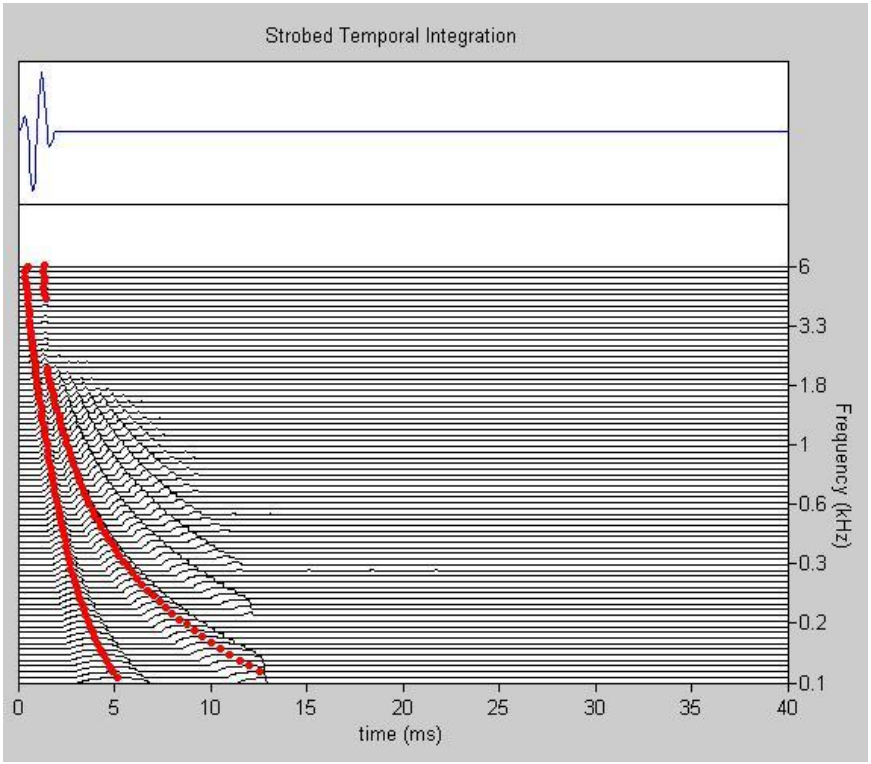

Figure 9: Displaying the position of strobe points on a click pulse taken from *aim-mat*

Stabilised auditory image: The collection of strobe points is used to create a stabilised auditory image. The current SAI process in *aim-mat* is called the 'ti2003'. It operates channel-by-channel by initiating the STI process when a strobe appears, and adds this to previous NAP values which have been scaled. Where a NAP value is entered into the SAI it is determined by the time interval between the strobe and any given NAP value. If no more strobes appear, the process continues unchanged for 35 ms and ends. However, in periodic sounds such as speech and music, more strobes do appear and so the weights of each previous strobe point are adjusted so that older processes contribute less to the SAI (Bleeck et al., 2004). The middle panel of figure 10 shows the SAI of the example (broadband) pulse. The lower panel displays the collapsed activity as a function of time, and the box on the right displays the collapsed spectral activity.

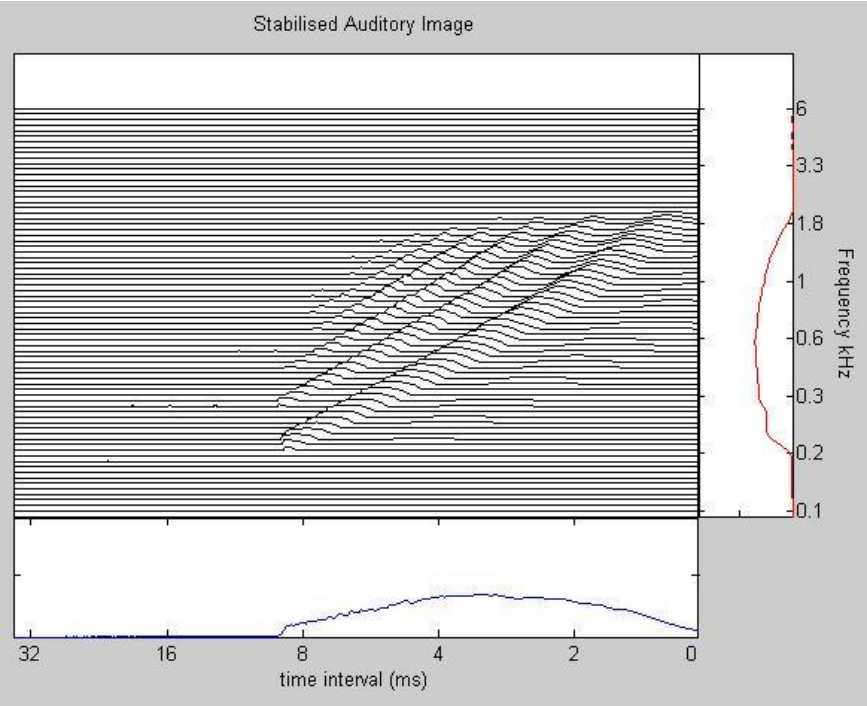

Figure 10: Stabilised Auditory Image of a broadband pulse taken from *aim-mat*

3.5 Mellin Image and Size normalisation in the auditory system

Irino and Patterson (2002) continued their research into the AIM and sound perception by investigating how the ear perceives vowels, more specifically how the ear can distinguish between vowels spoken by speakers of different vocal tract lengths (VTLs). It was mentioned earlier that some sort of scaling mechanism may be in use in the auditory system that may allow the size of an object to be separated from the shape of the object when the object resonates. In speech, this enables humans to recognise vowels uttered by speakers of different sizes. An attempt at simulating this ability for auditory modelling purposes is the Mellin transform.

The Mellin transform has been considered as a form of scaling mechanism, a method that could be used when modelling the size normalisation in the auditory system (Irino and Patterson, 1997). In contrast to the Fourier transform which specifies the energy in each frequency band, the Mellin transform displays scale

invariant properties, and in the case of speech, properties in a vowel that allows the listener to categorise the vowel regardless of the length of vocal tract from where it originated. In other words, the Mellin transform can produce the same magnitude distribution from the same vowel spoken by two different speakers.

In terms of the AIM, the Mellin transform measures the distances between the resonances on the vertical axis of the SAI. The power of the resonances is measured in terms of time, and the distance between the resonances calculated, and it is this relative distance that the magnitude distribution of the Mellin transform is created; these are known as special frequencies. However, for an accurate representation of this, the power of each resonance at each point must be aligned for each vertical column of the auditory image. This is the alignment of the peaks and troughs in each channel of the SAI that creates the Size-Shape Image (SSI). A Fourier transform of each column will show peaks in the spatial frequencies and display these in a Mellin image.

To clarify, imagine a complex periodic signal with a fundamental frequency of 50 Hz, and resonances occurring at 450-, 500-, 550-, 600 and 650 Hz; this is its frequency spectrum. The strongest spatial frequency here is 50 Hz, as this distance repeats five times and so the highest peak in spatial frequency will occur at 50 Hz. The peak will be at 100 Hz, as it repeats twice; 450 to 550 Hz, and again at 550 to 650 Hz. As spatial frequency is a representation of the distance between resonances, the position of the frequency spectrum does not matter. It is in this way that the Mellin transform can normalise for the size of an object. The spectral envelope of the object shifts higher or lower as a result of its size, but the relative positions of its resonances remain the same, and hence the peaks in the Mellin image of the object are the same for different sized objects. A mathematical explanation of this concept follows.

The Mellin transform is a fourier transform performed on a signal that has been filtered using a gammatone or gammachirp filter and the Stabilised Auditory Image has been obtained. One period in the pattern of the SAI is taken to be an Auditory Image (AI) and from here the transform can begin. The Mellin transform of a function, *f*, is:

$$M_I(p) = \int_0^{\infty} f(t) t^{p-1} \mathrm{d}t$$

Equation 6

where *p* is a complex argument and the Mellin parameter (De Sena and Rocchesso, 2005). In terms of the AI, where it represents the resonance information of the object, *f(t)* is replaced with a column of the AI and is designated $A_f(\alpha f_b, \tau)$. αf_b is the centre-frequency of one auditory filter, and τ is the time-interval axis of the SAI. The path of integration is along the lines of *h*, where *h* is the product of time-interval and centre frequency:

$$\alpha f_b \cdot \tau = h$$

Equation 7

and creates the Size Shape Image on which the integration is performed. This is due to a need for the representation to be invariant of scale change. The SSI allows the impulse response of the basilar membrane to be separated from the resonances in the signal. The complete Mellin transform of the AI is:

$$M_I(h, c) = \int_0^{t_p} A_F(\alpha f_b, \tau) e^{(-jc+\mu-1/2)\ln\tau} \mathrm{d}t$$

Equation 8

where t_p is the pitch period is the sound is periodic, and *h* is a constant (Irino and Patterson, 1999). The resulting image is called the "Mellin Image", and the vertical axis is cycles/best-frequency-range, $c/2\pi$. The values in the Mellin image correspond to the spatial frequencies in the SSI where the full frequency range of the SSI from 100 to 6000 Hz is equal to one period (Irino and Patterson, 1999).

Irino and Patterson (2002) investigated how the MI changes with shape. They compared three types of click trains and their resulting MIs. First was a click train of 100 Hz with no resonances. The second signal was a damped sinusoid that was repeated every 10 ms, very similar to a single formant vowel sound at a glottal pulse rate of 100 Hz. The sinusoid has a frequency of 1100 Hz, and an exponentially

decaying envelope with a half-life of 2 ms. The third and final signal was a simplified simulation of a vowel sound with the first two formants (resonances). It consisted of the same signal as before with a second added; a sinusoid of 2600 Hz with an exponentially decaying envelope of half-life 1.5 ms. The repetition rate remains at 100 Hz.

The SAIs showed the lack of resonances in the first signal and the formants caused by the addition of the damped sinusoids in the second and third. The SSI separated the impulse responses from the formants and the relative positioning of the formants could be seen in the second and third. The MI showed a lack of spatial frequency interaction in the first signal due to the lack of formants. The second MI displayed an addition of energy due to the addition of a resonance. Finally, the third MI showed the interaction between the two resonances in the form of broken bands along the spatial frequency axis. The distance between the bands is proportional to the distance between the formants.

Of more interest here is the value of the Mellin transform and the Mellin image for the normalisation of sound-source size. To show the difference between how the SSI displays size differences and the MI normalises these differences, Irino and Patterson (2002) synthesised two of the same vowel sounds, /a/, one with a glottal pulse rate of 100 Hz to simulate a male speaker, and the other of 160 Hz to emulate a female vocal tract which is one third shorter than the males. Figure 11 shows the stabilised auditory images of the vowels and the positioning of the formants are very clear. The second and third formants of the male vowel occur at 1100 and 2500 Hz respectively and the female vowel at 1600 and 3900 Hz, having moved up by a factor of 50% due to the shortening of the vocal tract (Irino and Patterson, 2002).

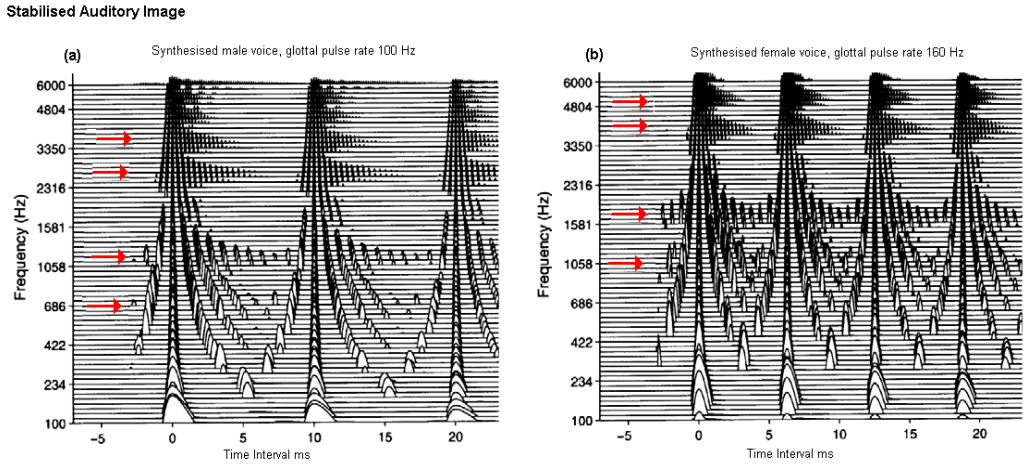

Figure 11: SAI of two synthesised vowel sounds /a/. Left: The SAI of a synthesised male /a/. Right: The SAI of a synthesised female /a/. The arrows show the formants. Reproduced from Irino and Patterson (2002) with permission from Elsevier.

The size-shape image of figure 12 which is a product of time-interval and peak frequency, shows the more emphasised formants of the vowels compared to the SAI, where they have been separated from the broadband impulse response of the basilar membrane. The relative positions of the vowels are the same, but the absolute positions in the shorter vocal tract have moved up by a factor of 3/2.

In the Mellin images shown in figure 13, spatial frequency, $c/2\pi$, is represented on the vertical axis. The initial broadband activity in the SSI due to the impulse response of the glottal pulse translates to low spatial frequencies, $c/2\pi \leq 4$, in the Mellin image. Moving along the horizontal axis, the formants in the vowel which are well spaced in the SSI, begin to show up as activity in the MI around $c/2\pi$ values of 6, 10, and 14. These features of spatial frequency appear in the same positions for the Mellin images of both vowels, showing how the MI can normalise for size or in this case, vocal tract length (Irino and Patterson, 1999).

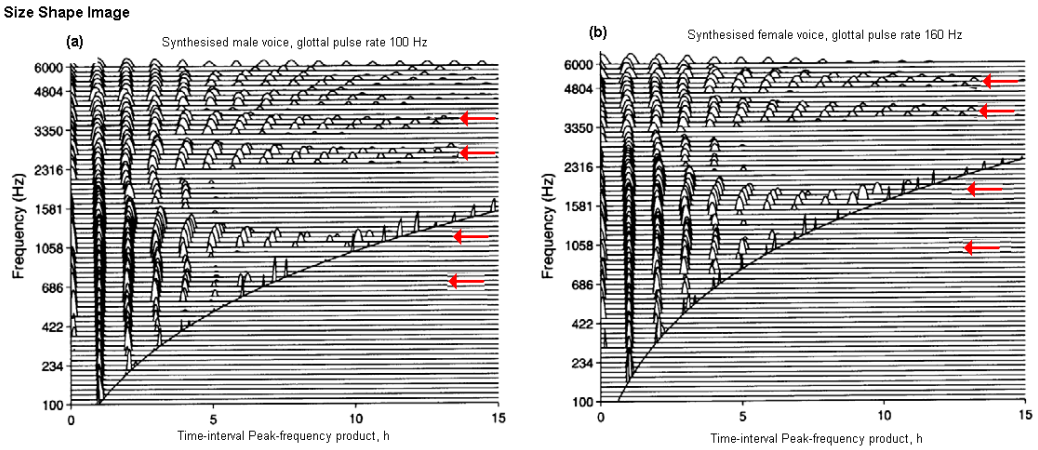

Figure 12: Size Shape Image of two synthesised vowel sounds /a/, male (left) and female (right). The arrows show the formants. Reproduced and adapted from Irino and Patterson (2002) with permission from Elsevier.

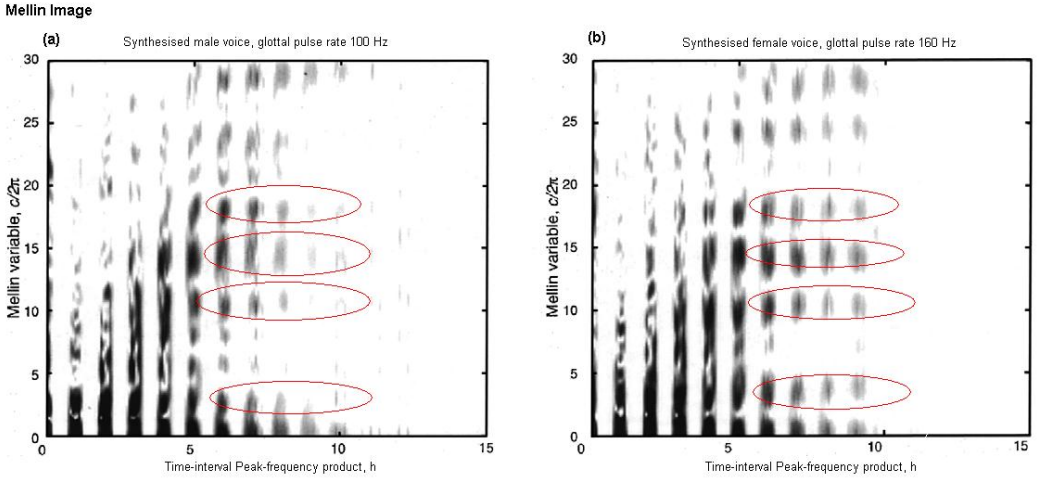

Figure 13: Mellin power images of two synthesised vowel sounds /a/, male (left) and female (right). Circles show the positions of the corresponding spatial frequencies, unaffected by size. Reproduced and adapted from Irino and Patterson (2002) with permission from Elsevier.

3.6 Assumptions of AIM and Mellin transform

The Auditory Image Model analyses speech by performing a spectral analysis using a gammachirp filterbank. It then takes an average of each glottal pulse and accompanying resonances in order to create a stabilised auditory image of a

continuous periodic signal. The Mellin image is created by separating the size information about the length of the vocal tract from the resonances that identify the vowel that is being uttered. The model assumes that for every size of VTL, and therefore person, the relative positioning of the resonances do not change when the same vowel is spoken. This is the underlying theory that allows the Mellin image to normalise for size, and produce an image containing information purely on the shape of the vocal tract, and the vowel spoken. It has been shown that the Mellin transform is successful with vocal tracts of different lengths, but it has not yet been used with transient sounds.

The assumption that the normalisation of speaker size is based on concerns the importance of the position of the spectral envelope, and that the envelope shape remains the same for different sizes. This study works towards investigating whether or not it is the same for transient signals, and if so, a model will be created for the analysis of transient signals in a similar way to the analysis of vowels carried out by AIM.

4. Scaling Transient Signals

The motivation behind this research is to understand the importance of spectral cues in the discrimination of transient signals from objects of different sizes. The term transient pitch was mentioned earlier as a way of distinguishing it from musical pitch, and refers to how one would identify one transient as sounding higher or lower than another. One way of looking at the importance of spectral cues is finding a method of scaling single pulse-resonance sounds so that an experiment may be carried out where transient is no longer a cue for a size discrimination task. As discussed in chapter 2, the most mentioned cue in the size perception experiments was spectral, but there were also mentions of timbre, loudness and duration. The purpose of this research is to establish the importance of spectral cues in a size discrimination task and create an auditory model for transient sounds. As well as ensuring the signals are of the same loudness and duration, scaling them could create a set of signals that sound the same in terms of their pitch, and would identify if there are any other cues that can be used to tell them apart.

A study by Smith and colleagues (2005) was mentioned in chapter 3 that demonstrated the ability of listeners to discriminate between vowels that were scaled to sound as though they were spoken by people with different length vocal tracts. The signals were scaled using the STRAIGHT processing package (Kawahara et al., 1999) that separates the spectral envelope from the individual glottal pulses of the voice sample. The glottal pulse rate can then be expanded or contracted with the spectral envelope superimposed over it to recreate the vowels as spoken by people with longer or shorter vocal tracts. This method proved successful, and the experiment for size discrimination and vowel recognition was

then achieved with positive results. This method has not yet been applied to transient signals.

Since voice is a periodic sound, the STRAIGHT method of resynthesizing vowel sounds separates the spectral envelope from the glottal pulse rate. The signals used here are transient, i.e. non-periodic single pulse-resonance sounds recorded from different sized tokens of the same object. There is no periodicity, thus each individual spectrum is the equivalent to the spectral envelope mentioned above. Care is taken to eliminate the other cues that affected the results in studies in the literature, such as intensity, temporal and TSP cues, in order to isolate spectral cues as much as possible.

4.1.1 Sound Source Recording

Polystyrene Styrofoam modelling spheres were chosen to create real in-air sounds. It is not the intention with this particular study to find the threshold of size discrimination and so the source sounds here will have clearly audible size differences. Due to the hollow nature of solid foam, each size of polystyrene sphere produces an obviously different sound to the others; hence their sizes are easily distinguishable. The spheres chosen had diameters of 70mm, 80mm, 90mm, 100mm and 120mm. Spheres smaller than these proved difficult to control during recording, and were discarded. Spheres larger than these were not manufactured as 'solid' whole spheres; they were moulded separately into two hemispheres and glued together. This affects the homogeneity of the spheres and would alter the transmission of the sound through them, and so they were not used for the experiment.

The equipment used to record the signals was:

Equipment:

- PCB High Sensitivity Free-field ½ inch microphone capsule - Model No. 377B02, Serial No. 108354
- PCB Microphone Pre-amplifier - Model No. 426E01, Serial No. 012866
- PCB ICP Sensor Signal Conditioner - Model No. 480E09, Serial No. 00028169
- 2 BNC Cables

- Dell Inspiron Laptop computer - Model: Latitude D520, Serial No. GSRZC2J
- Clamp to hold microphone
- 15 solid polystyrene Styrofoam spheres
- 3 x 70 mm, 80 mm, 90 mm, 100 mm and 120 mm diameter
- 1 x 40cm wooden rod, 10 mm diameter – Experiment 1
- 2 steel ball bearings of 10mm diameter – Experiments 2-#
- Spool of 0.5mm nylon thread

Software:

- Windows XP Professional
- Adobe Audition V.3

Figure 14: The five different sizes of polystyrene Styrofoam spheres used in the experiment. The numbers indicate the size of the diameter in millimetres. The ruler is there to represent the scale of the picture; it is 30 cm in length.

This description of the recording process is for all experiments except the first; that will be discussed within the method of experiment 1. The recording of the signals took place in a sound insulated room with anechoic properties. The sounds used were created by the impact between the spheres and a striking ball, a steel ball

bearing of 10mm in diameter, suspended 175 cm from the ceiling at a distance of 10 cm from the centre of the polystyrene sphere. A plumb line was created at a distance of 60 cm from the striking ball to be used as a guide for the point from where it should begin its fall. The distance of the striking ball from the ceiling was constant; the length of the thread attached to the polystyrene sphere was adjusted depending on the size of the sphere so that the ball bearing always struck as close to the central axis of the sphere as possible. It is known that the position of impact has an effect on the spectral content of a sound (van den Doel and Pai, 1998), and so it was deemed important to the uniformity of the recordings that the impact position and force applied remained as constant as possible.

Recording Set-up

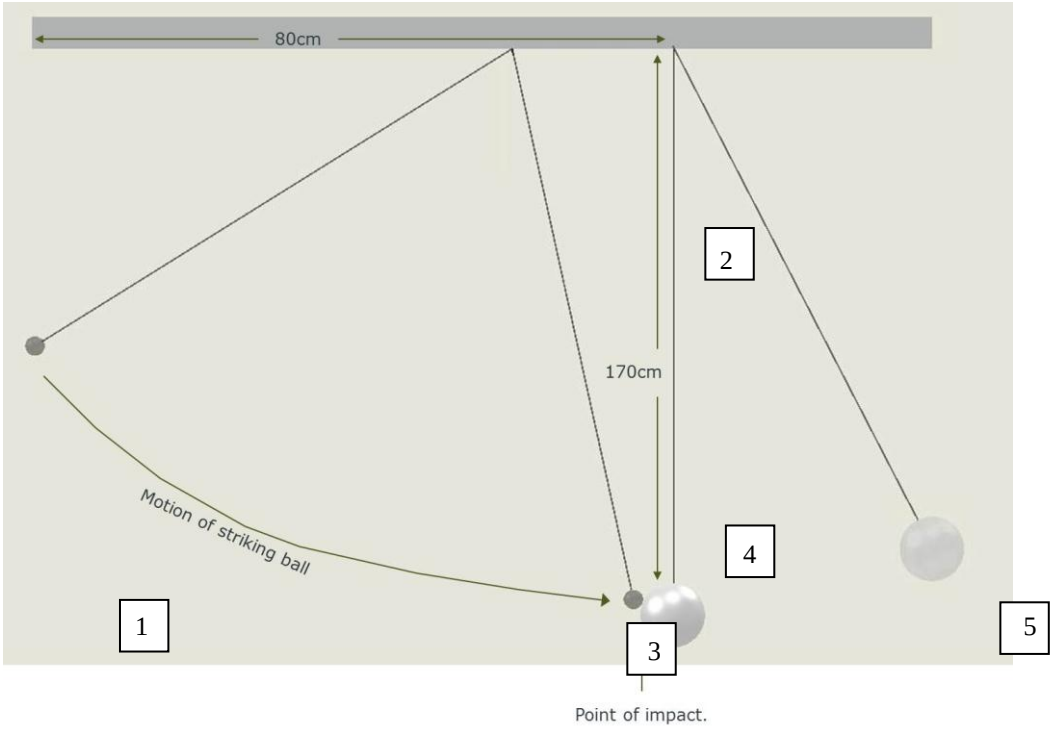

Figure 15: A sketch of how the sound was created. (1) A plumb line was used as a guide for where the striking ball is dropped from. (2) Here indicates the resting position of the striking ball, suspended 175cm from the ceiling at a distance of 70cm from the plumb line. (3) This is the point of impact between the striking ball and the static polystyrene sphere. (4) Here is the resting position of the polystyrene sphere, suspended ~170cm from the ceiling, its centre positioned 10cm from the striking ball. (5) This is the movement of the sphere after it has been struck.

The sound was created by allowing the striking ball to fall from the line of the plumb weight and impact with the sphere as close to along its diameter as possible. This method ensured the force of the impact was constant, and the size and density of the striking ball ensured that the signal contained energy from vibrations in the polystyrene sphere only. The estimated force of the strikes to each suspended sphere was ~0.055 N, calculated using the mass of the ball-bearing and the angular acceleration of its swing. Each of the three samples of each of the five sphere sizes were struck 100 times and recorded, making a total of 1500 recordings. For both types of sounds, the software Adobe Audition V3.0 was used to capture the 16-bit recordings at a sampling rate of 44.1 kHz. A diagram of the set-up is shown in figure 15.

4.1.2 Properties of Polystyrene

Tables 1 and 2 show the physical properties of the polystyrene spheres and steel ball-bearings used to create the sounds in the experiment. The descriptive diameters given by the manufacturer were checked for accuracy and found to be different, so the measured diameters are also listed, as well as the measured mass and calculated densities of each.

Properties of polystyrene spheres

Diameter (descriptive / measured)	Mass	Density (kg / m^3)
120mm / 119mm	17.95 g	20.35
100 mm / 99.1 mm	11.36 g	22.91
90 mm / 89.1 mm	8.22 g	22.19
80 mm / 79.2 mm	5.61 g	21.54
70 mm / 67 mm	3.40 g	21.53

Table 1 This table shows the descriptive (stated by manufacturer) and measured diameters, mass and calculated densities of the five polystyrene spheres used for recording.

Properties of steel ball-bearing

Diameter	Mass	Density (kg/ m^3)
10 mm	5.65 g	10,733

Table 2 A steel ball-bearing was used to strike the polystyrene spheres. This table shows its measured diameter and mass and calculated density.

Using information about the properties of the polystyrene material as well as the size and shape of the spheres used, the position of the first resonance in each spectrum can be estimated, and will be compared to power spectral density calculations later. The size, shape and densities of the spheres are already known, but to calculate where resonant frequencies will occur, the speed of sound through polystyrene is required. The equation to find the speed of sound in a solid is:

$$c_s = \sqrt{G/\rho}$$

Equation 9

where c_s is the speed of sound in a solid, *G* represents the Shear Modulus of the material, and ρ is the density of the material. The Shear Modulus, *G*, describes the response of a material to shear strain and the equation to calculate this is:

$$G = 2E(1 + v)$$

Equation 10

where *E* is Young's Modulus, and *v* is the Poisson's ratio of the material. Young's Modulus is also known as the elastic modulus and it is a measure of the stiffness of an elastic material, and the Poisson's ratio is the ratio of the corresponding expansion in other directions of a material due to compression in one direction. These three parameters: Shear Modulus, Young's Modulus and Poisson's Ratio, all affect how vibrations propagate through a material and thus how fast the sound moves through it and around it. From various studies, an average of Young's Modulus for polystyrene of density 20 kg/m^3 (Miki, 1996a; Horvath, 1995; Duskov, 1997; Eriksson and Trank, 1991; all compiled by Elragi,

2006), and the Poisson's Ratio (Yamanaka et al, 1991; Negussey and Sun, 1996; GeoTeck, 1999a; Duskov et al, 1998; Ooe et al 1996; all compiled by Elragi, 2006), the table below could be compiled. The speed of sound calculated in the spheres here is used in section 4.1.4 to show how the properties of the spheres influenced the frequency spectrum.

Mechanical Properties of the Polystyrene spheres

Sphere	Density ρ	Poisson's Ratio v	Young's Mod E	Shear Modulus	Speed of sound
XL	20.35 kg/m^3	0.1958	6 MPa	9 MPa	839.7 m/s
L	22.91 kg/m^3	0.1958	6 MPa	9 MPa	791.4 m/s
M	22.19 kg/m^3	0.1958	6 MPa	9 MPa	804.2 m/s
S	21.54 kg/m^3	0.1958	6 MPa	9 MPa	816.2 m/s
XS	21.53 kg/m^3	0.1958	6 MPa	9 MPa	816.4 m/s

Table 3 Using measurements, the average values taken from studies online and the formulae mentioned in the text, this table shows the mechanical properties of polystyrene that enables the calculation of the speed of sound through the spheres.

4.1.3 Signal Processing

Before any experiments could take place, the recordings had to be processed. First, a Butterworth band pass filter of order 4 was chosen because of its flat frequency response and steep roll-off. This was applied to all the recordings with a pass-band from 100 - 16,000 Hz so as to eliminate any unwanted low-frequency sounds, and to limit the upper frequencies to within the audible range. Each recording was chopped into the individual impact sounds; the waveforms aligned to the point of the first negative peak, and then cropped so all the recordings are of the same length. Being from the same type of object, the signals tended to have a general impulse response pattern, and can be seen from the left panel of figure 16: an initial small positive peak followed by a large negative peak, which could be from

the moment of impact, followed by a slowly diminishing wave of resonances. In the larger of the spheres, the shape of the response changed slightly according to where on the sphere was struck. Although every effort was made to ensure this happened as little as possible some signals resulted in a different sound and waveform. Figure 16 shows the general shape compared with an example of the more unusual waveform shape, and followed by the spectrogram form taken from Adobe Audition. The second signal shows more fluctuation in the time-series representation, likely due to the addition of high frequency energy as a result of being struck at the wrong point. The signals that did not sound clear enough were removed from the signal set at the discretion of the author.

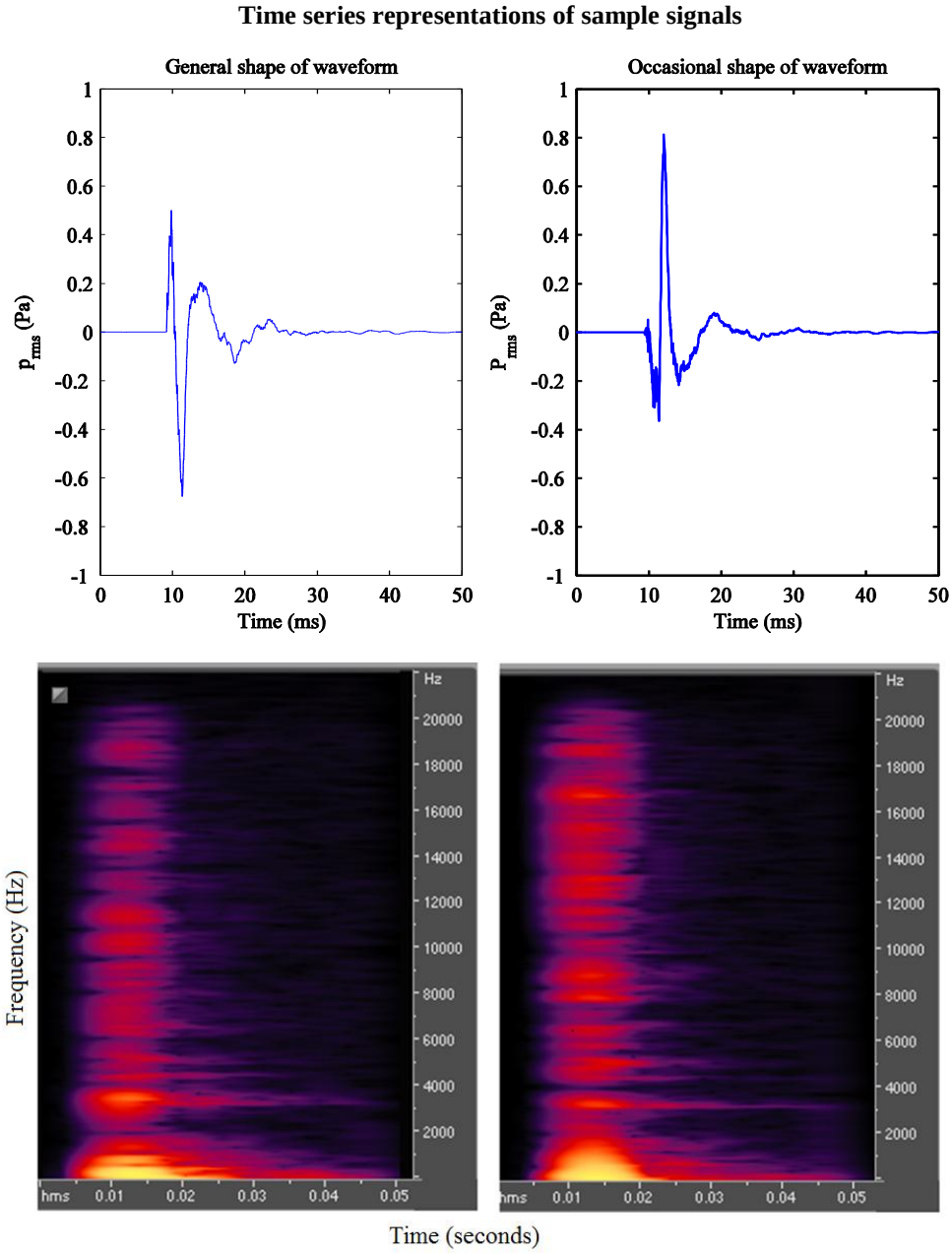

Figure 16: Time series and spectrogram representations of two of the different types of signals from the extra-large polystyrene spheres. The top left panel shows the general shape and frequency content of the signals when hit correctly along the central axis of the sphere, but the panel on the top right show the waveform of the signal which has been hit off the central axis, which has altered the temporal shape of the waveform and contains more high frequency energy. The lower two panels of the figure show the spectrograms from Adobe Audition, and the high frequency energy in the irregular signal is visible in the top 10kHz of the bottom right panel.

The next step in the signal processing was equating the root-mean-square values of the signals. The RMS of each signal was calculated, and mean RMS per size group and those outside a standard deviation of two from this mean were found. Figure 17 shows boxplots of the variation in the RMS values for each group of recordings. As the signals differed somewhat in form, but retain some similarities, all those that lay outside two standard deviations from the mean were discarded; these values are indicated by blue stars. The variety in the signals allow for a more realistic auditory representation of the spheres, as opposed each recording being identical. After the removal of those signals with an RMS of more than two standard deviations from the mean, each signal was normalised to an RMS of 0.1. The boxplots of figure 17 shows the decreasing trend in average RMS value as the size of the sphere decreases before normalisation. Normalising the signals ensured that the intensity differences were not used as a cue for differentiating between the spheres, and checks were made by the author throughout testing to guarantee this as much as possible.

Boxplots showing the range of RMS values for each group of recordings.

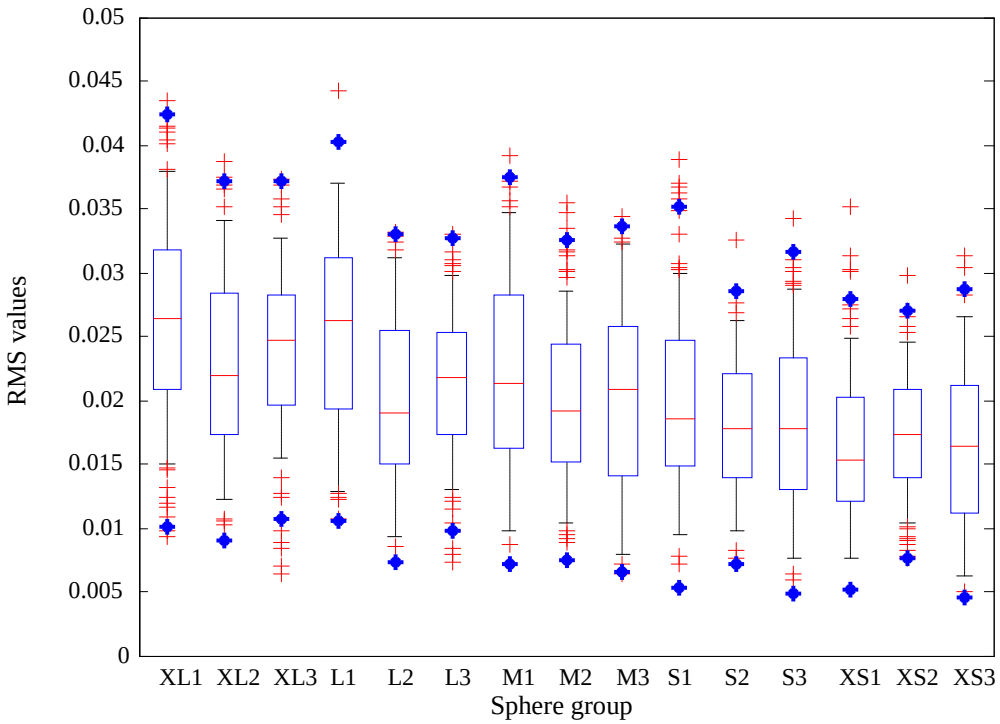

Figure 17: Boxplots showing the distribution of RMS values for all signals of each size, before being normalised to an RMS of 0.1. The median RMS value is indicated for each group of spheres by the red line. The edges of the blue boxes mark the 25th and 75th percentiles. The blue stars indicate the points of 2 standard

deviations from the mean; the red crosses outside the range of these blue stars represent the RMS values of signals that are left out of the experiment.

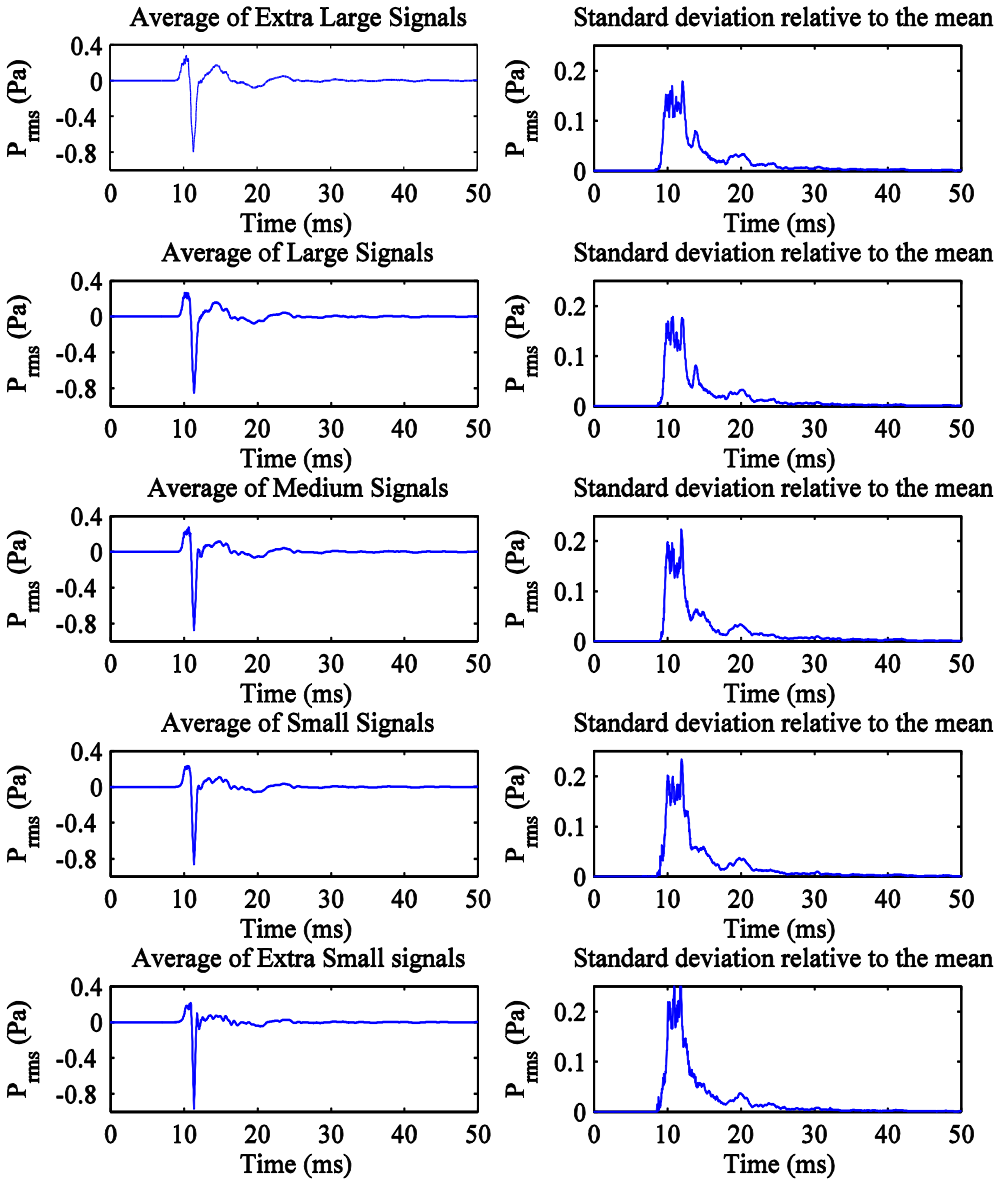

Figure 18: Averaged time series of all the recordings for each size group are shown in the left column. Signals with an RMS of more than two standard deviations from the mean RMS of that group were removed. All remaining signals, shown here, are aligned to the first negative peak, and normalised to an RMS of 0.1. The column on the right shows the standard deviation from the average signal for each size. The averaged signals show a small peak in amplitude followed by a steep decrease before oscillating gently around the zero line and coming to a stop.

After the processing and eliminating of the unwanted signals, a selection of approximately 290 recordings for each group of sphere size remained. Figure 18 shows the time history of the averaged signals on the left column, and the right column displays the standard deviation of all the signals relative to the averaged signal. The initial peaks of the signals tend to vary by about 0.2 Pa, but the variation levels off at zero towards the end of the signal.

The signal processing so far has ensured that intensity and loudness cues have been accounted for. As for temporal cues, the waveforms have the same duration of 50 ms, wherein the majority of the resonance rings for ~20 ms. Figure 19 below shows the time-series waveforms of the averaged signal for each sphere plotted one on top of the other to show how the duration and length of the resonances are similar for all sizes. The method of recording caused this, as the struck sphere swung away from the microphone and so the amount of time the resonances were recorded was limited.

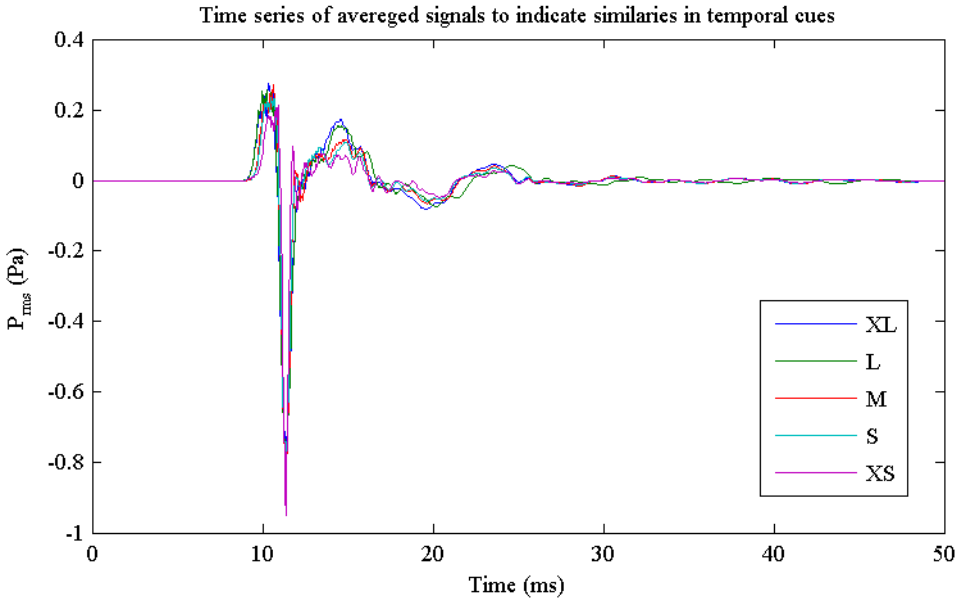

Figure 19: Time series representations of the averaged versions of the 5 sphere signals, plotted on the same axis in order to display the similarities in resonance duration.

4.1.4 Spectral Content

In order to view the spectral information of each signal, a Welch power spectral density (PSD) estimate was calculated through MATLAB, (nfft = 2048, window = 512), and the resulting frequency spectra are shown in figure 20. The spectra appear to have a similar shape, but this shape moves higher along the frequency axis as the size of the sphere gets smaller, as though it is a scaled spectral envelope. The arrows indicate the first and second resonant peaks for each of the spectra. The relative positions of the resonances shift along the frequency axis according to the size of the sphere, with the larger spheres having lower resonances than the smaller spheres.

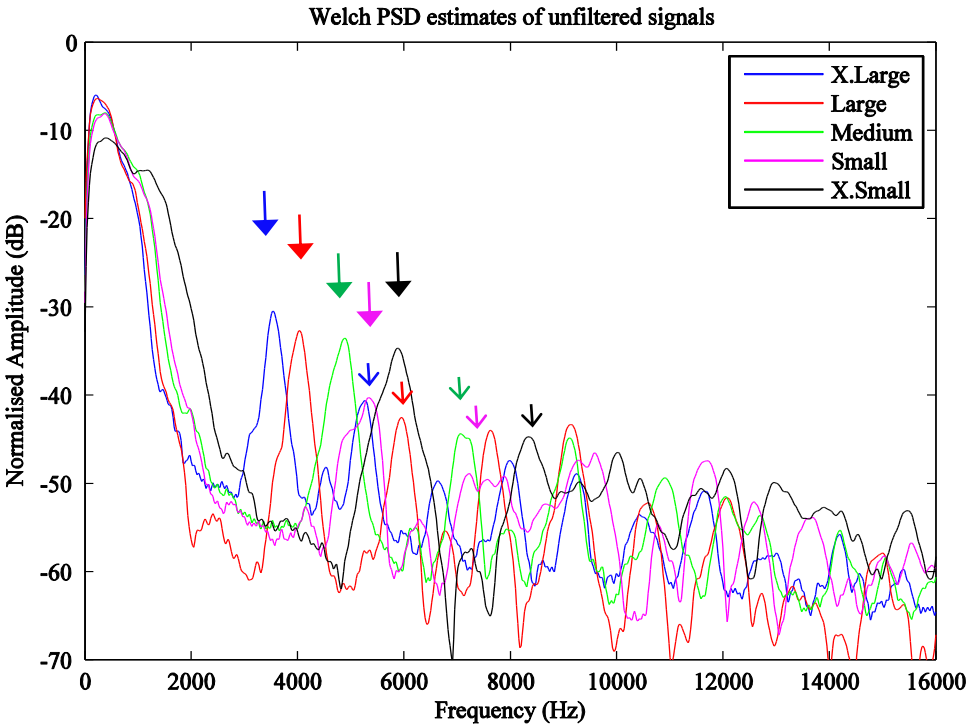

Figure 20: Welch Power Spectral Density (PSD) plots of each of the five averaged signals. Each coloured line represents the PSDs for the five different sized spheres. The coloured arrows point to the first (🡻) and second (🡻) resonant peaks for each PSD.

Table 4a below displays the values of the first and second resonances for each sphere, as well as the difference between the first resonances for each neighbouring sphere. The last row shows the difference between the first and second resonance for each sphere.

Frequency values and differences between F1 and F2

Table 4a

	F1	F2	F1 / F2	F2/F1
X.Large	3531 Hz	5254 Hz	0.672	1.49
Large	4048 Hz	5943 Hz	0.681	1.47
Medium	4909 Hz	7062 Hz	0.695	1.44
Small	5340 Hz	7235 Hz	0.738	1.35
X.Small	5857 Hz	8354 Hz	0.701	1.43

Table 4b

	XL to L	L to M	M to S	S to XS
F1 differences	517 Hz	861 Hz	431 Hz	487 Hz

Table 4a shows the frequency values of the F1 and F2 in each sphere spectrum, and ratio between the F1 and F2 for each size. 4b shows the distance between F1s in neighbouring sizes.

The average ratio between first and second resonances is 0.694, with a standard deviation of 4.13% of the mean. The small sphere has resonances that are least like the other spectra, and it also has the F2/F1 ratio that does not follow the same pattern as the others. This becomes very clear when F1 is plotted against F2 in figure 21 below, where there is an obvious linear relationship between these resonances for all sphere sizes except the Small sphere.

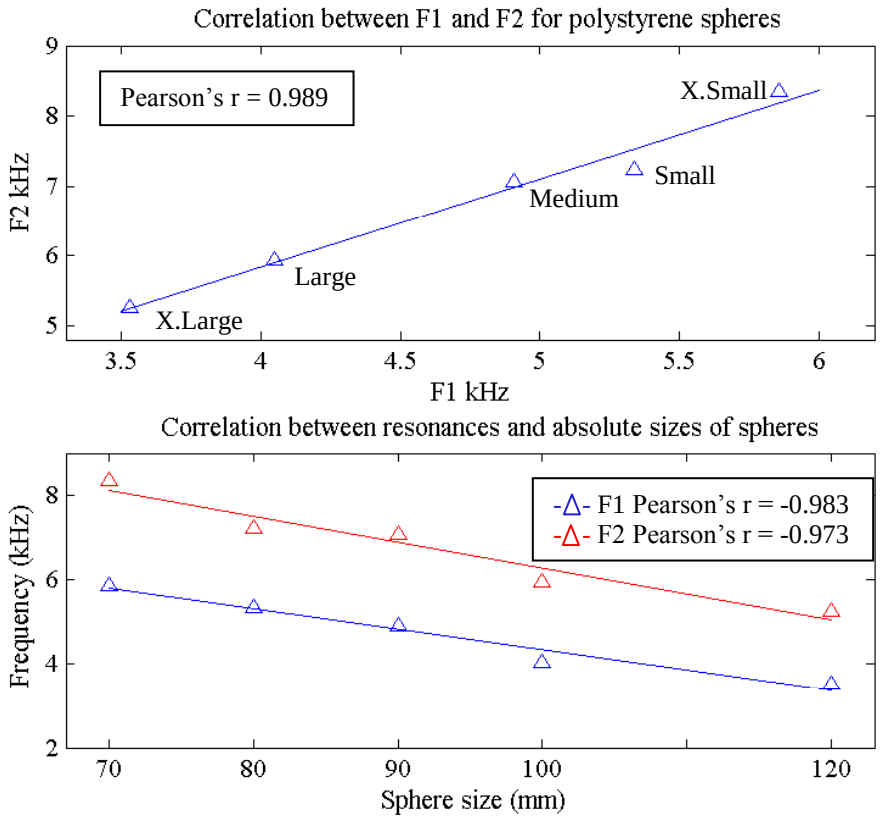

Figure 21: A diagram of F1 against F2 for all sphere sizes, and the individual F1 and F2 resonances against absolute size of sphere. The Pearson's correlation coefficients, r, show strong linear correlations for all relationships above.

The F1 for each sphere are related to the dimensions and the properties of polystyrene. When sound propagates through a sphere, the wave travelling across the diameter of the sphere and back again is the wavelength of the first resonance in the spectrum. Using the speed of sound calculated in section 4.1.2 and the diameter of each of the spheres, an estimate of the first formant can be calculated and compared with the F1s above. The equation used is:

$$f = \frac{c_s}{\lambda}$$

Equation 11

where λ is twice the diameter of the sphere and also the length of the wave propagating through the sphere; c_s is the speed of sound through the sphere taken from table 3; and f is the frequency of F1.

Table 5 below shows the comparisons:

Sphere	Diameter	Corresponding Frequency Diameter = λ/2	F1 from PSD
X.Large	119mm	3528 Hz	3531 Hz
Large	99.1 mm	3993 Hz	4048 Hz
Medium	89.1 mm	4513 Hz	4910 Hz
Small	79.24 mm	5150 Hz	5340 Hz
X.Small	67.06 mm	6087 Hz	5857 Hz

Table 5 The calculated resonance corresponding to the diameter of the spheres and the measured F1 values from the sphere spectra. The chart below shows the similarities more clearly.

Speculation was made for other resonances in the spectra, for example F2 and F0 – the wide band of low energy occurring in the spectrum below ~ 1.4 kHz. F2 works out to be just under 1.5 times the value of F1 for most sphere resonances. Using the equation above but with the speed of sound in air and the circumference of the spheres as the wavelength, frequency values that are very similar to the upper end of the F0 band are calculated. However, this is not the correct method of calculating surface waves of a sphere, and the mathematics of vibrational patterns in solid objects is not within the scope of this research, so no further calculations were made. Figure 22 below shows the similarity between the calculated F1 and the observed F1, where the diameter of the sphere is half the wavelength of F1.

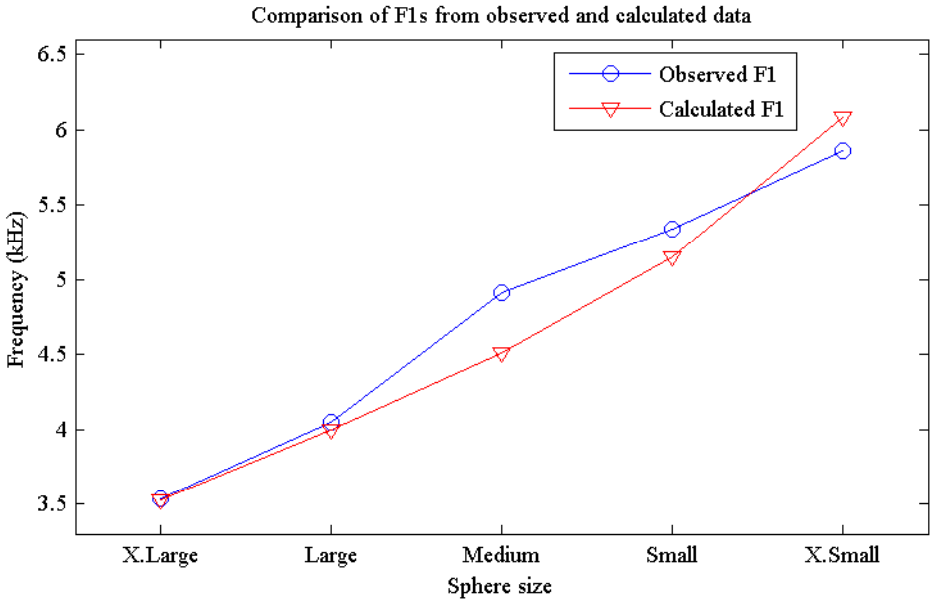

Figure 22 A chart of the F1s observed from the PSDs of the spheres, and the calculated F1s using Equation 11, where the diameter of the spheres represent half a wavelength of F1.

4.2 Experiment 1 – Scaling using PSR

Despite the differences between vowels and a sphere being struck, the structures of their spectra are similar in their linear quality. Turner and colleagues (2009) analysed correlations between formant frequencies of vowels and the sizes of speakers and found there to be a linear relationship between the two. The spectral analysis of the spheres in section 4.1.4 also shows there is a linear relationship between resonances and size. At the beginning of this chapter the STRAIGHT method of scaling vowels was mentioned, and that this had yet to be attempted using transient signals. The method involved separating the GPR from the formants and scaling the latter, the spectral envelope, separately. Transient signals do not have a GPR as they are not periodic. Therefore a method of simply increasing or decreasing the playback sample rate (PSR) of the recorded signals was proposed as a suitable method of scaling the signal. The objective of experiment 1 is to test this scaling method to see if a shift in the spectral envelope in this manner is sufficient for transient signals, as it has been shown to be for vowel sounds.

In order to test the method, participants were presented with recorded signals of two different sizes of struck objects in order to alter the playback sample-rate (PSR) of one to sound like the other. They were specifically asked to listen to the transient pitch of each size of sphere, and adjust the PSR so that the pitch of each sounds the same. This resulted in an average scaling factor for each size, which was used to create a set of scaled signals to be presented in a size discrimination task. The success of the scaling method was judged on the difficulty of the scaling process and the change in discrimination ability after the signals are scaled. The theory that spectral cues are the most important cue in size discrimination suggests that this scaling task would prove to be easy for the participants, however it is not a pure tone or a musical tone with harmonics and therefore the complexity of the signals may present difficulties.

4.2.1 Methodology

Five postgraduate students with more than fifteen years of musical training each volunteered to participate in this experiment. It was deemed necessary to have musical training as a condition of participation in this experiment due to evidence previously found relating increased pitch discrimination abilities to musical experience (Micheyl et al., 2006), and since this was quick test of the viability of the scaling method it was not considered important to use a larger number of participants. Two males and three females took part, all participants reported normal hearing, and the average age of the group was 26.8 years. The participants were asked to familiarise themselves with the sound sources by playing with the spheres, tapping them, bouncing them on the table, and indicating that they were able to tell a difference between the sounds each sphere size made.

Testing took place in a quiet room with only the participant and the tester present. Participants were asked to sit at a desk in front of a computer monitor and given circumaural Beyer Dynamic DT-990 headphones and a computer mouse. These were connected to a Dell Inspiron laptop computer that the experimenter accessed to run each test. At the start of the first session the method for the experiment was explained to the participants and then they were given one minute to familiarise themselves with the five different sized spheres; the sound sources

used for the experiment. Instructions were given to the participants to bounce each of the spheres off the table, tap them with a pencil or finger, and rotate them in their hands. They were then asked to indicate whether they could hear a difference between each of the spheres; all said they could easily hear the differences between all the sizes.

For this experiment, the sounds used were created by striking the spheres with a wooden rod, and the experimenter tried to maintain as consistent a speed and force of strike as possible throughout. The sphere to be struck was suspended at 170 cm from the ceiling by a fine nylon thread to a distance of 60 cm from the floor. A microphone stand was positioned on the floor and held a PCB High Sensitivity Free-field ½ inch microphone 11 centimetres away from the stationary sphere, but perpendicular to the direction in which the sphere would swing when struck. The microphone was connected to a Bruel & Kjaer Charge Amp, and then to an Edirol hub which was connected via USB to a Dell Inspiron laptop computer. The signal recorded was the sound made by the rod striking the sphere directly in front of the mic. The sphere was allowed to swing after each strike to allow for any lasting resonances to be recorded. Each of the five spheres was struck fifty times in total, and the recorded signals were then edited into signals of the same length before presented to participants.

The testing procedure used was a 2AFC method, using the same GUI as in Experiment 1. A comparison pair consisted of two recorded signals, each repeated five times in close succession, presented with a gap of 0.5 seconds between the first and second signal. The GUI had four buttons: two large ones were labelled 'Sound 1' and 'Sound 2' indicating the first and second sound, which lit up in green depending on which sound was playing through the headphones at the time. The third button was labelled 'Start' and the fourth was a 'Play again' button so the participant could repeat the pair of signals if required. To choose an answer, the participant simply had to click on either the 'Sound 1' or 'Sound 2' button. They were aware of how many pairs were in each round and how far through their round they were by way of a counter at the bottom of the GUI. The question at the top of

the GUI read 'Which sound comes from the bigger object?' Figure 23 shows the graphical user interface used for gathering the data.

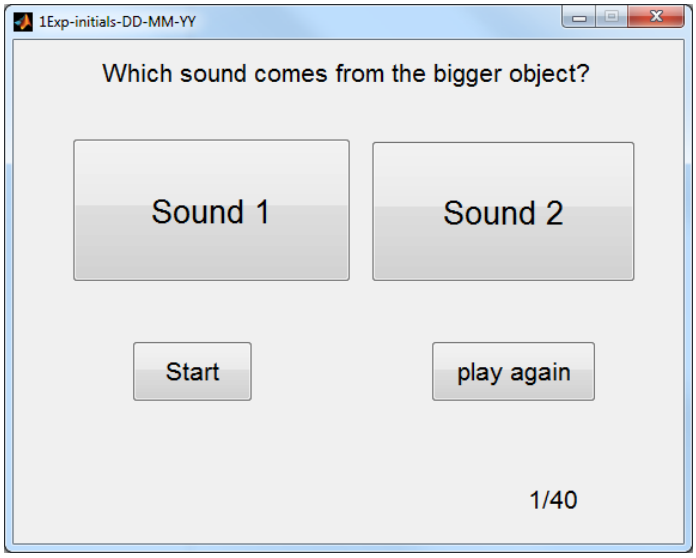

Figure 23: Graphical user interface used in experiment, created using MATLAB. The buttons 'Sound 1' and 'Sound 2' lit up as the signals were playing, and the participant could play the signals again if they so needed.

The experiment was separated into two parts. First, the participants were required to compare a recording of each size of sphere to a reference sphere. To do this they were asked to judge the pitch of each sphere and indicate which one sounded lower. Gaver (1993) refers to this as musical listening, when attention is paid to the musical attributes of a sound. The Medium sphere was assigned as the reference sphere to which the other sizes were compared. The PSR of the comparison sphere was adjusted until the perceived pitch of the two signals sounded the same. This first task is a loose form of a pitch comparison experiment, and so the musical experience of all the participants would have guaranteed a greater degree of accuracy in matching the pitch of the spheres. Five sets of ten pairs of comparisons were presented to each participant in a 2-alternative forced choice (2AFC) method, with 8 reversals in each trial to achieve the subjective scaling factor for the comparison signals for each participant. The average over each set found the scaling factor for each signal for each participant. A method of 'roving amplitude' was employed, where each signal presentation was adjusted randomly by ±2 dB SPL to avoid loudness cues from the signals. A short listening

test of the signals was carried out by the author to their satisfaction. The signals were repeated five times in succession with a gap of 0.1s between each repetition for the participant to get a good sense of any pitch contained within the resonances, but at a rate of 10 per second they were far enough apart not to produce any TSP.

The second part of the study consisted of two size discrimination tasks: the first to determine how well the participants could discriminate between the original signals, and then to establish how successful the scaling was; their performance should decrease significantly. The first task required the participants to compare pairs of the original unscaled signals, and answer which of the signals sounded larger. The signals were those of the recorded spheres, unaltered or scaled. Since the participants had already indicated their ability to tell the spheres apart, only one set of results was taken from each participant for the unscaled task. The second size discrimination task used the signals which had been scaled, and four sets of comparisons were presented 20 times to each participant in order to answer the same question "which object sounds larger?" Both tasks were presented in a 2AFC method, and again within each presentation the signal was repeated five times in succession with a gap of 0.1s between each repetition.

4.2.2 Results

This experiment turned out to be more difficult than expected. The purpose of the experiment was to test if shifting the spectral envelope of a transient signal allowed for the matching of transient pitch, thereby suggesting that size discrimination worked for transient signals the same way as it does for the size of a speaker. Therefore this task should have been a simple case of pitch comparison. However, all five of the participants found the task difficult, reporting that they heard "diverging harmonics" that confused their ability to match the transient pitch of the signals. For example, as the PSR of the Large sphere was increased in order to make the pitch of the signal higher and the size of the sphere appear smaller, the participants described hearing two different resonances in the scaled signal: one resonance appeared to be moving upwards in pitch and while another moved downwards. But in order to complete the task the participants chose to concentrate on one of these resonances to match the scaled signal with the reference. It is

worth noting that for one subject a scaling factor was never achieved due to confusion and so an average was taken from all the other participants for that size.

The expected result was that for a smaller signal to appear larger a scaling factor of >1 would be required, with a factor of <1 needed for the larger signals, and the factors would get closer to 1 the nearer in size the comparison sphere was to the reference sphere. The medium sphere was not compared to itself and was automatically assigned a scaling factor of 1. Figure 24 shows the individual scaling factors preferred by each of the participants for each size of sphere, with table 6 below showing the corresponding values. In accordance with the comments from the participants about hearing diverging harmonics, and the difficulty with deciding upon one particular transient pitch for each sphere, there is a degree of confusion that can be seen from the results with scaling factors of >1 chosen by a couple of the participants for some of the larger spheres. In addition, two participants chose factors for the Small sphere that were lower than the factor for the X.Small sphere, contradictory to expectations. These results were more variable than expected probably due to the decisions made by the participants on which pitch to follow in the signals, and so it was decided to keep the individual results for each person to be used as parameters for part two of the experiment.

Individual Scaling Factors found for each participant

Participant	X.Small	Small	Medium	Large	X.Large
Red	0.94	0.96	1	1.013	1.036
Blue	0.75	0.96	1	0.98	1.1
Green	0.97	0.98	1	1.02	1.01
Magenta	0.97	0.92	1	0.965	0.96
Black	0.98	0.96	1	1.01	1.34
Average	0.922	0.956	1	0.9976	1.089

Table 6 The individual scaling factors for each sphere found for each participant to achieve the same pitch as the Medium sphere.

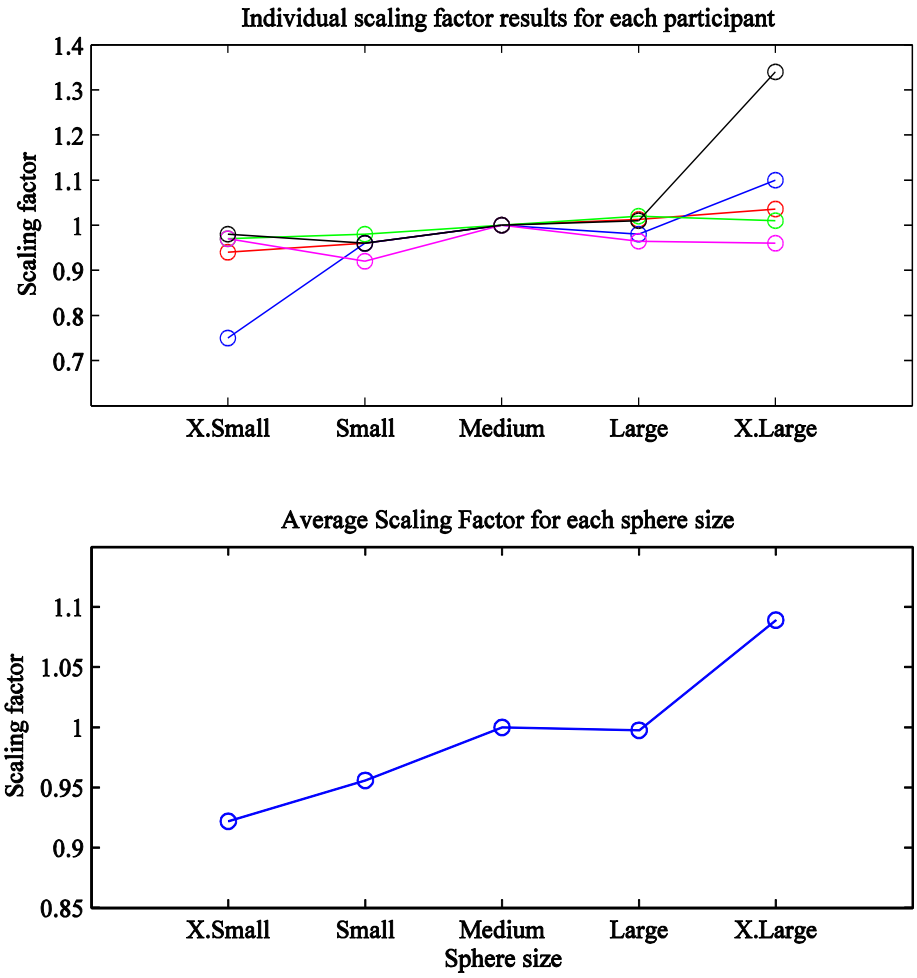

Figure 24: Individual and averaged scaling factors for adjusting the PSR of the signals to sound like the Medium sphere. Confusion while listening to the signals and hearing 'diverging harmonics' resulted in scaling factors that are unexpected.

Figure 25 shows the results of the two size discrimination tasks that make up the second part of the preliminary study. The green bars show the average percentage correct for unscaled signals, indicating the simple size discrimination abilities of the participants. Their scores are above the chance level of 50% for each paired comparison. All participants correctly judged the larger sphere when compared to the smallest, but found it more difficult to discriminate between spheres that were more similar in size.

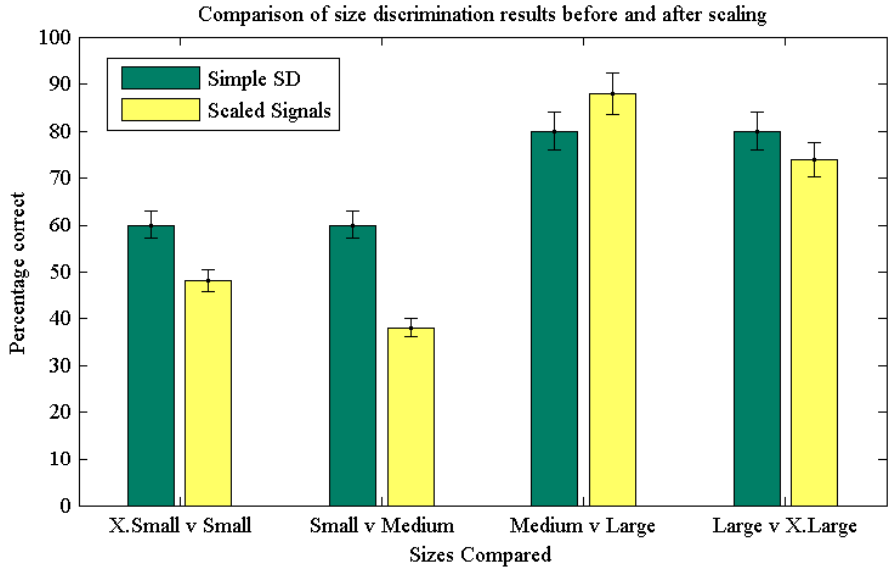

Figure 25: Average results for percentage correct for the original (green) and scaled (yellow) signals. In all cases except for the Medium vs Large, discrimination abilities decreased when the signals were scaled to have the same apparent pitch. The errorbars represent 95% confidence intervals.

The scaling factors for each sphere for each participant were used instead of an average due to the variability among the participants. Compared to the results of the unscaled task there is an overall drop in performance of 20%, apart from an increase in performance in the Medium v Large comparison. Ignoring the unscaled results, the scaled Large and X.Large signals were identifiable as larger than their comparison at a score higher than 50%, indicating that either the scaling did not work, or signals with altered transient pitch still contain size information can still be extracted using other cues that may remain. Nevertheless, these results go towards showing that with a transient signal of this type, unlike a single glottal-pulse from a vowel sound, the spectral envelope cannot simply be scaled up or down in order to change the size of the object, that there are other aspects to the transient that interferes with this scaling method. Paired sample t-tests were carried out to test for any statistical similarities between means, none were found and so no further statistical analyses were undertaken.

4.2.3 Discussion

The results of this study question the success of scaling the entire spectrum of a transient signal of this type. Firstly, the scaling method used caused considerable confusion, with the participants indicating that they could not easily compare the pitches of the signals. Secondly, the number of trials used does not give a good indication of how good or bad the size discrimination abilities of the participants are, either with or without the pitch cue. It is worthy of mention that the participants here were not trained in size discrimination beforehand; they were briefly introduced to the sound sources in order to understand what they would be hearing in the study.

In this study only five signals were used, one for each size of sphere. For the next study, different recordings of the same sphere could be used to give the participant a more realistic presentation of the sound, since it is unlikely that every time a sphere is hit, it sounds the same. Alternatively, signals could be created from an average of many recordings of the same sphere; this may give a fuller representation of the resonances in the sound. The amplitudes of the signals were not adjusted here to be equal for all signals. Instead, during the presentation of the signals, a method of 'roving amplitude' was employed. This was intended to force the listener to use other cues instead of amplitude in their size discrimination task. Even though the author listened for loudness cues the signals in next experiments were equated to the same RMS, as well as employing roving amplitude, so that they are all of equal amplitudes as has been done previously in the literature (Carello, 1998).

The final method that needs to be addressed is the chosen scaling method. Adjusting the PSR was used here to increase or decrease the perceived pitch but it was found to be difficult and confusing, and diverging harmonics created conflicting pitch cues. Perhaps a sort of interpolation method could have proven more useful. All the signals were adjusted to sound as though they were the same pitch as the medium sphere, and the medium sphere was used in the size discrimination task. As this was the only sphere not to have been scaled, it may have had some effect on how the participants listened to the comparison signal.

This does identify that shifting the spectrum up or down does not do enough to alter the perceived size, as is the case with vowel spectra, indicating that the spectra of transient signals are more complex.

5. Spectral Cues of Transient Signals

It is important to understand exactly how much transient pitch contributes to discriminating for size; it has been suggested as a cue in most of the studies in the literature. If two obviously different sized bottles are struck, and a listener asked the question "Which object is bigger and why?" the correct answer is given accompanied with the reason "because the bigger one sounds lower". While it is true that transient pitch plays an important part in size discrimination, one question posed here is just how important is it? Experiment 1 already tried out one unsuccessful method of scaling signals, but if a signal can be scaled in order to shift the spectrum higher or lower to obtain pitch ambiguity, then perhaps other important cues providing size information will be revealed; perhaps there are temporal cues other than the length of the signal itself. If other cues are to be found, it is of interest at this point in the research to understand just how important transient pitch is in size discrimination. The question we pose here is whether this pitch can be removed as a cue for a discrimination task and if so, is size discrimination still possible without this apparently important cue?

The purpose of this study is to identify if size discrimination is still possible without a pitch cue, but to also find out where the most important information lies in the spectrum. It is hypothesised that size discrimination is still possible without the frequency as a cue. This has been hinted at in experiment 1 with the Medium vs Large and Large vs X.Large scaled signals results scoring well above chance for the scaled signals. A simple clue to why this may be so is if one considers the same note on a violin and a viola being plucked, they are still perceived as different instruments, even though the fundamental frequency of the note is the same. The

following experiments will hopefully show that this theory holds true for a single pulse-resonance sound of a resonant object being struck.

The signals for the next three experiments are taken from the batch of recordings that used the metal ball bearing as the striking object. The recording method (section 4.1.1) controlled the amount of force applied to the spheres as much as possible. The large number of recordings of each sphere (~300) allowed for the creation of an average recording to be presented as a typical signal. As already mentioned, along with roving the amplitude on presentation the signals are normalised to the same RMS to remove any loudness cues, and the duration of the signals are the same limiting any temporal cues with respect to the length of the signals. The creation of the sound source has been addressed and changed to a more controllable method of a falling ball bearing. The prominence of F1 in the spectral analysis prompted the use of this as a possible method of scaling. And finally, to compliment the literature, the signals will be filtered in a number of different ways – lowpass, highpass and bandstop – in order to identify if size discrimination is possible with limited information.

The mention of SCF as a cue in the literature (Carello, 1998) is addressed in experiment 2 in a simple size discrimination of the spheres. SCF is the centre of gravity of a spectrum and so it is expected that the larger the sphere, the lower the expected centre of gravity. It was hypothesised that the greater the difference between SCF values, the fewer the number of errors in size discrimination would occur. To test for this, an extended version of the size discrimination task of experiment 1 was carried out.

The method of scaling used involves identifying the F1 value in each signal's spectrum and scaling the signal according to ratio between them; this is tested in experiment 3. The importance of F1 and F2 in speech is well known, and so experiment 4 presents signals that are filtered with more specific attention paid to F1 and F2 in order to highlight whether or not these are also important for transient signals. The scaling and filtering methods are discussed in the methods of each individual experiment. The box in figure 26 below shows all the signals, scaled and unscaled involved in the experiment.

Matrix of original and scaled signals

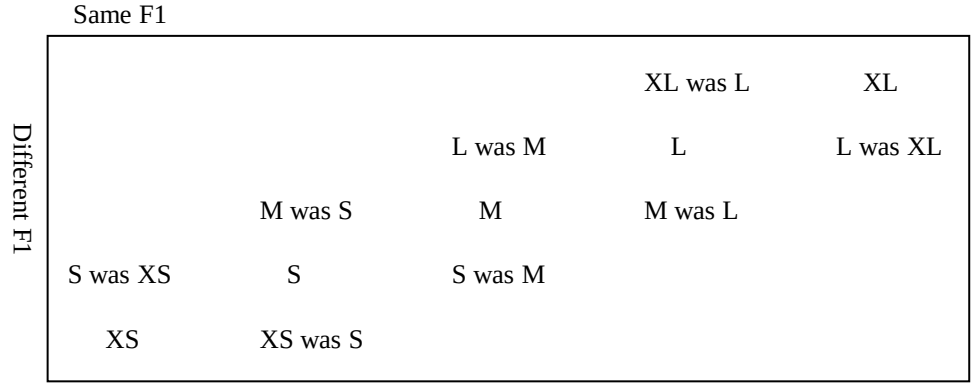

Figure 26: Matrix of signals to be used in the experiment. The horizontal axis represents signals that have the same F1, and the vertical axis represents the scaled signals that had the same original F1. In the top row, for example, 'XL was L' signifies that it was the L sphere signal, now scaled to have the same F1

5.1 Testing Procedure

The data for the experiments below were collected from one group of participants of 40 participants between the ages of 18 and 35 years, who were recruited from within the University to participate in this experiment. They completed a consent form and questionnaire before they began (see Appendix), and were paid £15 each for their participation. The questionnaire was handed back after testing for some further questions on their experience of the signals. The group consisted of 24 males and 16 females, and the average age was 26 years. One of the five participants in experiment 1 also took part in this experiment. All reported to have normal hearing at the time of the experiment. The number of participants with musical ability was 23; 17 reported no experience of learning a musical instrument of any kind.

The testing took place in a quiet room where only the participant and experimenter were present. The testing procedure was the same as used in experiment 1 (see section 4.2.1). To acquire the data for all the following experiments, the testing was split into five 30 minute sessions which took place over a number of weeks, with at least one week between each session to avoid any learning effects. The 30 minute long testing sessions were broken up into sets

according to types of signals presented, and each pair was presented to the participant a number of times within each set. The testing sets were short, approximately five minutes per set depending on how often the participant chose to repeat the pairs they heard. At the start of the session, the participant was allowed to adjust the volume of the signals in order to be at a comfortable listening level for them. The maximum level of the signals was controlled by MATLAB, and the maximum noise exposure for each participant per session was under 65 dB (A), well within the guidelines for noise exposure (see appendix). All experiments were carried out under the approval of the Human Experimentation Safety and Ethics Committee, Institute of Sound and Vibration Research.

5.2 Experiment 2 – Spectral Centroid Frequency

A simple size discrimination task was carried out to confirm the participants' abilities to tell the sphere signals apart. The test also served as a method of judging the importance of the SCF cue. The set-up for this experiment was as stated above. In testing sets, only neighbouring pairs were compared to each other, i.e. X.Small v Small, Small v Medium, Medium v Large, and Large v X.Large. Test sets involved signals taken from the batch of averaged signals, or signals taken at random from the library of recordings. There were eight presentations of each size from the random signals and six presentations from the averaged signals heard by each participant.

The follow is hypothesised:

"Participants will perform above chance in a simple size discrimination task of non-periodic stimuli."

"Spectral centroid frequency differences will have an effect on amount of errors made in size discrimination."

It is expected that due to the nature of the sounds chosen, the participants would show good abilities to discriminate for size. The spectral centroid frequency of each signal is expected to be lower for larger signals, and that the difference between SCFs of compared signals would affect performance.

5.2.1 Results

To test these hypotheses, a simple size discrimination task was carried out involving comparisons between random signals taken from the library of 1500 recordings, and comparisons between averaged signals for each size of sphere. The red arrows in figure 27 below show the signals which are compared in this basic test for size discrimination ability. Each unaltered signal is compared with its neighbour in order to determine the SD abilities of the participants for the spheres used in this study.

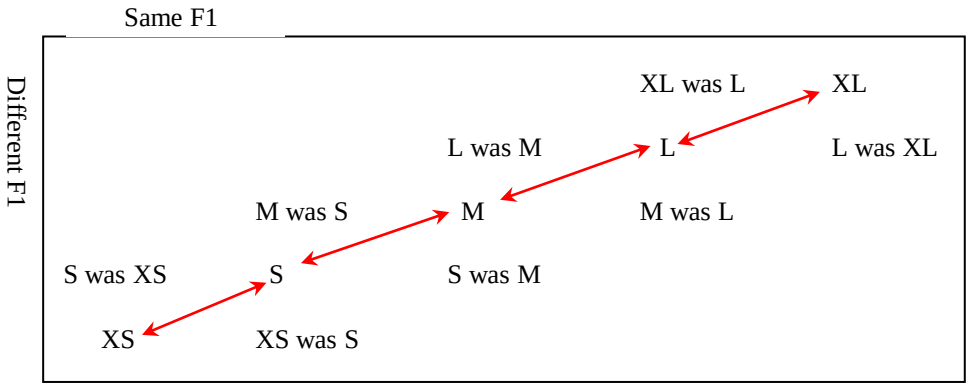

Figure 27: Chart indicating the signals used for this simple size discrimination task. Only neighbouring signals were compared, marked with red arrows.

The results here are from data acquired by 19 participants; the other participants either failed to complete all the rounds of testing or were deemed unreliable (see section 6.1 for reliability tests). In order to see the effect of differences between SCF values of the signals, the SCF was calculated using the equation described in 2.3.3. The resulting values are plotted in figure 28, and the differences between them are visible. With increasing size of sphere, there is a decrease in SCF value. The biggest difference between SCF occurs in the X.Small and Small spheres, and the smallest is between Large and X.Large.

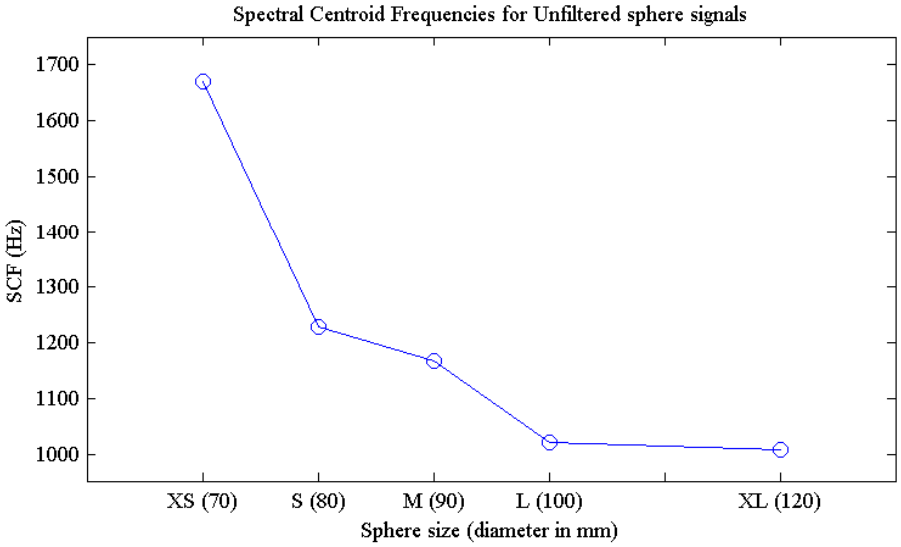

Figure 28: Spectral Centroid Frequencies (SCF) for each of the five different sized spheres, calculated from the averaged signals of each. As the spheres increase in size, the SCF decreases in value. There is a bigger difference between the X.Small and the Small spheres than any other pair.

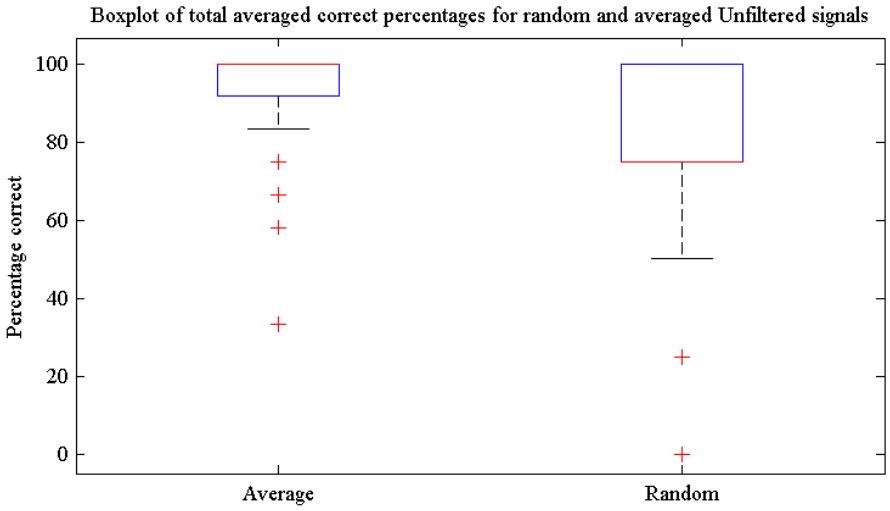

Figure 29: Results for the simple size discrmination test. The boxplot on the left shows the results for the averaged signals. The right boxplot shows the total results for the signals chosen at random from the library of signals. Participants performed significantly better when listening to the averaged signals, but results are well above chance for both sets indicating good size discrimination abilties overall.

The boxplot above contains the data for unfiltered signals, grouped into random signals and averaged signals. Shapiro-Wilk test of normality shows that neither set is normally distributed ($p>0.05$). For each set of comparisons, participants performed

significantly above the chance level of 50% (one-sample T-test: Random $t(18)=10.016$, $p<0.05$, $r=0.921$; Averaged $t(18)=22.6$, $p<0.05$, $r=0.983$) consistent with the hypothesis. The mean for the Random signals was 80.62 % (SD ±13.3), and for Averaged signals the mean was 90.8 % (SD ±7.9). A Wilcoxon signed-rank test indicates that the mean score for the averaged signals is significantly higher ($p<0.05$) than signals taken at random from the library of recordings, probably due to the variability between the individual recordings. The results were then analysed in order to compare individual comparison pairs with the differences between the SCF values for each signal. The graph below shows that the biggest difference in SCF values occurs between the X.Small and Small sphere signals (442.7Hz), with the next biggest difference occurring between the Medium and Large sphere signals (145.2Hz). The difference between Small and Medium is 61.5Hz, and the smallest difference is only 13.1Hz between the Large and X.Large sphere signals.

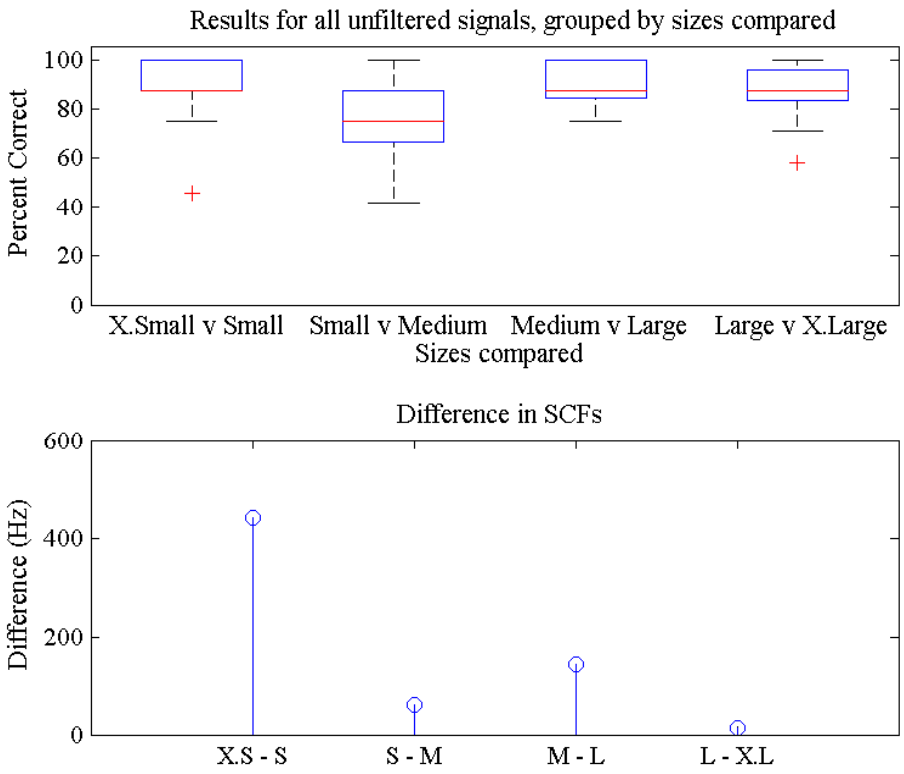

Figure 30: The results for the unfiltered size discrimination task compared with the differences between SCFs for neighbouring sizes of sphere. There is more variance and lower mean scores for comparisons with the smallest difference in SCF.

5.3 Experiment 3 – Scaled Signals

5.3.1 Method

The longest sets were those involving scaled signals (15 minutes) and so this set contained specific comparison pairs order for the researcher to ensure the participants were not losing concentration. These were unfiltered average signals, which the participants had already shown their ability to discriminate between in the first round of testing. The results of these pairs went towards indicating the reliability of the participants, and they were also included in the overall results.

Experiment 1 attempted a way of subjectively scaling the signals, presenting the a signal to the participants and adaptively increasing or decreasing the PSR of the comparison signal to the point at which the participant deemed the signals to have the same transient pitch. Although it involved very few participants, they each encountered difficulties with this, saying they could not just pick out one "pitch" to listen to for their comparison. Due to the difficulties in subjective scaling, and since it was observed that the first resonance in each signal was more than 10 dB louder than any of the other resonances, it was decided that a more objective approach to scaling would be to use this first resonance as a value to create ratios between the signals. These ratios would be the factors by which the signals would be altered.

Table 7 shows the frequency increase and decrease factors between first resonances of the signals. Signals were only scaled to their neighbouring sizes, so as to avoid any distortion due to large scaling factors; i.e. the Large sphere signal was scaled to both the Medium and the X.Large sphere, but the X.Small sphere signal was only scaled to the Small sphere. As mentioned above, the signals were only scaled to the resonances of their neighbours.

Size 1	F1 Hz	Size 2	F1 Hz	Size 1→2	Increase factor	Size 2→1	Decrease factor
X.Large	3531	Large	4048	XL → L	1.1464	L → XL	0.8723
Large	4048	Medium	4910	L → M	1.2129	M → L	0.8244
Medium	4910	Small	5340	M → S	1.0876	S → M	0.9195
Small	5340	X.Small	5857	S → XS	1.0968	XS → S	0.9117

Table 7 The frequency values of the first resonances of neighbouring signals and the increase and decrease factors by which they would be multiplied to create the scaled signals.

Signals were scaled in MATLAB as follows:

1 - The existing sample rate was multiplied by the scaling factor (see Table 7) to create a new sample rate.
2 - A vector of new samples was created using this new sample rate and the length of the existing signal.
3 - The existing signal was then interpolated to fit into the newly created sample points to create the scaled signal.

This scaling method ensured that length of the signal was neither stretched nor compressed, as it would have been if the scaling method used was simply altering the play-back sample rate. This would have compromised the experiment as the listeners could possibly have heard any difference in the length of the signal. Creating a new signal as described above ensured that only the frequency spectrum was altered. Interpolation in this sense is a method of calculating the value of a non-existing sample using the values of the existing samples either side, and thereby creating a new sample. As the original sample rate was 44.1 kHz and the scaling factors are relatively small (between 0.82 and 1.21), any distortion or error introduced by interpolation is minimal.

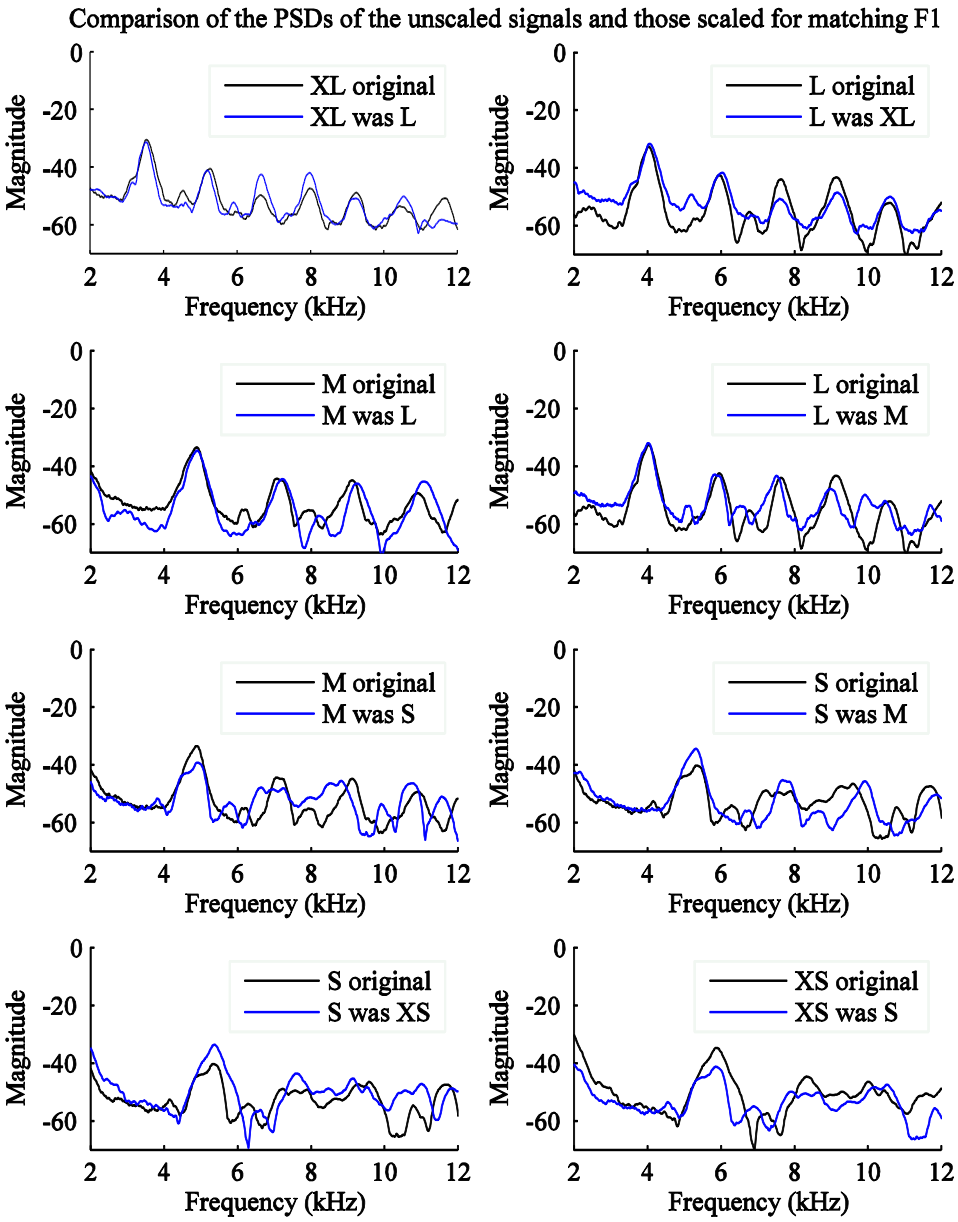

Figure 31: Comparisons of the Welch PSDs of the original signals with the signals that were scaled. The black lines represent the original un-scaled signals, and the blue lines are the signals scaled to have the same first resonance. These plots are zoomed to a resolution of 2-12 kHz in order to see more clearly how the newly scaled signals now have first resonances very similar in value to those to which they are scaled.

From Figure 31 it is clear that the scaling method used above aligns the frequency values of the resonances well. The level differences and the distances between the resonances remain unchanged, and again the length of the signal was not altered. For these reasons it was concluded that the primary frequency cue was effectively removed, that being the first resonance as opposed to any subsequent others, due to the first being 10dB louder than the rest. Thus, if any discrimination abilities were to appear from using these signals it would be due to the listener using cues other than frequency to tell the signals apart.

The following two points for Experiment 3 were hypothesised:

3A : 'When signals are scaled to have the same F1, participants will correctly identify the larger signal by using other cues.'

3B : 'When signals are compared to scaled versions of themselves, the value of F1 will have a significant effect on the results.'

5.3.2 Exp. 3A Results

Scaled versions of all the signals were created by resampling them according to the ratios between the F1 values. The signals were scaled to either the F1 of next size smaller or larger than them to prevent artefacts from over-scaling. The signals with the same F1 values were compared in a test for the success of the scaling method, and to expose the possibility of other size discrimination cues. Signals were compared with scaled versions of themselves in a test for the importance of the spectral cue; i.e. if the participants only listen to transient pitch, they will be fooled by the scaled signal and choose the signal with the largest F1 regardless of its unscaled value. Each scaled signal appeared 12 times to each participant in total.

The chart below shows the pairs in which the signals had matching F1s. The blue line compares one size with the size just smaller which is scaled to have the same F1, i.e. signal versus signal-was-smaller. The magenta line compares one size with the size just bigger than it which is also scaled to have the same F1, i.e. signal versus signal-was-bigger. Finally, the orange line in this group compares the signal-was-smaller with the signal-was-bigger where, again, both have the same F1. It is

proposed that participants will correctly identify which signal comes from the larger sphere because cues other than frequency and F1-matching will reveal the original size of the sphere.

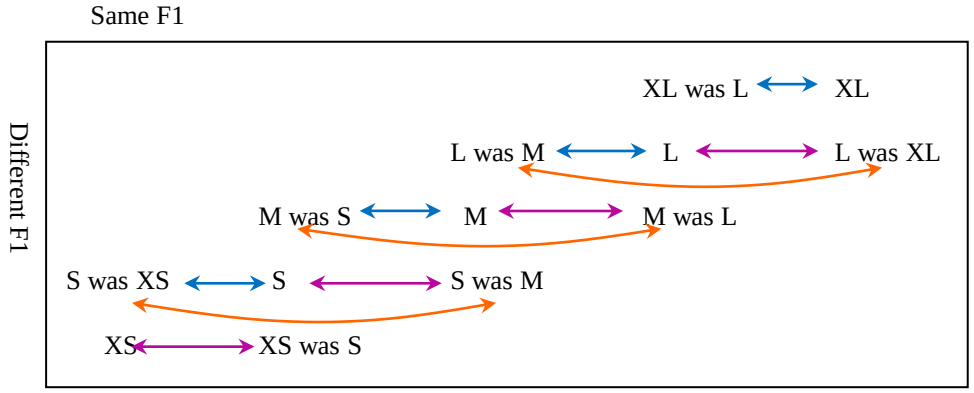

Figure 32 This chart from before now shows the groups of pairs that are used in Exp. 3a. Each original signal is compared with its neighbouring signals which have been scaled to have the same F1. The blue line compares the signal with its previously smaller neighbour, and the magenta line compares the signal with its previously bigger neighbour. The orange line compares the signal-was-smaller with the signal-was-bigger.

A Shapiro-Wilk test for normality shows that none of the data sets are normally distributed ($p>0.05$). All results sets are found to be significantly above or below chance level using a one-sample t-test (Blue-unscaled: $t(18)=-7.195$, $p<0.05$, $r=0.86$; Blue-scaled: $t(18)=7.195$, $p<0.05$, $r=0.86$; Magenta-unscaled: $t(18)=9.798$, $p<0.05$, $r=0.9177$; Magenta-scaled: $t(18)=-9.798$, $p<0.05$, $r=0.9177$; Orange-scaled-from-bigger: $t(18)=-16.2$, $p<0.05$, $r=0.967$; Orange-scaled-from-smaller: $t(18)=16.2$, $p<0.05$, $r=0.967$). Means comparisons are made using a repeated-samples Wilcoxon Signed Rank test. The blue line compares a signal with its smaller neighbouring signal which is scaled to have the same F1, and so according to the hypothesis, the participants should correctly identify the unscaled signal as the bigger signal. However, the boxplots in Figure 33 show that a significantly larger percentage of answers indicate the 'scaled from smaller' signal to be perceived as the bigger signal. In addition to this, the magenta line also showed results contrary to the hypothesis. The comparisons made here are a signal with its bigger neighbour, again scaled to

have the same F1. The results indicate a significantly larger percentage of answers show the unscaled signal as the bigger signal, instead of its 'scaled from bigger' neighbour. The results for the orange line supports the preference towards signals scaled from the size below, where a significantly larger proportion of answers indicate the signal 'scaled from smaller' to be larger that which is 'scaled from bigger'. Despite the similarities in the ratios between F1 and F2 in each signal, this suggests that aligning F1 was not sufficient for scaling the signals to have the same transient pitch.

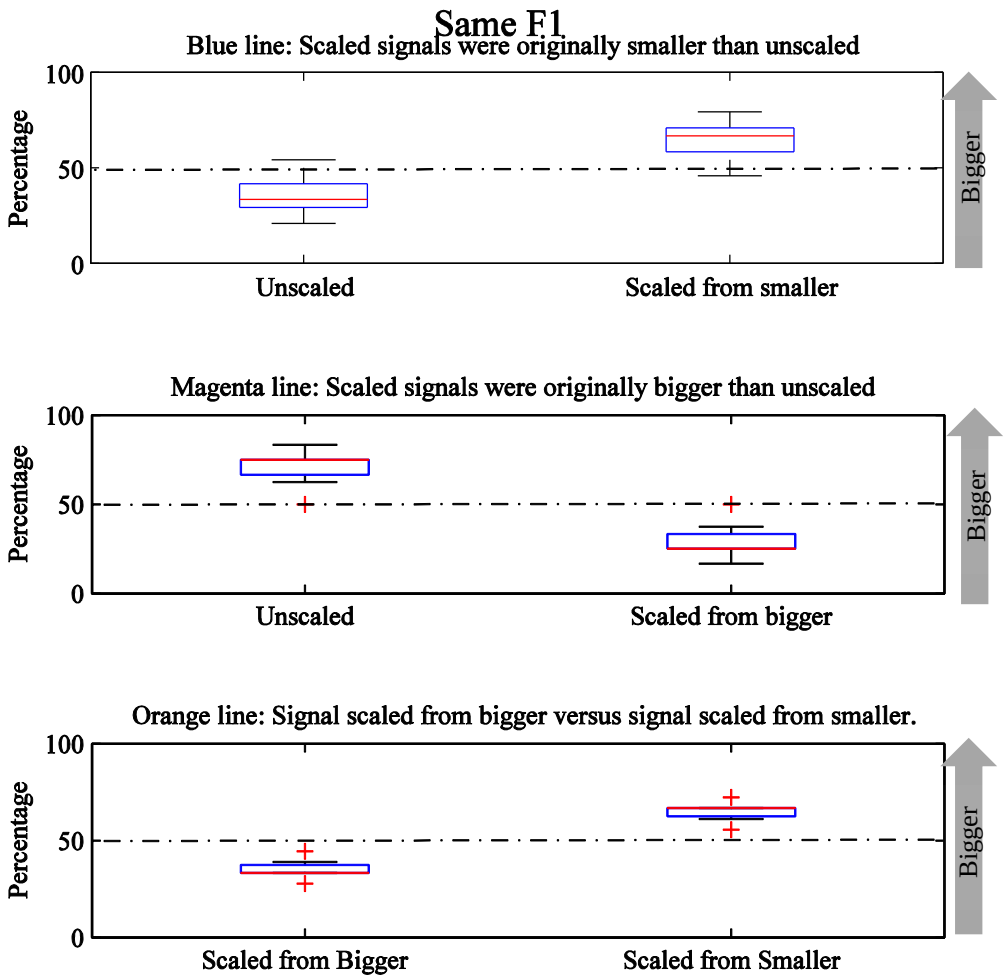

Figure 33: Boxplots of results from the Same F1 SD tasks of Exp. 3A. The top box shows the results for the blue line, where participants choose the signal 'scaled from smaller' to sound bigger than the unscaled bigger signal. The middle box shows magenta line, where a signal is compared with its 'scaled from bigger' neighbour. Participants choose the unscaled smaller signal to sound bigger. The

lowest box is a comparison of the 'scaled from bigger' and 'scaled from smaller' signals, both resulting in the same F1. Participants favour the 'scaled from smaller' signal as that which sounds bigger.

5.3.3 Exp. 3B Results

The second hypothesis compares each signal with a scaled version of itself, and assumes that participants will be fooled by the spectral cues given, and will choose the signal with the lowest F1 as the larger signal. The chart below indicates the pairs which are used in this 'Different F1' task. The light blue line compares the unscaled signal with a version of itself that is scaled for a lower F1. The yellow line pairs up the unscaled signal with the version that is scaled for a higher F1. Finally, the green line compares the two scaled signals (scaled higher and scaled lower), both of which were originally the same signal. Each scaled signal in this set appeared eight times to each participant in total.

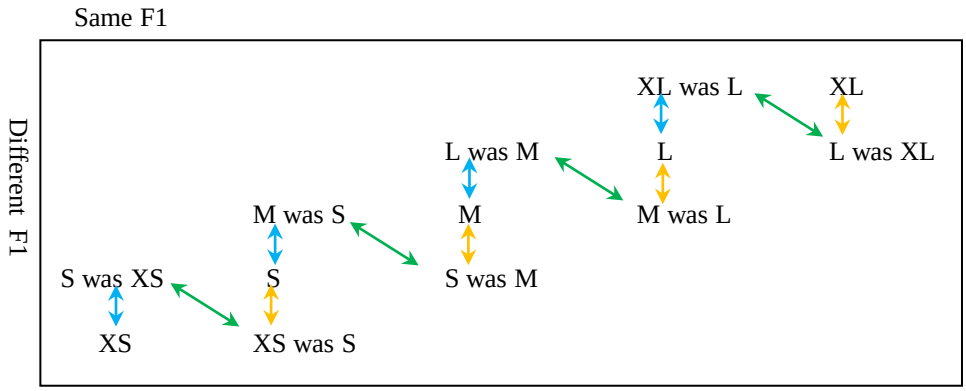

Figure 36: The chart above shows the groups of pairs that are used in the 'Different F1' SD task of Exp. 3B. Each original signal is compared a version of itself that is scaled for either a lower F1 (light blue line) or a higher F1 (yellow line). The green line compares a signal which is scaled for a lower F1 with the same signal that is scaled for a higher F1.

Apart from the data set corresponding to the Yellow line, Shapiro-Wilk tests for normality shows that none of the data sets are normally distributed ($p<0.05$). Means comparisons are made using a repeated-samples Wilcoxon Signed Rank test,

except for the Yellow line where a paired-samples t-test was carried out. All results are found to be significantly above or below chance level using a one-sample t-test (LightBlue-unscaled: $t(18)=-30.32$, $p<0.05$, $r=0.99$; LightBlue-scaled: $t(18)=30.32$, $p<0.05$, $r=0.99$; Yellow-unscaled: $t(18)=10.552$, $p<0.05$, $r=0.9278$; Yellow-scaled: $t(18)=-10.552$, $p<0.05$, $r=0.9278$; Green-scaled-Lower-F1: $t(18)=13.449$, $p<0.05$, $r=0.954$; Green-scaled-Higher-F1: $t(18)=-13.449$, $p<0.05$, $r=0.954$).indicating that participants choose the signal with the lower F1 as that which sounds bigger. The signals scaled to have a lower F1 (light blue) shows the most consistency across participants and less variation in the results, with an average of 94.5% of the results showing a preference for the signal that is scaled for a lower F1 as sounding bigger. The signals scaled to have a higher F1 (yellow), while having slightly more variation in the results, also shows a significantly larger percentage of participants choosing the signal which sounds bigger; in this case it is the unscaled signal, as the scaled signal is altered to have a higher F1. Lastly, the data corresponding to the green line compares two signals which were originally the same signal, one scaled higher and the other scaled lower. The boxplots below again show a significantly larger percentage choosing the signal that is scaled to have a lower F1 as the signal that sounds bigger.

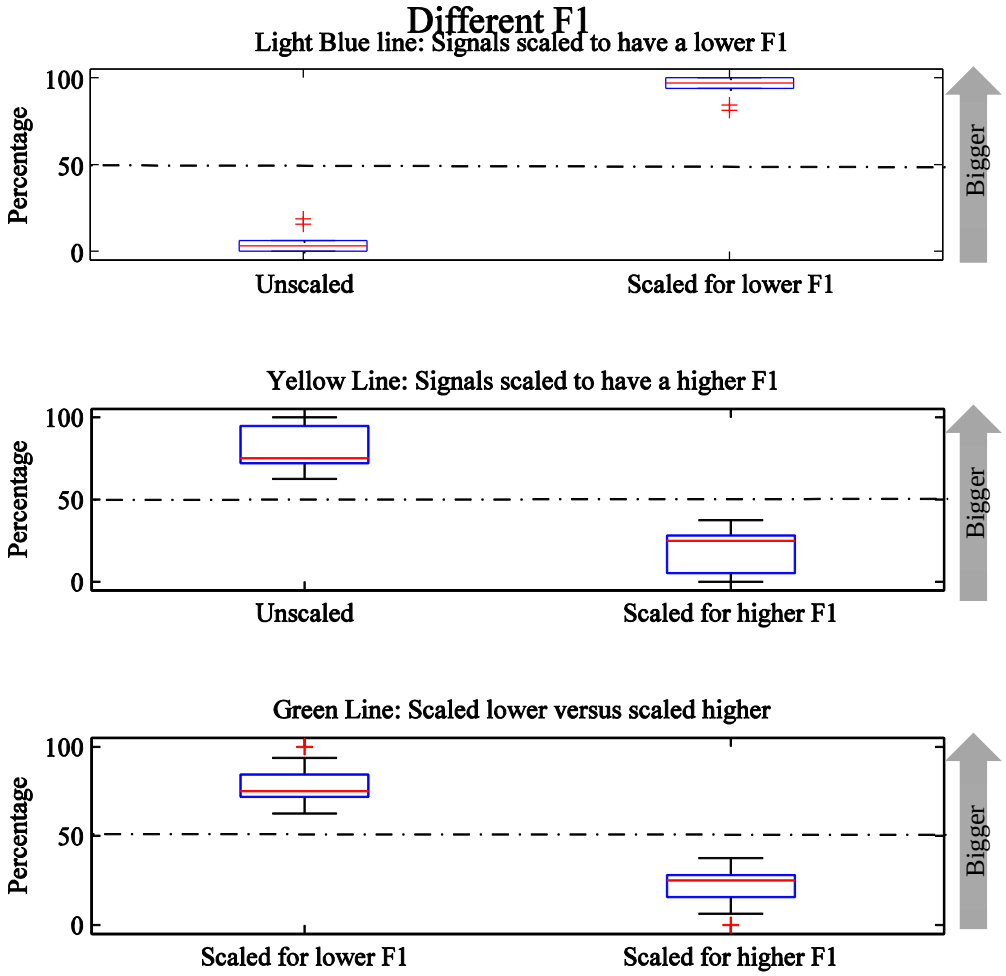

Figure 34: Boxplots of results from the Different F1 SD tasks of Exp. 3B. The light blue line at the top compares a signal with a version of itself that is scaled for a lower F1. The yellow line in the middle then compares a signal with a version that is scaled for a higher F1. Finally the green line boxplots show the results for signals that were originally the same signal but are now scaled for either a lower or higher F1.

5.4 Experiment 4 – Filtered Signals

5.4.1 Method

As well as using scaled signals in this experiment, the signals were filtered in several ways in order to find if some parts of the spectrum of a signal were more important that others in discriminating for size. The importance of F1 and F2 is already known for vowel sounds, and here the signals are filtered in order to

highlight if it is the same case for transient sounds. The signals were filtered through Adobe Audition, all using a 4th-order Butterworth filter except for the first high-pass filter mentioned below due to the need for a steeper roll-off. The signals were filtered as follows:

- Low Pass filtered
    - c/o 5% lower than first resonance
    - c/o 5% higher than first resonance
- High Pass filtered
    - c/o 5% lower than first resonance
    - c/o 5% higher than first resonance
- Band-stop filtered
    - F1 removed
    - F1 and F2 removed

In the literature, Grassi (2002) has manipulated the wooden ball recordings by applying low-pass and high-pass filters both with cut-off frequencies at 5 kHz. This seemingly arbitrary cut-off point was discussed as emulating how sounds are low-pass filtered when they are heard through a wall, for example. No explanation was given for the exact reason behind the cut-off frequency of 5 kHz. In this study, however, low-pass filtering the signals below F1 will show up if there is any size information contained within the wide band of low frequencies that appear to be very similar in the PSDs. Filtering above the F1 value will show the importance of F1 for size discrimination. Figure 35 below displays the power spectral density plots for both the No-F1 and With-F1 LPF signals.

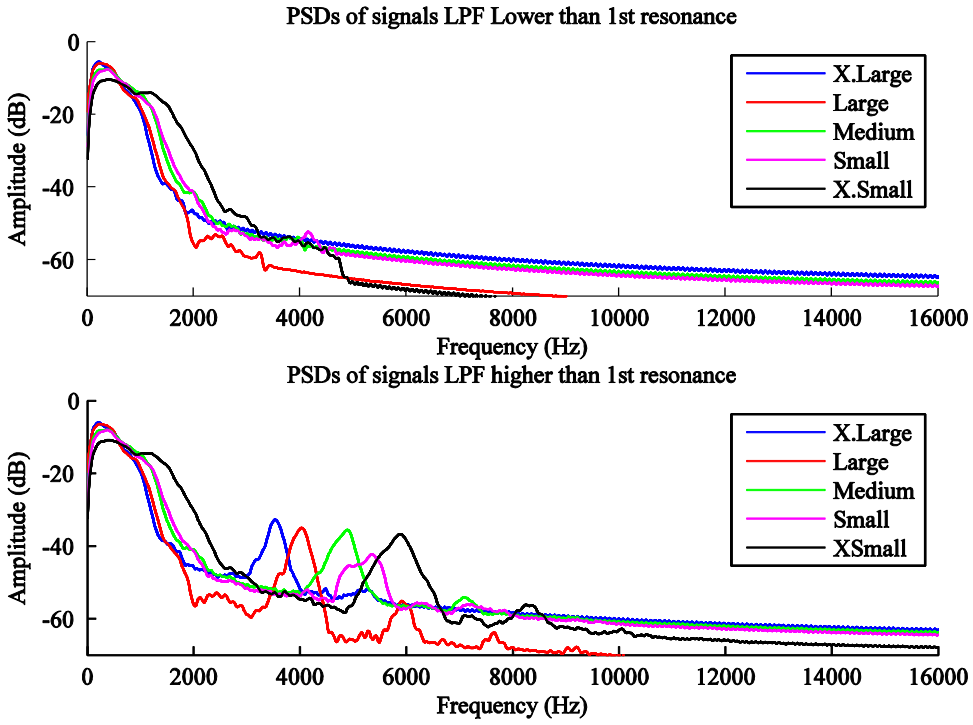

Figure 35: Power spectral density plots of the low-pass filtered sphere signals, the top panel shows the signals filtered with the c/o at 5% below F1, and the lower panel includes F1 with the c/o frequency at 5% above F1.

Again, in the literature, Grassi (2002) high-pass filters his signals at 5 kHz with no given reason for the value. Here, the signals are high-pass filtered with cut-off frequencies of either 5 % above or below F1 in each spectrum. The frequency resolution of the auditory system weakens with increasing frequency, so it is no surprise that for Grassi the results of the high-pass filtered discrimination task were lower than the other results. However, high-pass filtering the signals in the method shown here will demonstrate if F1 and F2 are audible as size cues when the wide-band of low frequency energy is removed. Figure 36 shows the power spectral density plots of the No-F1 and With-F1 HPF sphere signals.

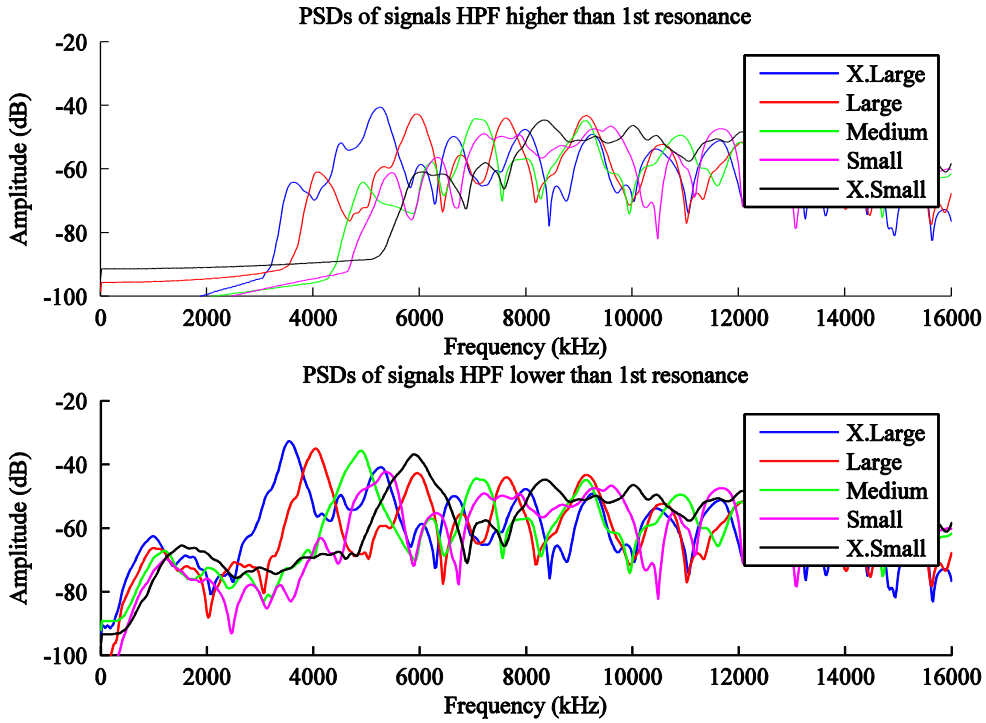

Figure 36: Power spectral densities of the high-pass filter sphere signals. The top panel is filtered with a c/o frequency of 5 % above F1, to test if listeners can hear and use F2 as a size cue. The lower panel has a c/o frequency of 5 % below F1.

The last type of filter applied to the signals was band-stop filtering of which there were two types: F1 removed, or F1 and F2 removed. The removal of this specific range of the spectrum would conclusively identify if the perception of formants in transient signals was similar to how much information F1 and F2 provides in speech sounds. Figure 37 shows the power spectral density plots of these signals and while there is still plenty of spectral information remaining in each spectrum, the resonances with the largest peaks have been removed.

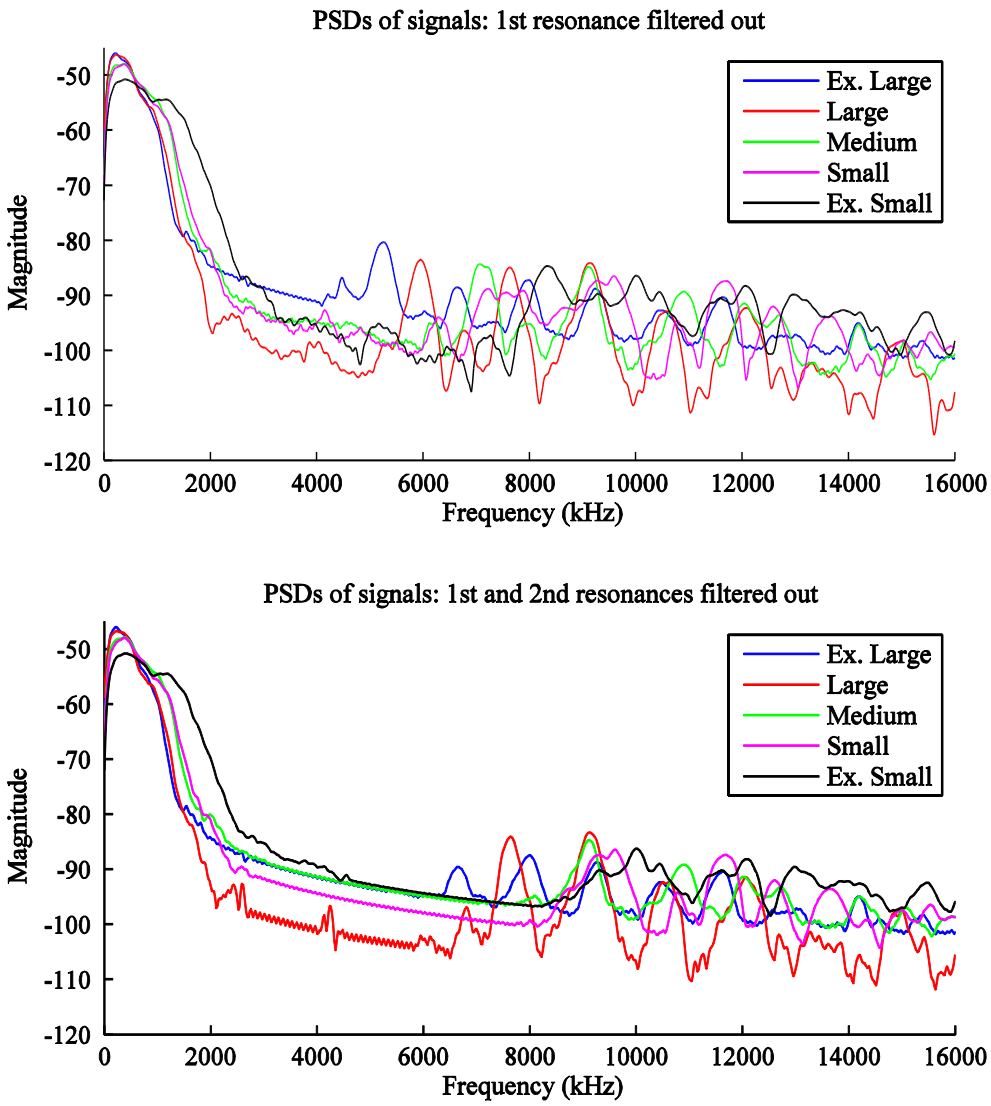

Figure 37: Power spectral density plots of the band-stop filtered sphere signals. The top panel shows F1 has been filtered out, and the lower panel no longer has bother F1 and F2.

Of the filtered signals experiment, the following were hypothesised:

'There will be a significant effect of all types of filtering; low-pass, high-pass and band-stop, on size discrimination abilities.'

'Removing F1 from signals will have an effect on size discrimination.'

5.4.2 Results

To identify if there are areas in the spectrum that are more important for size discrimination, and also to test the robustness of size discrimination abilities, the signals were filtered using low-pass, high-pass and band-stop filters. Two different cut-off frequencies were used for the low-pass and high-pass in order to include F1 or filter it out. Band-stop filtered signals filtered out either F1 alone, or both F1 and F2. The number of presentations of each signal to each participant was as follows:

Type of signals in the set		Total no. of presentations
Low-pass filtered	With F1	12
	Without F1	8
High-pass Filtered	With F1	12
	Without F1	8
Band-stop Filtered	Without F1	8
	Without F1 & F2	8

Table 2: The total number of presentations of each signal in experiment 4.

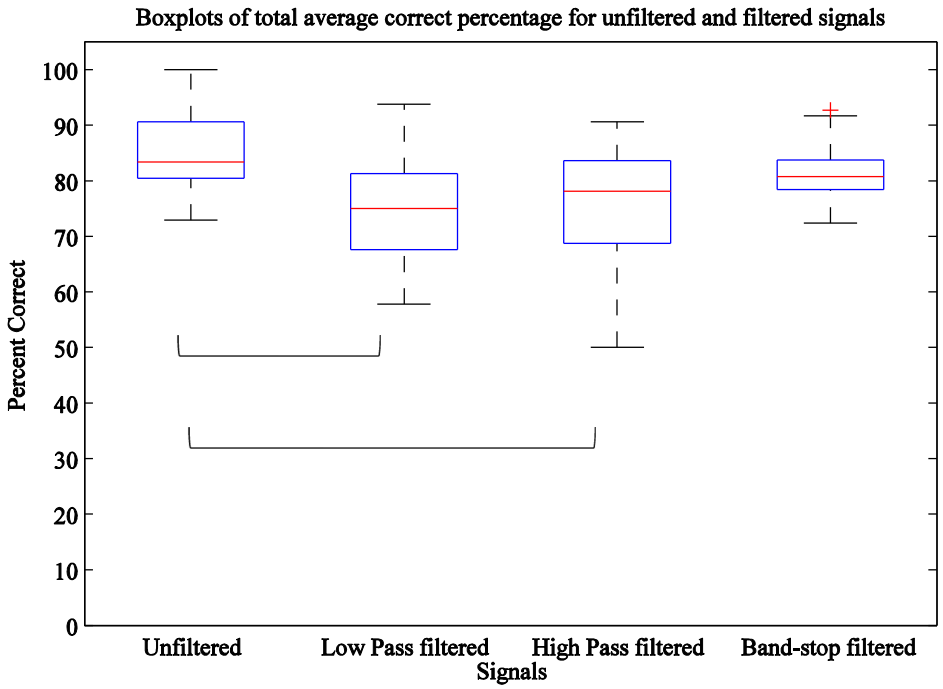

Figure 38: Boxplots of results from total UnF, LPF, HPF, and BSF signals from all 19 participants. The blue box indicates the interquartile range while the dashed black line shows 1.5 times this value. The red line shows the median for each

range of data and the red crosses reveal any outliers in the data. LPF and HPF were significantly lower than Unfiltered ($p<0.05$). There is no significant difference between UnF and LPF, or between UnF and BSF. Note- the BSF results include both no-F1 and noF1-F2 results.

The unfiltered (UnF) results here are taken from the combined totals for the SD tasks using random signals and the averaged signals. The low-pass (LPF), high-pass (HPF), and band-stop (BSF) boxes contain the results from the Without F1 signals only. All sets of data shows equal variance (Levene's test; $p>0.05$) and a Shapiro-Wilk test shows all sets are normally distributed ($p>0.05$). The results, as seen in the boxplots above, show that the participants are capable of carrying out the simple size discrimination task, with UnF, LPF, HPF and BSF all showing mean scores to be significantly above the chance level of 50%. A repeated-measures ANOVA with a Greenhouse-Geisser correction method ($F(2.414,43.443)=23.844$, $p<0.0005$) showed there to be a significant difference between the four filter types. Bonferroni post-hoc test shows that there is a significant effect of high-pass filtering the signals ($p<0.0005$). There is no effect on the results by applying low-pass or band-stop filters. There is a significant difference between the scores for LPF and HPF ($p<0.01$), and HPF and BSF ($p<0.0005$).

The results above include both with F1 and without F1 filtered signals. From the PSDs in section 4.1.4, it is clear the F1 was the strongest resonance in each of the signals, and so the next test investigates the importance of F1. Within filter categories, signals are grouped so that one set has F1 filtered out using a band-stop filter which included signals with F1 and F1&F2 filtered out (Stop-F1, StopF1F2), one is low-passed filtered with a cut-off frequency below F1 (LPF no F1), and another is high-pass filtered with a cut-off frequency higher than F1 (HPF no F1). The results analysing the two types of band-stop filtered signals separately follow these current results.

The boxplots below show the results within these filter categories. All data sets show equal variance according to Levene's test ($p>0.05$), and the Shapiro-Wilk test for normality shows the HPF data set is not normally distributed ($p<0.05$). Paired t-

tests were carried out for the LPF and the BSF sets, and a Wilcoxon test for the HPF set. The results show a significant effect of filtering out F1 for all results sets. A repeated measures ANOVA was carried out on the data below and a Greenhouse-Geisser correction method showed there to be a significant difference between groups ($F(3.031,54.561)=20.597$, $p<0.0005$). Bonferroni post-hoc tests compared data within the filter groups. The decrease in scores between UnF and Stop-F1 is not significant. Removing F1 from the LPF signals results in a significant decrease in mean scores ($p<0.0005$) compared to LPF with F1, and the difference between UnF and LPF No F1 is also significant ($p<0.05$). The opposite occurs when removing F1 from the HPF signals where the participants appear to improve significantly in their SD abilities ($p=0.01$). All HPF scores were significantly below the UnF mean (HPF With F1, $p<0.0005$, HPF No F1, $p<0.01$). All scores are well above chance for this test of importance of F1, indicating that even without F1 size discrimination is possible.

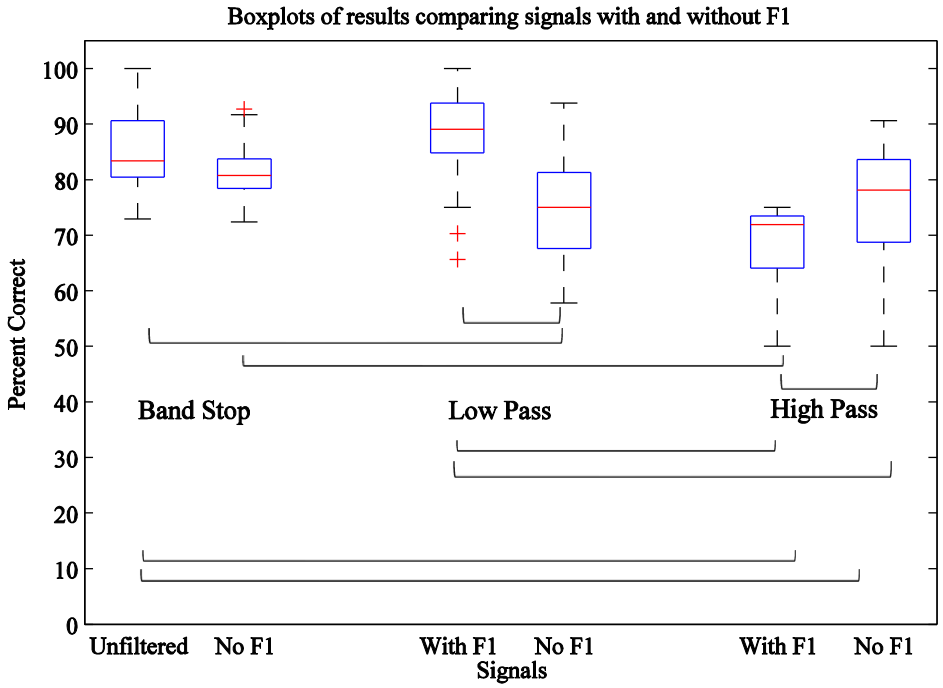

Figure 39: Results from the Effect of F1 on Size Discrimination. Removing F1 from unfiltered signals shows no significant decrease in mean scores (These results include StopF1 and StopF1F2 data). However, there is a significant decrease in

ability in LPF signals after F1 is removed, as well as UnF signals. The reverse happens when F1 is removed from HPF signals with an increase in mean scores from HPF With F1 to HPF No F1. The horizontal brackets in the figure indicate significant differences in the means.

The No F1 results above included both StopF1 and StopF1F2 data. Here, the data is separated for further analysis. The boxplot below shows the comparison between UnF, and BSF without F1 (BSF no F1) and without F1 and F2 (BSF noF1-F2). Levene's test shows the BSF noF1-F2 data set to be of equal variance ($p>0.05$), and the Shapiro-Wilk test shows it to be normally distributed ($p<0.05$). From before it is known there is no significant decrease in the mean scores by applying a band-stop filter on the signals, but the results above included both No F1 and No F1F2 signals. The results below show the mean scores for UnF compared with BSF No F1 and BSF No F1F2 signals. A repeated-measures ANOVA shows there to be a significant difference between scores ($F(1.825,32.858)=4.671$, $p<0.05$). A Bonferroni post-hoc test shows the mean score decreases significantly when the BSF No F1 was applied to the signals ($p<0.05$). Surprisingly, there is a significant increase in means from BSF No F1 and BSF No F1F2 ($p<0.05$), despite two of the most prominent resonances and presumably the most informative frequencies having been filtered out. There is also no significant difference between the means of the UnF and the BSF No F1F2 data sets.

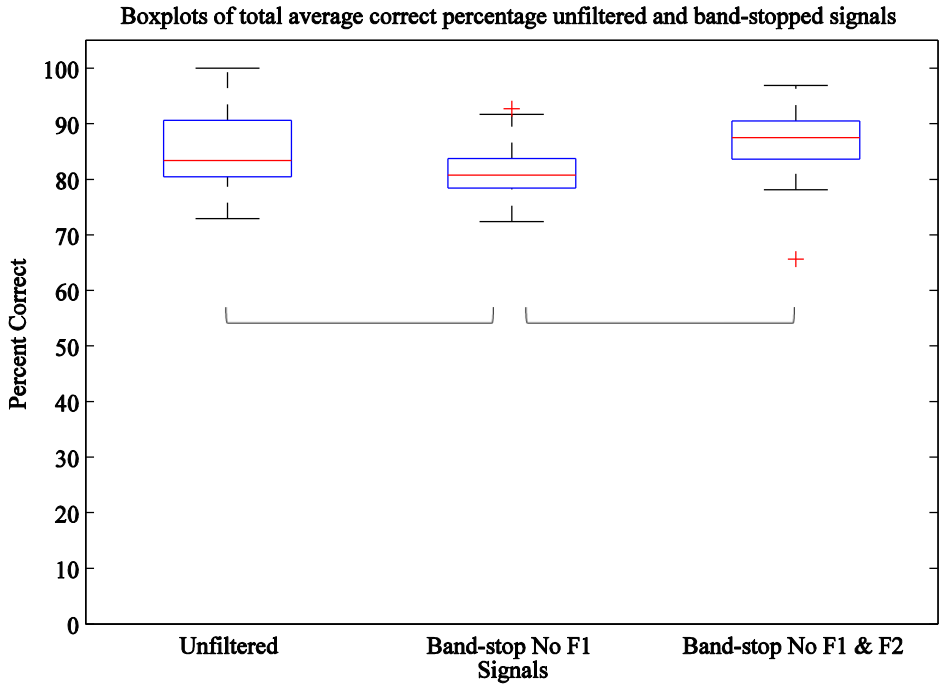

Figure 43: Boxplot of percentage correct answers for unfiltered signals compared with BSF no F1 and no F1-F2. The brackets indicate significant differences between data sets.

With respect to the SCF of the signals, the effect of comparison pair involving all the different types of filtered signals was also analysed. Figure 44 below shows all the SCFs for the different types of filtered signals used in the experiment. In almost all cases, the SCF decreases with increasing size, except for the LPF with F1 signals between Small and Medium. The trend for larger differences between X.Small and Small, and Medium and Large seems to hold, with this being reflected in the data with fewer errors in discrimination and less variance.

All scores are significantly above chance for the pairs compared, which includes the pairs in each of the following filter categories and sub-categories: UnF, LPF (with F1 and without F1), HPF (with F1 and without F1), and BSF (without F1 and without F1 & F2). The data shows equal variance based on the mean (Levene's test; $p>0.05$) and a Shapiro-Wilk test shows the Small v Medium and Large v X.Large sets to be normally distributed ($p>0.05$).

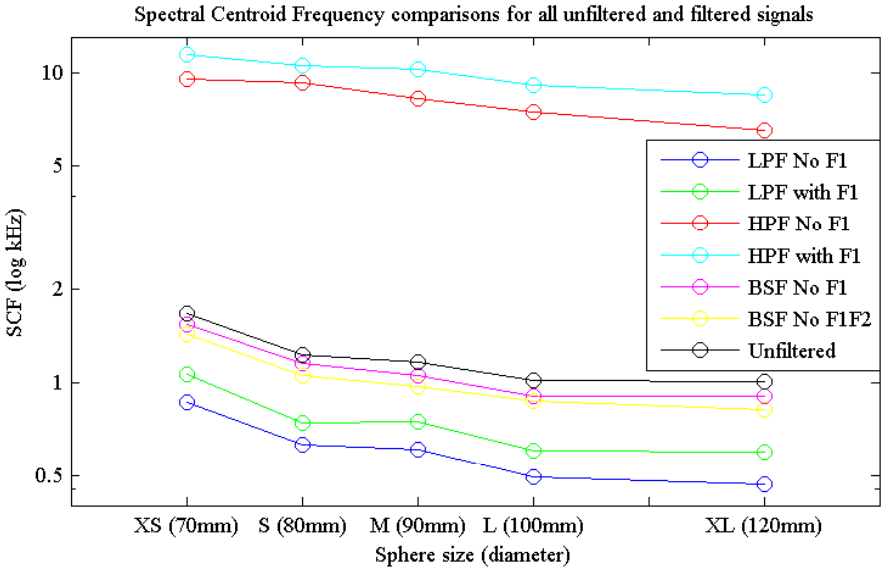

Figure 44: SCF values for each type of filtered signal. The trend for decreasing SCF value with increasing size is visible.

A repeated measures ANOVA with a Greenhouse-Geisser correction indicates there is a significant difference between the groups ($F(2.035,36.632)$, $p<0.0005$). The results of the Bonferroni post-hoc tests are shown in the boxplots of figure 45, and it appears that spheres that have the largest differences between SCF have the highest scores: i.e. X.Small v Small, and Medium v Large, and the lowest scores go to the comparisons that have the lowest SCF: i.e. Small v Medium, and Large v X.Large. Paired samples t-tests show that the means for the high-scoring sets are significantly higher than those of the low-scoring sets (XSvS and SvM: $p<0.0005$; XSvS and LvXL: $p<0.0005$; MvL and SvM: $p<0.0005$; MvL and LvXL: $p<0.0005$). This positively enforces our hypothesis that larger differences between SCFs leads to easier discrimination between signals. There is no significant difference within the high-scoring or the low-scoring sets (XSvS and MvL: $p>0.05$, SvM and LvXL: $p>0.05$).

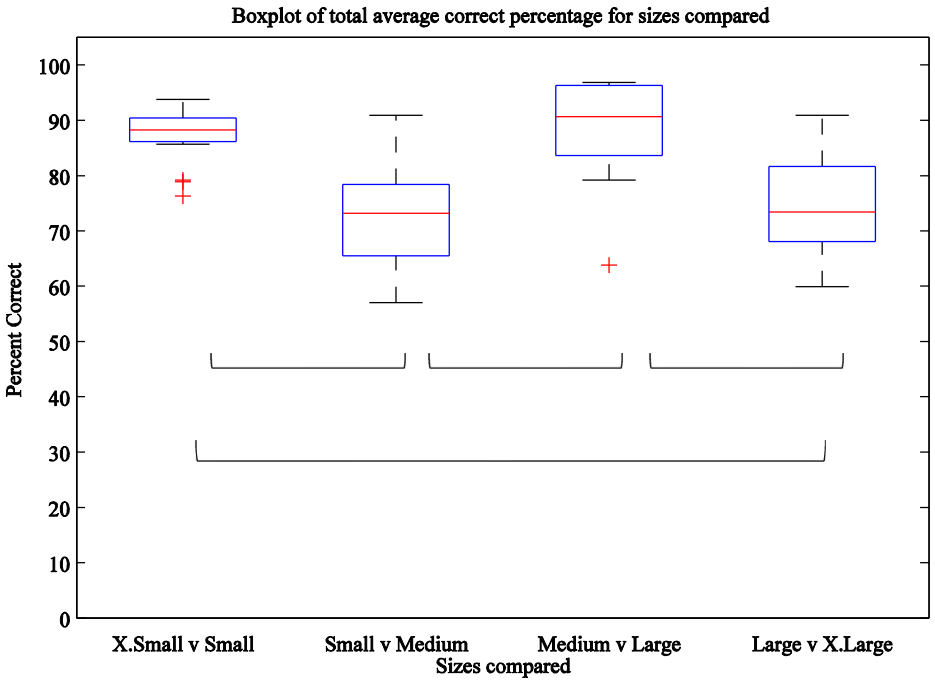

Figure 45: The effect of sizes compared. The lines indicate significant differences between means using Bonferroni post-hoc tests. The high scores for X.Small v Small and Medium v Large correspond with the two largest SCF differences, and the two smallest SCF differences correspond with the lower scoring Small v Medium and Large v X.Large pairs.

6. Discussion

The motive behind this experiment was to study the importance of spectral cues in the size discrimination of transient signals. Three methods were followed to achieve this: comparing the results of a simple size discrimination task to the SCF of the signals; scaling the signals in an attempt to remove the spectral cue; and filtering the signals in a number of ways to determine the importance of the different areas of the frequency spectrum. The SCF of each signal was calculated in a way described in section 2.2. Experiment 1 discovered that adjusting the PSR was an unsuccessful scaling technique, and so after signal analysis a method using the ratios between the F1 in each signal was used to scale the signals instead. The signals that were scaled to have the same F1 in Exp. 3A were compared in order to test the success of the scaling technique, and then to check if other cues were available after the spectral cue was removed. Signals were also compared to scaled versions of themselves (Exp. 3B) to test whether or not participants would always choose to listen to spectral cues for size discrimination. Finally, the signals were low-pass and high-pass filtered either retaining F1 in the spectrum or not, in order to determine the importance of this resonance.

6.1 Repeatability

Measures were taken to improve the repeatability of the experiment by assessing the degree of variability within individual results sets. Of the 40 people that took part, eight people were unable to complete all the rounds of testing and 13 were inconsistent in their test responses. These results from all the experiments in chapter 5 are from 19 of the remaining participants that were found to be the most reliable after a test to gauge the consistency in their answers. Reliability was tested by gathering the data for the unfiltered and unscaled size discrimination questions in the experiment. Within size pairs (e.g. Small v Medium) the participant was given a score of 1 for every correct answer, and for every incorrect answer they scored -1.

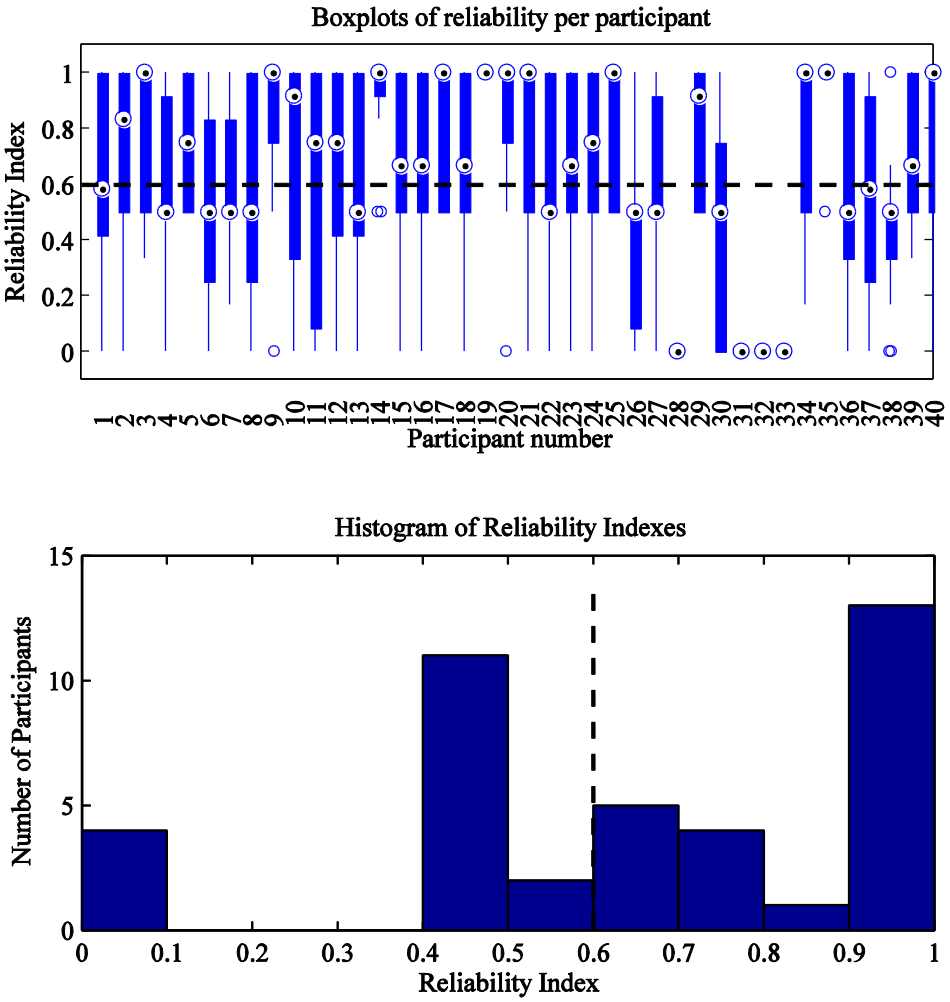

Figure 46: The boxplots above show the absolute reliability index for each of the 40 participants that took part in the experiment. The thick blue line indicates the interquartile range for the RI, with the thin blue line showing 1.5 times that. Empty circles indicate outliers in the data. The dot within the circle marks the median of the RI range for that participant. The participants whose median is at zero did not complete the testing. The histogram shows the number of participants per value of RI. The dashed black line at RI=0.6 in both figures above indicates the cut-off point for reliability; when the median occurs below this line, the participant is deemed unreliable as the chance they choose the same answer each time they are presented with the same comparison pair falls below 80%. As well as those that fell under the line, participant numbers 16, 24, 25 and 40, although reliable, were unable to complete all rounds of testing.

The total was calculated and normalised to between -1 and 1; this was the initial reliability index (RI) for that size pair for that participant. The further this value was

from zero the more often the participant chose the same answer in that pair, and therefore that participant was labelled as reliable. In accordance with the hypothesis that musical people have better pitch discrimination abilities (Micheyl et al., 2006; mentioned in section 4.2.1), the reliability of the participants seems to be closely related to whether or not they were musical. 60% of the non-musical participants were unreliable according to this test compared to 31.8 % of those that had years of musical experience.

All the absolute values of RI for each participant were added together and normalised to between zero and one. An RI value of 0.6 corresponds to an 80% chance that the participant chose the same answer every time they were given a specific comparison pair. Those with an RI of 0.6 or above were considered to be reliable in their size discrimination abilities, and their results were used for statistical analysis. The figure above shows boxplots of all the absolute RIs for each participant, and a histogram of the number of participants per RI value, with dashed lines to indicate the 0.6RI cut-off point in each panel

6.2 Discussion

In brief, the results show that discriminating between the spheres comes easily to the participants. The SCF is of great importance when it comes to correctly judging the size of the signals. The greater the difference between the SCF of the signals, the easier it is to identify which signal is bigger. Scaling the signals using the ratios between F1s as a method of removing the 'pitch' cue was not successful: the participants chose the signal what had been scaled up from smaller. When signals were compared to scaled versions of themselves, participants always chose the signal with the lower F1. After filtering, the presence of F1 in the signals helps those which have been low-pass filtered, but with high-pass filtered signals the differences between F2s are greater and so results are higher without F1. In band-stop signals, there is a significant decrease in scores after filtering out F1, but results were significantly better than UnF after both F1 and F2 were filtered out. These results will now be discussed and conclusions will be drawn about what cues are important in a size discrimination task.

6.2.1 Experiment 2 - Simple size discrimination abilities

First and foremost, participants showed they can discriminate between the different sized spheres, after having scored significantly above chance in all sets. They showed better skill when listening to the averaged signals compared with signals from the database, which is possibly caused by the variability in the signals due to human error while recording the signals. For a perfect signal to have been recorded each time, the sphere would have to have been struck along the line of the sphere's diameter (as discussed in section 4.1.1). When struck along this line, the length of the diameter corresponds to half the wavelength of the first and strongest resonance in the spectrum. The experimenter's involvement during recording resulted in not all of the signals being created as such.

When the sphere is struck off this line, the first resonance is not as prominent, and the resulting waveform is more variable compared with the ideal. After the RMS normalising process and the database had been pruned of the unwanted signals (section 4.1.1), a number of the signals that remained still had a slightly irregular waveform shape and so there was some variability in the signals presented to the participants. This variability could have caused some confusion in the discrimination test. Creating a set of averaged signals for each size removed any variability resulting in each F1 being clearer and thus easier to discriminate. However, in the real world of single pulse-resonance sounds, the listener often does not get to choose the best representation of a sound he or she listens to and often has to make do with a less than perfect representation of what an object's spectrum could be. For this reason, the results of the random UnF signals test is important as it demonstrates that even with a flawed representation of an object's spectrum, the listener can still discriminate for size.

The SCF of the spectrum had an effect on the perception of the size of the object. The results showed there to be greater variance in the results when comparing sizes that had the greatest difference between their SCF values. Figure 30 shows the boxplots of results for the unfiltered signals aligned with the difference between the corresponding spheres' SCF values. Figure 30 showed the results of the effect of sizes compared from all the different types of filtered signals,

and again this showed that SCF had an effect on abilities. The larger the difference in SCF values, the fewer the number of errors. These results are important as they show that the auditory system does not need a periodic representation of a pulse-resonance sound in order to extract size information. The signals here were presented to the participants five times in succession, but with a gap of 0.5 seconds between each instance, so that no fundamental frequency could be made from the repetitions. The results prove that a signal does not have to be periodic to be able to discriminate for size easily.

6.2.2 Experiment 3 - Scaled signals

Experiment 3A

After the participants were shown to have good size discrimination skills, the signals were scaled and presented to the participants in order to prove they could not be fooled simply by scaling the frequency spectrum. The signals were scaled to have the same F1, in an attempt to be perceived as having the same transient pitch. The scaling of the signals was to show that frequency was not the only cue used by a listener in a size discrimination task to decide which object was bigger. Signals were scaled using the ratios between F1s as a scaling factor and new signals were created with the same spectrum shape as before but scaled so that F1 matched that of a sphere of another size. It was hypothesised that in the task where F1s were matched, participants would be able to hear the signal which was originally the larger sphere, i.e. they would use cues other than frequency to hear size. The expected opposite result to the hypothesis would be confusion in the task due to both signals now having the same F1 value. However, neither of these outcomes occurred; there was little confusion and the majority of the answers were for the wrong signal. The fact that the wrong answer was chosen so often prompted an analysis of the scaling process. The signals were scaled so that F1s matched, but of the scaled signals the participants had a preference towards the signals which had been scaled from the smaller sphere, and in general never chose the signal that had

been scaled from a bigger sphere. A review of the results is as follows, with those signals that were perceived as bigger marked with an asterisk[1]:

Unscaled Signal	VS	Signal was Smaller*
*Unscaled Signal	VS	Signal was Bigger
*Signal was Smaller	VS	Signal was Bigger

In the tests where a signal that had not been scaled was compared with one that had, there was the concern that the scaling process may have created some noise or other distinguishing feature in the signal to set it apart, and then caused the subjects to either choose one or the other consistently as their answer. The results showed that when the comparison signal was scaled from being smaller, they chose the scaled signal, but when it was the signal scaled from bigger, the participants chose the unscaled signal, thereby suggesting there was no giveaway feature in the scaled signals to confuse the participants. The preference for the 'signal-was-smaller' was shown in the final comparison task, and thus there is reason to compare the spectra for cues other than the matched F1s.

Figure 47 below shows two examples in order to more closely see how the unscaled and scaled spectra compare. The first plot in the figure compares the unscaled signal from the Medium sphere with the signal that was scaled from Large. While the F1 peaks line up, with increasing formant number the peaks begin to lose this alignment and the spectrum of the scaled-from-bigger signal shifts to the right. This gives an overall SCF that is higher than the unscaled signal, thereby causing the participants to disregard it as the smaller sphere. A different effect happens when the signal is scaled from smaller. The F1 of the unscaled X.Large sphere is again aligned with that of the scaled signal, X.Large was Large, but the peaks of the scaled-from-smaller signal following F1 are much smaller in magnitude than the same peaks of the unscaled signal. Figure 48 shows the resulting SCF due to these

[1] To clarify, Signal was Smaller refers to a signal that was originally smaller and has been scaled to have an F1 of the size bigger than it. Signal was Bigger refers to the opposite: a signal that was larger, and has been scaled to have an F1 of a smaller signal.

magnitude differences; the scaled signal has more power in the low formants and results in a lower SCF, and thus causing the participants to judge the unscaled signal as the smaller signal, and like before, disregards it.

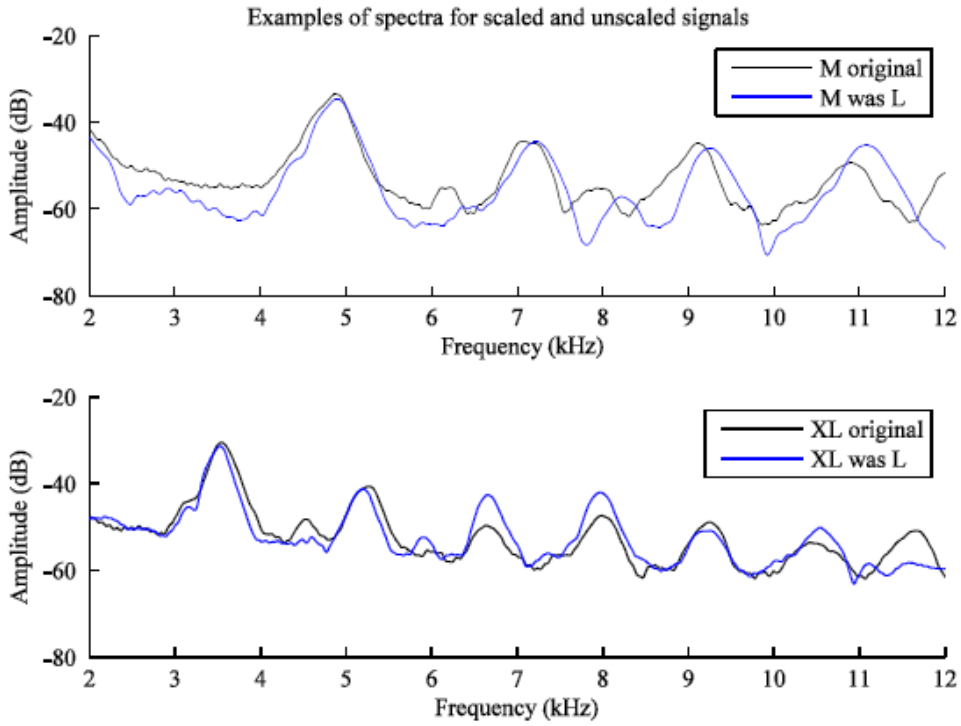

Figure 47: A zoomed in representation of a selection of PSDs from figure 16, to show how the formants differ after scaling from bigger (upper plot) and from smaller (lower plot) compared with the unscaled signal, when both signals have been scaled to have the same F1. The signal-was-bigger shows formants shifting to the right and thereby resulting in the signal having a higher SCF. The signal-was-smaller shows stronger formants in the lower end of the spectrum and is chosen as the larger sphere because it works out to have a lower SCF.

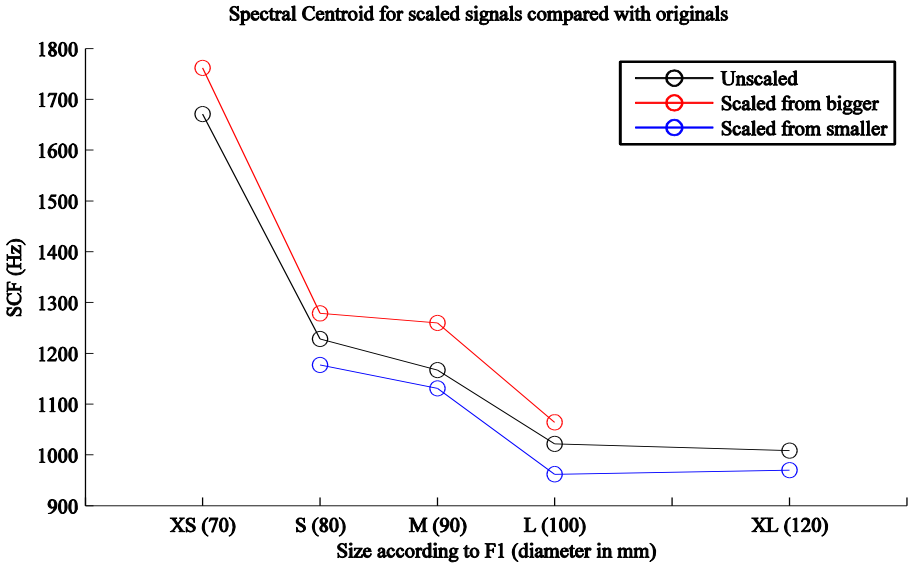

Figure 48: Spectral Centroid Frequencies of unscaled (black) and and signals which has been scaled from bigger (red) or smaller (red) in order to have the same F1. The red line shows that when signals were scaled from bigger, they ended up with a higher SCF than the unscaled and scaled from smaller signals, and therefore were judged as smaller. The blue line shows that when the signals were scaled from smaller, their spectra resulted in a lower SCF and caused the participants to judge the signals as from larger spheres.

Experiment 3B

The second part of this hypothesis was to compare signals which were originally the same size but one of them had been scaled to be 'made Bigger', or 'made Smaller'. It was expected that due to the strength of the spectral cues, participants would use this as the cue, whether it was incorrect of not. One signal was scaled so that F1 in the spectrum matched a sphere of a different size either smaller or bigger, but if participants were to use another cue rather than frequency they would be confused by the fact that both signals were the same size before scaling. A review of the signals that were compared is as follows, and again the signal chosen to sound the largest most often is marked with an asterisk:

Unscaled Signal	VS	Signal made Bigger*
*Unscaled Signal	VS	Signal made Smaller
*Signal made Bigger	VS	Signal made Smaller

The results were as expected with the participants almost always chosing the signal with the lower F1 as their preferred larger signal, i.e. signal made bigger. Figure 49 below gives two example comparison pairs. In the top plot the unscaled signal was the X.Large sphere, and the scaled signal was the X.Large signal scaled to have the F1 of the Large sphere. The lower plot shows the unscaled Medium signal compared with the Medium signal that has been scaled to have the F1 of the Large sphere. In both cases, the scaling has caused a large shift in the spectrum, and in accordance with the previous results for using frequency as a cue for size, it is clear why the participants chose the unscaled XL signal as the largest in the first plot, and the scaled Large-was-Medium signal in the second.

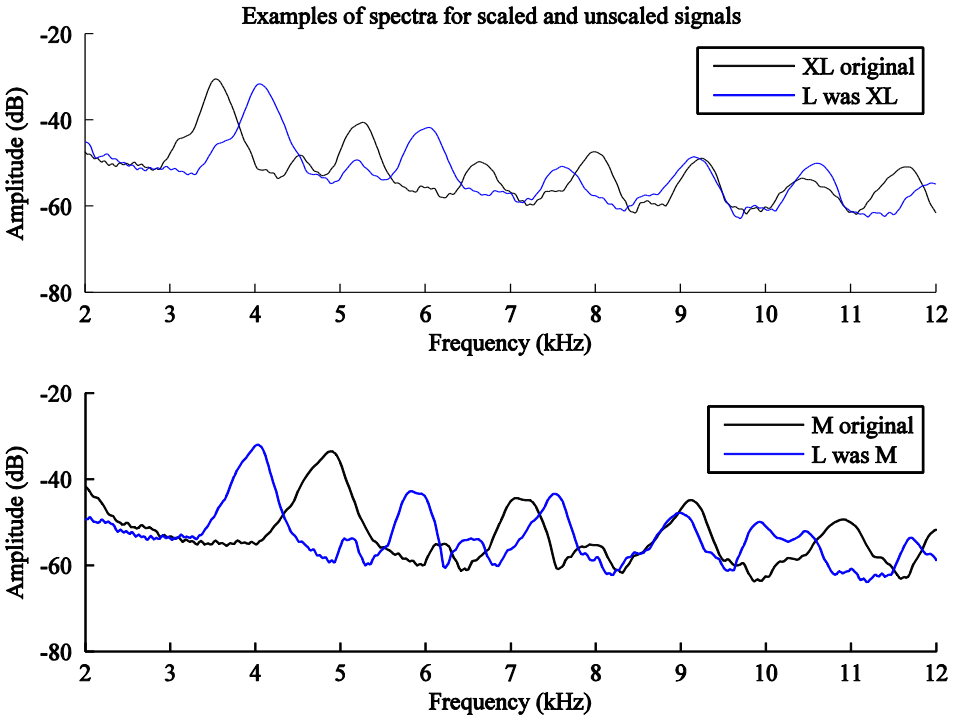

Figure 49: Zoomed in examples of comparisons made in the scaled with different F1s task. The scaling shifts the spectrum considerably, and the participants tended to choose the signal with the lowest F1 as the largest, regardless of their original size.

6.2.3 Experiment 4 - Filtered Signals

The results of the scaled signals task proved that participants did not look for another cue for size discrimination if they could choose to listen to frequency. In no cases were they confused by the tasks, they always chose the signal with either the lowest F1, or if F1s matched the participants chose the signal with the lowest SCF as the larger signal. Accepting this as the most important size cue, the next step in the experiment was to uncover where exactly in the spectrum was the most important for the listeners to determine size.

In order to do this, the averaged signals were filtered and presented to the participants in the same experimental format. For each of the filters used - low-pass, high-pass and band-stop - F1 was the deciding factor for positioning the cut-off frequencies so that the filtered signals no longer included F1. The LPF and HPF cut-off frequencies were also adjusted to create signals that included F1. Figure 50 below compares the time-series waveforms for the X.Large sphere for all filtering types. There was a significant difference in ability when discriminating between the BSF no F1 signals compared with the UnF signals. In the first row of figure 32, the UnF and BSF no F1 waveforms differ slightly in the first and second upwards rising peaks; the UnF waveform shows more variation in these sections of the waveform compared with the same sections in the BSF no F1 waveform.

Figure 51 shows the spectrograms of X.Large signals for all types of filtering. Again, the first row shows the UnF and BSF no F1 X.Large signal. A large band of low frequency energy spans almost the full length of the time axis – due to the impulse response of the sphere being struck, with shorter bands of energy occurring at ~1.5 kHz intervals from ~3.5 kHz and above, gradually fading out towards 12 kHz. These shorter bands are the resonances of the sphere, and their frequency value is determined by the size and shape of the resonating object (see section 3.5). The first two plots shows the removal of the band of energy around 4000Hz after BSF no F1 has been applied. This resonance, F1, is the strongest resonance, and removing it significantly lowered discrimination abilities. The spectrograms show there to be plenty of information left in the signals, however,

and this allowed for the results to remain significantly above chance whether F1 remained or was filtered out.

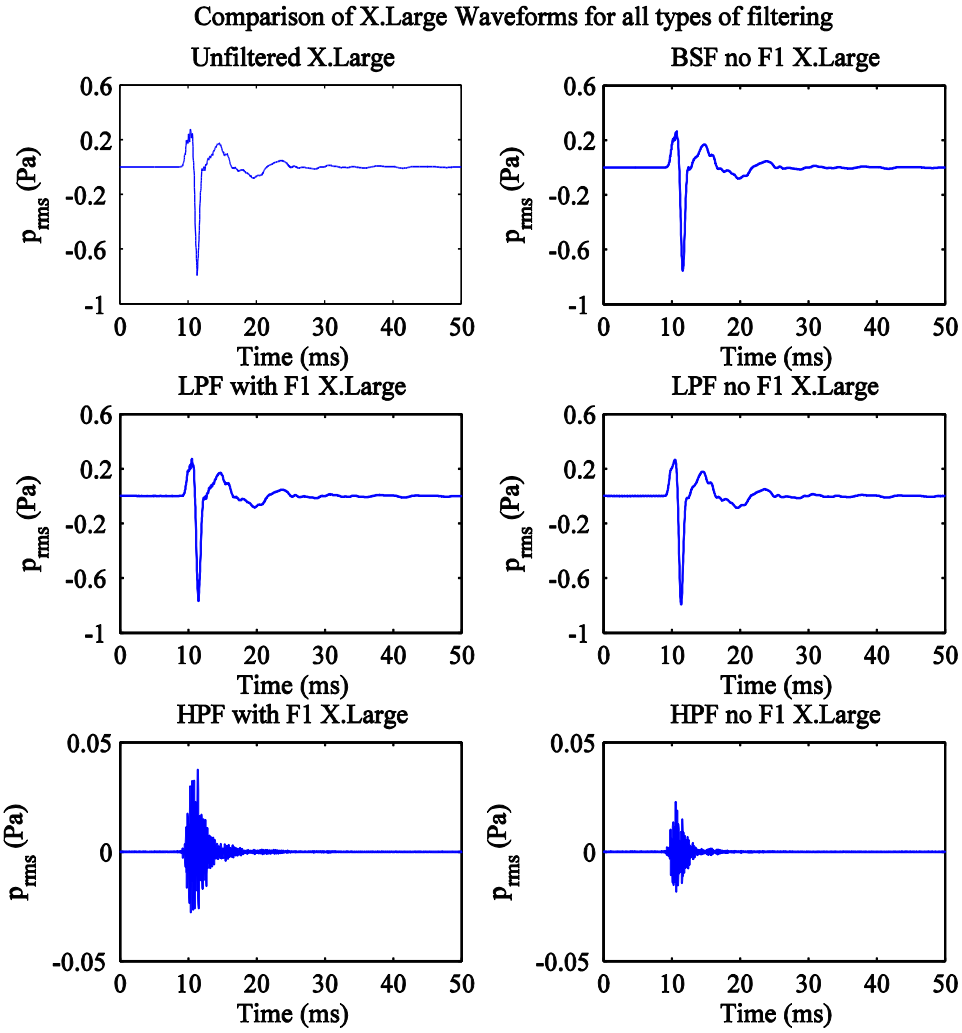

Figure 50: Comparison of the waveforms from unfiltered and filtered X.Large signals. The first row shows the UnF and BSF no F1 waveforms, and that the latter has a slightly smoother shape. The middle row compares the LPF with F1 and no F1 signals, again the latter has a smoother shape, and both have less noise in the first peak of the waveform compared with the first row. The last row shows the HPF with F1 and no F1 signals. The waveforms have lost their characteristic shape and are almost unrecognisable.

The middle row in both figures 49 and 51 show the low-pass filtered signals; one has a cut-off frequency just above F1 (LPF with F1), and the other just below F1 (LPF no F1). On close inspection, the LPF waveforms are smoother than all the

others, with LPF no F1 showing a slightly smoother shape, this is a minute difference in the waveform and can be seen as a smoother tip to the first peak of each waveform. This can be seen more clearly below in figure 50. LPF with F1 still contains F1, and LPF no F1 contains only the impulse response from the strike but none of the resonances, and even though almost three quarters of the frequency information in the spectrum has been filtered out, the general shape of the waveform has remained intact. The spectrograms show that a considerable amount of information has been removed after applying the low-pass filters, which resulted in a significant decrease in discrimination ability among the participants. However, the results showed there to be no significant difference between UnF and LPF with F1 (Figure 42). There is enough loss of information only in LPF no F1 for the signals to be altered enough to cause discrimination difficulties, suggesting the importance of F1 information for LPF signals.

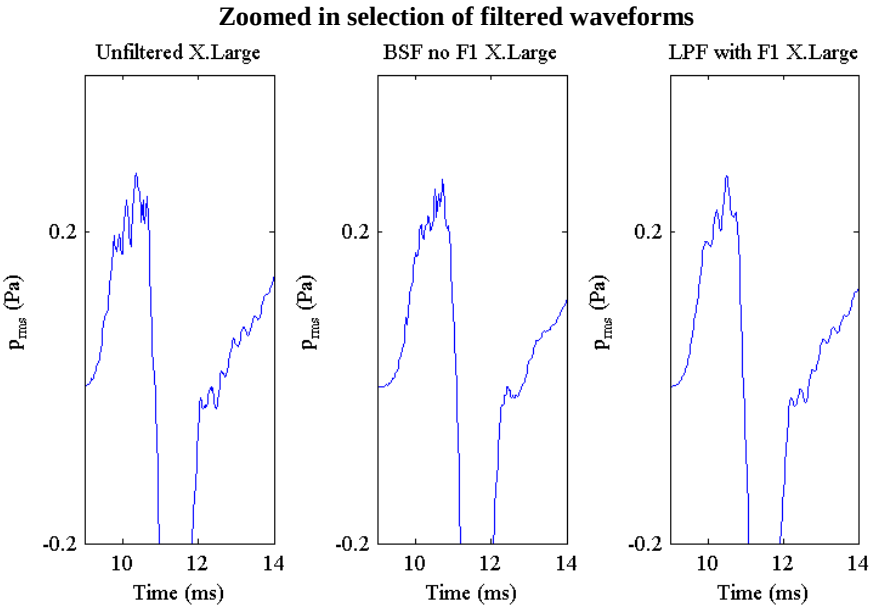

Figure 51: A zoomed in selection of the Unfiltered, BSF no F1 and LPF with F1 X.Large sphere signals, in order to show the differences the filtering has caused to the time series waveforms.

The last rows in each of figures 49 and 51 show the waveforms and spectrograms for the high-pass filters, and again the cut-off frequency was just below F1 in one set (HPF with F1) and just above F1 in the other (HPF no F1). The waveforms of the X.Large HPF waveforms has drastically changed from the UnF and LFP waveforms due to the absence of the low-frequencies that give it its characteristic shape, even though more than two thirds of the spectrum still remains. The HPF spectrograms show the resonances from F1 and above still remain in the HPF with F1, but all the information below this has been removed. In the second image, F1 is no longer seen.

In the discrimination of filtered signals test, HPF signals caused the lowest scores among the participants due to the removal of low-frequency information and the resulting difference in waveform shape. However, participants scored lower when F1 was still contained within the signal than after it was removed. The spectrum of the HPF no F1 signal contains less power than that of the HPF with F1 signal due to high frequency resonances being weaker, and despite the reduced energy and information the scores for these are higher.

Alternatively, the presence of F1 improves the scores for discrimination of LPF signals. The hypothesis assumes that F1 is the most important frequency component for effective size discrimination, and removing F1 from the BSF signals resulted in a drop in discrimination abilities. However, without F1 the scores for HPF are higher than with, thus suggesting that F1 may not be a necessary component of the spectrum for size discrimination for some signals. Relevant spectral information must be stored elsewhere: in the relationships between the formants for the HPF signals, or in the impulse response frequencies (0-2 kHz) for the LPF and BSF signals. Table 4 already looked at the values of F1 and F2 and their relationship, but it may be worth analysing this information in a different way.

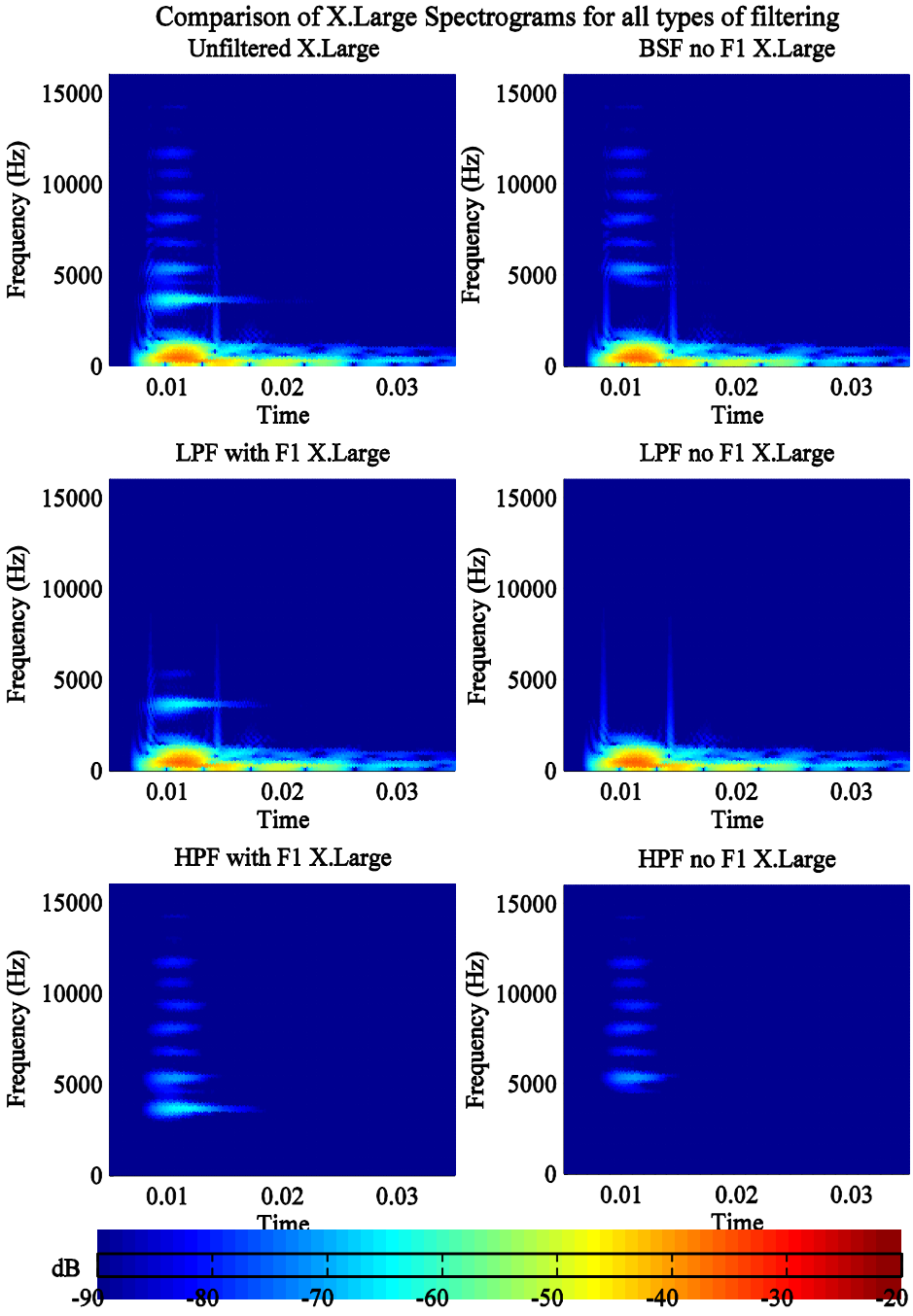

Figure 52: The spectrograms from unfiltered and filtered X.Large signals that correspond with the waveforms in figure 17. The first row shows the removal of the band of energy at ~4 kHz which is referred to as F1. The middle row displays how much information has been removed by the low-pass filters, with only just the impulse response left from the LPF no F1. The last row shows the clear high frequency formants that remain after the HPF process.

Table 9 below contains the formants indicated by black arrows on the PSD plots in figure 53, with the help of the frequency analysis tool in Adobe Audition for precise values, along with the ratios between the neighbouring formants. As it was seen earlier, the formants in the three largest spheres are clearer and more evenly spaced than the smaller spheres.

The table makes it easier to see that there is a tendency for the ratios, i.e. the distances between neighbouring formants, to increase with formant number. In other words, the higher the formats, the further apart in frequency they tend to become. While this may be of use for the shape of the object (Bergman, 1994), it does not explain why F1 works differently for HPF signals than it does for LPF signals.

	F1 (Hz)	F2 (Hz)	F3 (Hz)	F4 (Hz)	F5 (Hz)	F1/F2	F2/F3	F3/F4	F4/F5
X.Large	3531	5254	6632	8010	9216	0.672	0.792	0.828	0.869
Large	4048	5943	7665	9130	10,590	0.681	0.775	0.84	0.862
Medium	4909	7062	9130	10,930	12,050	0.695	0.774	0.835	0.907
Small	5340	7235	9560	11,710	12,570	0.738	0.757	0.816	0.932
X.Small	5857	8354	9991	12,050	13,000	0.701	0.836	0.829	0.927

Table 9 Values of F1-F5 for each size of sphere, and the ratios between neighbouring formants. The ratios and therefore the distances between formants increase with increasing formant number.

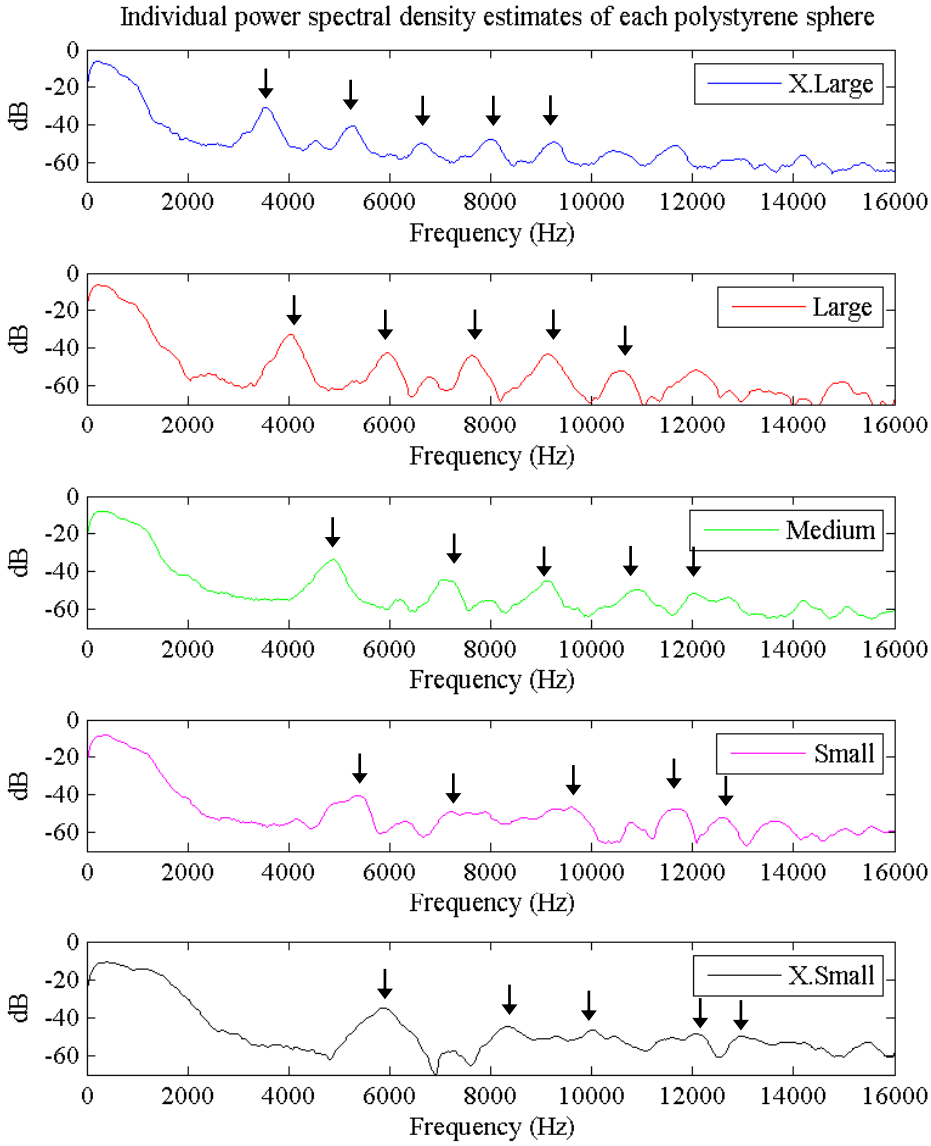

Figure 53: PSD plots for each unfiltered signal recorded from the spheres. The arrows point to where the formants F1-F5 lie on the spectrum as referred to in table 9.

	X.Large	Large	Medium	Small	X.Small	X.L – L	L - M	M - S	S -X.S	Average
F1 (Hz)	3531	4048	4909	5340	5857	517	861	431	517	581.5
F2 (Hz)	5254	5943	7062	7235	8354	689	1119	173	1119	775
F3 (Hz)	6632	7665	9130	9560	9991	1033	1465	430	431	839.75

Table 10 A different view of the formants F1-F3 in the sphere spectra, with a breakdown of the distances between the same formants in each size. The final column shows the difference between the formants in the largest and the smallest sphere. The results for HPF no F1 are higher than HPF with F1 due to the easier task of discriminating between F2s, rather than F1s, as they tend to be further apart.

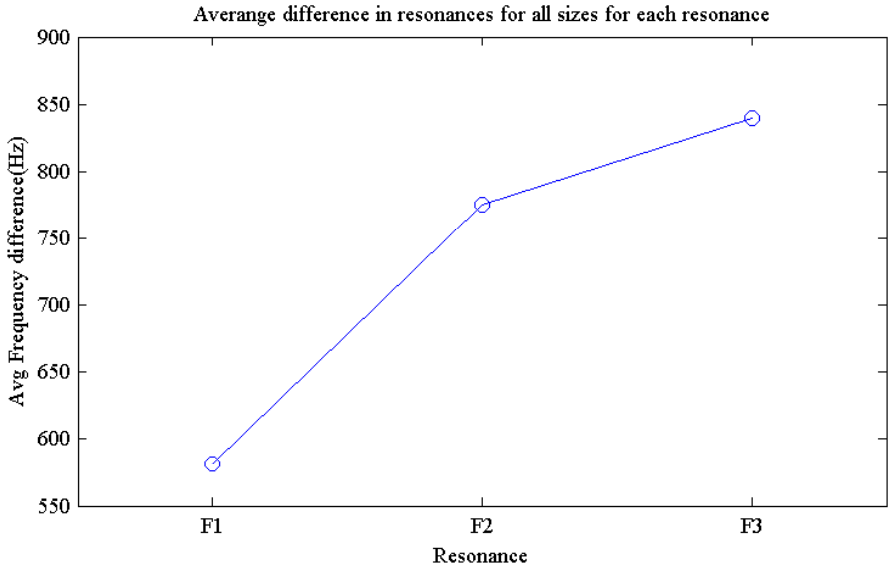

Figure 54: A plot of the averages from table 10, showing the average distance between F1s across all sizes is the lowest, F2 is higher and F3 is the highest. This explains the results of the Band-stop filtered signals discrimination test, where filtering out F1 and F2 showed higher mean scores than filtering out F1 alone. The reason is due to the larger difference in resonance values.

Another way to look at the table is the distances between the formants across all the sizes. From X.Large to X.Small, the frequency values of F1 range from 3531 Hz to 5857 Hz this is a difference of 2326 Hz. The difference between F2 values is a much greater 3100 Hz. Table 10 shows the distances between the F1s and the F2s from size to size. The general trend shows that there is a greater distance between

F2 values than F1 as the averages show in figure 54, apart from the Medium to Small pair. The results showed that participants were capable of discriminating for size in both cases of HPF signals, but that they better when F1 was not in the signal. If F1 is removed, the participants could have used F2 as their discrimination cue. Even though the F2 values are higher, the differences between them are still well outside the JND range for those frequencies, as illustrated by figure 55 below (Weir et al 1977). And since the distances between F2 in each size are greater than the distances between F1, it is likely that the participants found the task using signals with F2 as the cue easier than the signals which included F1.

The second band-stop filter used removed both F1 and F2 from the signals, and participants performed the size discrimination task significantly better in this than with the BSF No F1. The explanation used for the LPF and HPF no F1 results may only explain this for the larger signals, as the differences between F3, the next available formant, is only much greater for the larger spheres. Another possibility is the spectral centroid frequency of the signals before and after filtering.

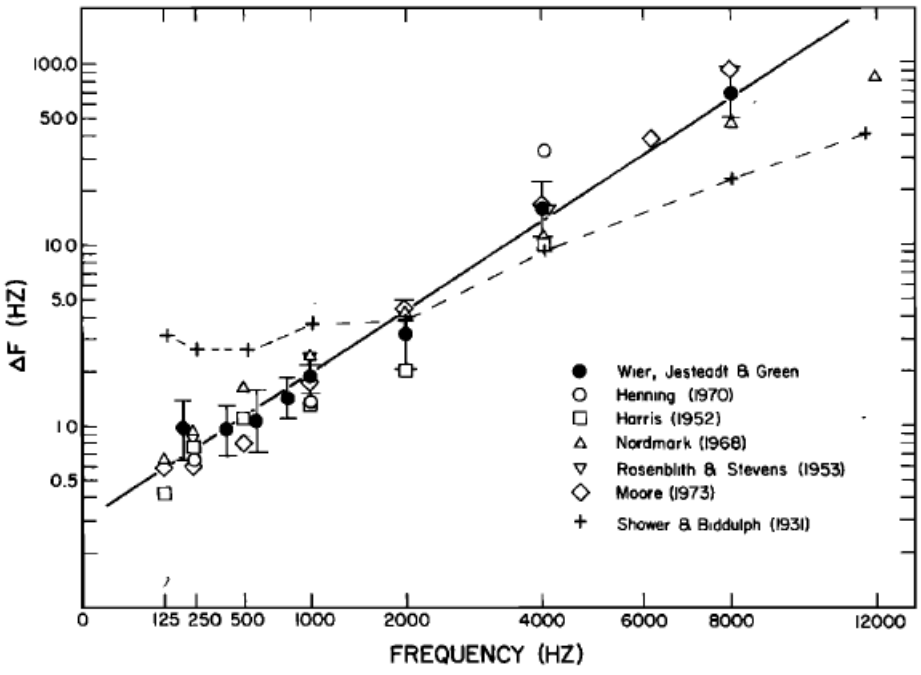

Figure 55: Results from studies that measured frequency discrimination thresholds using 40 db SPL tones. The thresholds, ΔF, are plotted as a function of frequency,

both in Hertz. (Weir et al, 1977. Reproduced with permission from the Acoustic Society of America).

Figure 56 below shows the SCFs for the UnF signals compared with those of the BSF No F1 and No F1F2 signals. The blue line for the unfiltered signals shows the decrease is SCF as the size of the sphere increases. As previously mentioned in section 5.2, the shallow decrease in SCF of 13 Hz between the Small and Medium spheres corresponds to more errors made in the task. When F1 is filtered out, the SCF decreases for all sizes as shown by the red line, but the SCF of the X.Large sphere has increased by 4.5 Hz from that of the Large sphere. This change in direction could cause some confusion in the task if the SCF is used as a cue. Finally, the green line shows the SCFs for the BSF NoF1F2 signals and it has the most consistent downward slope, with the smallest decrease in SCF at 51 Hz. If the spectral centroid frequency of a signal was used by participants as their preferred cue in the discrimination task, these BSF signals with both F1 and F2 filtered out would have been the easiest to discrminate between, hence explaining why they made fewer errors.

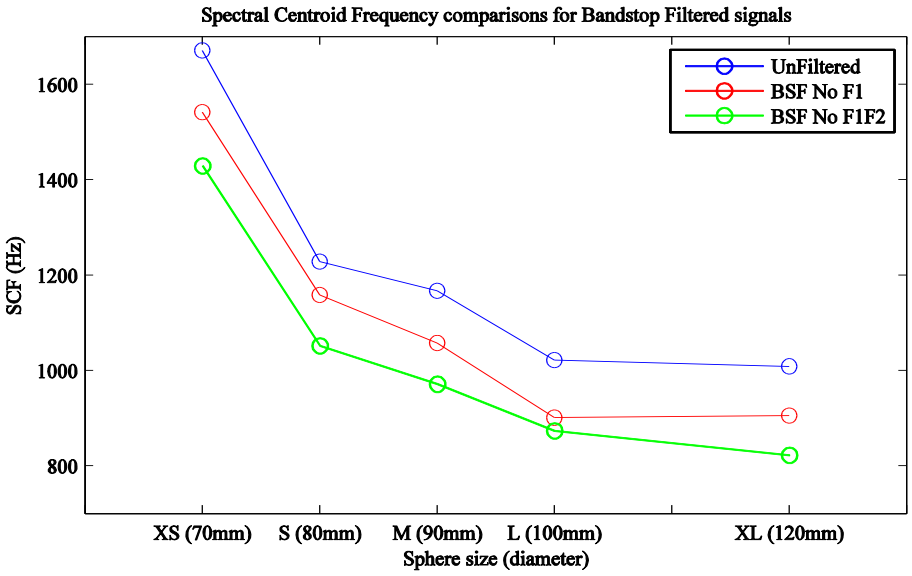

Figure 56: The change in spectral centroid frequencies after the two BSF filters were applied. The BSF no F1F2 has the most consistent decreasing slope as the size of sphere increases, accounting for the lower number of errors in the results.

For this task it appears that the participants chose to ignore the F3 values in favour of the SCF to make their choice. It could be the fact that the F3 values were too high in the spectrum for them to be a useful cue, or perhaps the strength of low frequency content in the signals caused enough of an upwards spread of masking to render F3 useless. In the BSF task, the SCF was the most informative cue. What about in the LPF and HPF signals size discrimination task?

Figure 57 below shows the SCF values for LPF with F1 and no F1 signals, and also the HPF with F1 and no F1 signals. It has already been discussed that participants used the differences in either F1 or F2 to determine which object was larger. Figure 57 shows us whether or not the SCF of the signals could also be a useful cue in the task. The top panel gives the SCF for both of the LPF signals. As expected, the SCFs for the signals without F1 are lower overall than those with F1, but it would also be expected that the values decrease with increasing sphere size. However, there are degrees of variability here for both filters used, even more so in the LPF with F1 even though the results showed fewer errors. In fact, the smallest decrease in SCF is 7.2 Hz between the Large and X.Large sphere, and the SCF increases by 7 Hz from the Small to Medium sphere. Considering the results from the BSF signals, it is conceivable that if SCF was the only cue used, there would be a higher number of errors in the results. The opposite occurred and the participants chose to ignore the SCF in this case and used what formants they could hear to perform the task. On the other hand, scores for the LPF no F1 were also above chance albeit significantly lower than those for LPF with F1. In this case, the lack of F1 meant the participants had to use another cue, and SCF for these signals provides the required information.

The spectral centroid frequencies for the HPF signals are shown in the second panel of figure 57. Contrary to those of the LFP, there is a lot more consistency in the slopes between SCFs in both HPF With F1 and No F1. In addition to this, the SCFs for the signals with F1 are lower than without. Since the lower the frequency the easier discrimination becomes (figure 55, Weir et al. 1977), this is not in accordance with the results as the participants made fewer errors with the signals that had the higher SCF values and did not contain F1. In this case it is may be

more difficult to say if SCF was the cue the participants favoured in order to perform the task. However the evidence for formant frequency values as a cue is stronger and so in this case the participants appear to have disregarded SCF in favour of the correct information the differences in the F1s and the F2s provided.

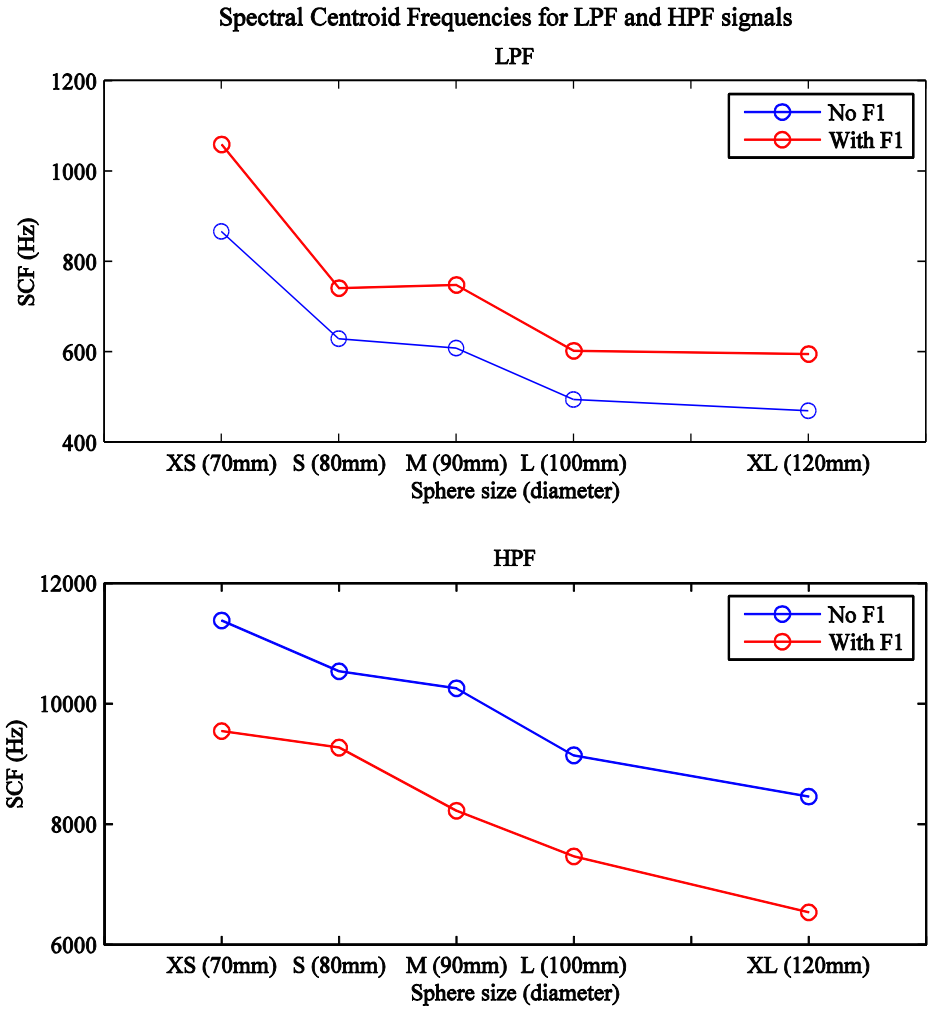

Figure 57: An examination of the spectral centroid frequencies for the LPF and HPF signals in order to judge their value as a size discrimination cue. Contrary to which results were higher, the LPF with F1 had least consistent downwards slope, which should have caused confusion. Likewise against the results, the HPF with F1 had overall lower SCFs, yet the participants performed fewer errors with the HPF no F1 signals.

The final point to discuss here is that although the formants had all been filtered out, participants were still capable of discriminating between signals that had been low-pass filtered with a cut-off frequency below F1. It has already been discussed that the SCF of these signals were used as a cue, however the PSDs seem to show just an impulse response of the impact on the spheres, and the frequency information was contained higher in the PSDs, i.e. from F1 and above. Figure 58 zooms in on a selection of the power spectral densities already seen of the unfiltered signals so as to more clearly see the frequencies below F1. Across most of the selection below, the power for each line corresponds to size; the larger the sphere the more power each frequency has. In terms of the experiment, this point is moot due to the RMS normalising process and amplitude roving in the presentation of the signals to the participants to ensure volume did not play a part in the discrimination process. The main point here is the width of the band of frequencies that occur in this selection is due to the impact of the ball bearing on the sphere. The smaller the sphere, the wider the band of frequencies that ring and this contributes to the SCF for each signal.

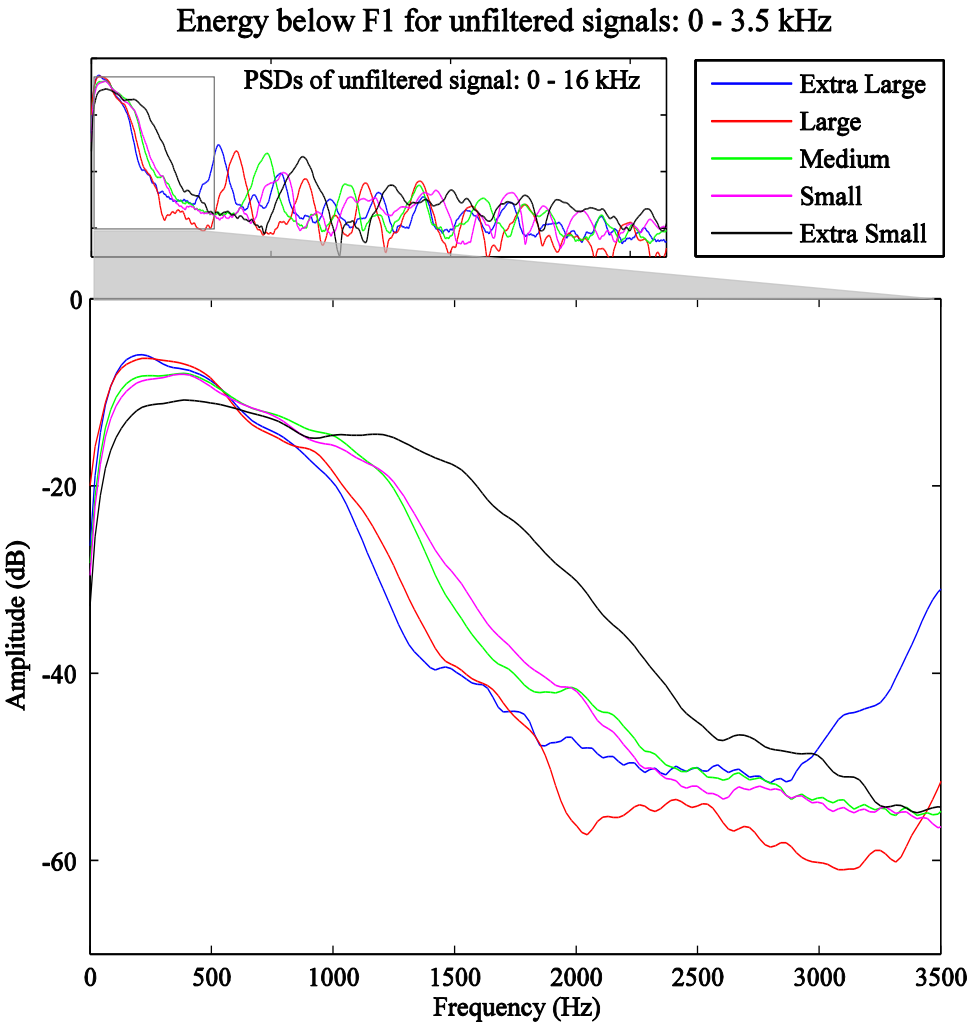

Figure 58: A zoomed in selection of the low frequency end of the spectrum for each unfiltered signal, from 0 – 3.5 kHz. It can be seen that impulse response of the impact on the spheres causes a specific bandwidth to ring, and this bandwidth gets wider with decreasing sphere size.

6.3 Conclusion

Experiment 1 turned out to be inconclusive for the most part; even though there was an overall drop in performance of 20% from unscaled to scaled signals, the results had little validity due to low trials and low numbers of participants. However scaling the spectrum by adjusting the PSR resulted in participants hearing diverging harmonics, and this suggested the sphere signal spectra were not as simple to scale

as vowel spectra despite both showing linear relationships between size and resonances. It is possible that there are different types of pulse resonance signals, those that are scalable in the manner shown by STRAIGHT, and attempted here, and those that are not scalable. The vocal tract is a hollow tube with an opening; the spheres used here are solid with no openings. Further investigation into this possibility is required, but it is not within the scope of this research. It is only possible here to suggest the possibility of signals that result in spectra that have a linear relationship with the size of their source, but are not scalable in a linear manner.

Experiment 2 improved on the simple size discrimination task of the first experiment, but was carried out using a much larger number of participants and higher numbers of repeated trials to ensure validity. The results confirmed that the participants were able to easily discriminate between the sizes of neighbouring spheres, but that also, the larger the difference in SCF values, the fewer errors they made. The participants were presented with the spheres before the experiment in order to familiarise them with the sound source, but also in an attempt to get them to hear the spheres instead of signals with different transient pitch. However, the participants reported "the bigger sphere sounded lower", and so transient pitch was their main cue. An interesting note of the experience of the author was that some participants that found the task difficult could be heard humming along to the signals, as if trying to create a pitch from what they heard.

Experiment 3 went on to test a different type of scaling method, based on the strength of F1 in each of the sphere signals, and the ratio between F1s for neighbouring sizes. The results of Exp. 3A where the signals that were scaled to have matching F1s showed that this scaling method was not sufficient to remove transient pitch cues, as the participants ignored the matching F1s and chose the signal with the lower SCF as the larger signal instead. In Exp. 3B, when the signals that were scaled to sound bigger or smaller than they originally were compared with their original signals, the transient pitch always affected the choice of bigger sphere, with the lowest F1 or SCF being chosen as larger.

Experiment 4 applied filters to different areas of the sphere spectra to test the effects of removing spectral information. There appears to be a trade-off between two cues that participants use in size discrimination. First of all, participants have proven their ability to discriminate for size regardless of the filtering applied to the signal and how little of the spectrum remains. When one cue is no longer available, another cue is found. For the unfiltered signals, the larger the difference between SCF value, the higher the mean scores for discrimination. In the LPF signals, the presence of F1 makes the task easier than when F1 is also filtered out. The difference in SCF values does little to help as in two of the comparison pairs the difference is very little, and one of them is an increase in value. The HPF signals are clearer when F1 is no longer present in the signal due to a larger distance in frequency between the F2 values in each size, and again the SCF is of little use. Finally, the band-stop signals showed a significant decrease in mean scores when F1 was filtered out, but an increase in means when F1 and F2 were filtered out; there was no significant difference between UnF and BSF No F1F2 signals. An analysis of the differences between F1s, F2s and F3s in the signals showed that the differences between F3 for each comparison pair was bigger than the others, which accounts for the significantly higher mean scores for the BSF no F1F2 filtered signals over the BSF NoF1 signal results There could be a possible explanation pertaining to the fact the SCFs of the BSF no F1F2 filtered signals were the lowest of all three – UnF, BSF NoF1 and BSF No F1F2 - but this does not hold true for the BSF no F1 filtered signals which had significantly lower mean scores for size discrimination than the unfiltered signals despite lower SCFs (see figure 56).

The high-pass filtered signals produced the most interesting results, where a comparison of mean scores for HPF signals with or without F1 showed higher scores for those without F1. This was explained by the larger differences between F2 values than between F1 values across size comparisons. The SCFs for the HPF no F1 signals are quite high, ~8.5 – 11.5 kHz, compared with the much lower SCFs for the HPF with F1 signals, ~6.1 – 9.8 kHz, and so it is suggested that the difference between resonances is the important cue for discrimination here, rather than the SCF values (see figure 57).

The questionnaires handed to the participants to be filled out before testing were also returned after testing in order to answer some questions on the cues the participants used in the tasks. The most common answer for the cues used was transient pitch, followed closely by timbre. To the question regarding the signals that sounded as though they were the same pitch, a number of answers suggested that the larger signal sounded more hollow or dull. In most cases, participants said they tried to listen to transient pitch as much as they could in order to make their choice. The cues mentioned by the participants seem to mirror the analysis of the results. In most cases, transient pitch is the most important cue, heavily influenced by the most prominent resonance in the signal. For signals with the same transient pitch, a lower timbre, or centre of gravity, indicates a larger sphere.

Now that it is known how size information can be extracted from the frequency spectrum, the importance of the differences between F1, F2 and SCF, this study will progress to create an auditory model for transient signals, based on the Auditory Image Model and implementing the gammachirp filterbank previously described in Chapter 3. The AIM uses periodic signals and creates an averaged auditory representation of the signal in order to extract information such as size from the signal. It is proposed that a model can be created that does not require a signal to be periodic in order to discriminate for size. The next chapter justifies the need for this type of model and endeavours to create an automated size discrimination program using the results from this experiment as the cues used by the auditory system.

7. tAIM - Transient Auditory Image Model

7.1 Justifying a new model

Size and shape information of a sound source are contained in bio-acoustic communication sounds, and one of the tasks the auditory system performs is a segregation of the size and shape information for auditory perception (Irino and Patterson, 2002). In order to achieve this segregation with an auditory model, the model must contain three components; frequency analysis to uncover where the resonances occur, pattern stabilisation in the form of an auditory image, and pattern normalisation to separate the size information from the shape information. This study aims to create a model that can mimic these processes for the purposes of analysing transient signals.

From the literature mentioned in chapter 2, it is known that humans are capable of discriminating between sizes of speakers by listening to vowel sounds, but also inanimate objects such as cylinders and wooden rods that have been made to resonate in some manner. The experiment carried out here showed how well humans can do the task even after spectral information has been removed. A transient signal therefore contains enough information for a person to know properties of the object. Periodicity has been shown to improve recognition but it is not necessary for a signal to be periodic in order to tell it apart from another of a different size (Lyon, T. 2010a, MSc thesis). The current version of *aim-mat* analyses multiple periods of vowel sounds in order to create Stabilised Auditory Images, and this has been expanded upon to create a size normalisation process by way of the Mellin transform. This was to model the normalisation that occurs in the auditory system which enables humans to identify the vowel spoken regardless of the size of

the speaker. In order to further expand the research into auditory models and more specifically auditory modelling related to size discrimination, it was decided to create a new auditory model based on *aim-mat*, for the purposes of analysing non-communication transient environmental sounds. Humans process these non-periodic types of sounds in everyday situations and are often capable of identifying their sources through sound alone (Bergman, 1994). The experiment described in chapter 4 has proven that humans are capable of discriminating between different sizes of spheres, and so a tool for the purposes of modelling the same ability is proposed. The objectives of this model, which will be referred to as tAIM (Transient AIM), are as follows:

- Create an auditory image of a transient, non-periodic signal using the same dynamic compressive gammachirp (dcGC) filterbank that is used in the original AIM (Irino, Walters & Patterson, 2007)
- Demonstrate that size normalisation is possible for non-periodic signals thereby showing that:
  - size information can be extracted from within one pulse-resonance of a signal using an auditory model, and
  - tAIM has preserved size information by creating Mellin images of control stimuli, simulated vocal and underwater stimuli, and real recorded signals
- Create an analysis tool using the output of tAIM and the results of the polystyrene sphere experiment for the discrimination of different sized objects.

The modules of *aim-mat* include pre-cochlear processing which filters a continuous signal in accordance to the equal loudness contours suggested by Rosen and Baker (1994). The motion of the basilar membrane is simulated next using the dcGC filterbank, followed by compression and half-wave rectification to emulate the work of the inner hair-cells. The next step is to analyse the vowel by averaging over each period of the waveform using the strobe method and creating a Stabilised Auditory Image. Further to this, the creators of the model went on to create Mellin images of vowels in order to model the size normalisation that occurs in the

auditory system. Although tAIM is inspired by *aim-mat*, it was thought more favourable for the purpose of this study to leave in as much information about the signal as possible. Thus, the pre-cochlear filtering was not implemented, and the compression and half-wave rectification was also avoided. *aim-mat* has a frequency range of 100-6400 Hz, but this is deemed too narrow considering the resonances emitted by the polystyrene spheres. The range of tAIM is 100-10k Hz. Frequencies below 100 Hz are not analysed due to limitations in the filterbank module. Finally, the use of a non-periodic signal allows us to omit the strobed temporal integration, since several iterations of the same pulse-resonance is no long available.

The purpose of this project is to create an automated size discrimination tool based on the processing of the human auditory system. It is clear from the spectral analysis of the signals in section 4.1.4 that the information used by participants in the polystyrene sphere experiment could be easily extracted from the PSDs of the signals, notably the values of F1 and the spectral centroid frequencies of the sphere spectra. However, this study aims to show that it is possible to extract size information from the auditory images of different types of objects including the spheres using tAIM, due to the use of the dcGC filterbank.

The following is proposed:

- A model based on AIM created for use with transient signals: tAIM.
- The Mellin transform can be used on the tAIM auditory images of non-periodic damped-sinusoids and successfully normalise for size.
- tAIM will be capable of discriminating between two sizes of objects by analysing the auditory images using the cue that was found to be the most important in the psychophysics experiment: the difference between the most prominent resonances.

In order to carry out the first two proposals, double-damped sinusoids were created at melodic intervals apart but corresponding to different sizes of VTL. It is expected that these signals will result in the same Mellin image, verifying the success of tAIM. The recordings of the polystyrene spheres will also be analysed. It is expected that after the result of the initial experiment where the spheres were

deemed 'unscalable', the Mellin process of this model will not be able to normalise for size. The physical composition of the spheres is very different to that of a vocal tract. The vocal tract is a cylinder of a specific diameter and length according to the size of the speakers, which changes shape due to the position of the tongue. Its length is independent of its diameter and tongue position, which means that the length can change and so the resonances will shift in proportion to the length. In contrast, spheres have a diameter and a circumference and they are dependent on each other, thus the circumference of the sphere cannot change without affecting the diameter. Section 4.2.3 has already shown how the diameters of the polystyrene spheres have a direct relationship to the F1 in the frequency spectrum. And also from a signal generation standpoint, a vocal tract and the spheres are very different. The pulse from a vocal tract is created from the vibration of the vocal folds at one end of the tract, and air is flowing through it as it is hollow. The spheres, however, are solid and the sound is generated by one point on the outside of the sphere being struck. It is these differences in structural properties and methods of signal generation that may prevent size normalisation using the Mellin transform on the spheres from being a possibility.

Finally, a tool will be created to discriminate between two objects of different size using the conclusions about size discrimination drawn from the subjective experiment earlier regarding the importance of spectral centroid frequency, F1 and F2 values, and the upper frequencies, and the information extracted by the Mellin transform. The magnitude of the Mellin power spectrum creates the patterns according to shape which shows size to be normalised, but the phase of the transform contains the size information. The model will extract the relevant size information and so will be capable of size discrimination regardless of object shape. This size-extraction section of the model will be discussed in the following chapter.

7.2 Outline of the model

As mentioned before, this version of the model does not include any pre-cochlear filtering or amplification in order to retain as much information about the input signal as possible. Thus, the first step of the tAIM begins with the spectral analysis; a dynamic compressive gammachirp filterbank is applied to a non-periodic

signal. This process breaks down the signal into a multichannel filterbank according to the users parameters; a representation of the basilar membrane motion. The parameters used here create a filterbank of 100 channels spaced from 100 Hz to 10 kHz on an ERB scale (Glasberg and Moore, 1990). As mentioned before in section 3.3, the dynamic compressive gammachirp auditory filter consists of a passive gammachirp filter, and a high-pass asymmetric function that causes the filter to widen and its centre frequency to increase slightly as the stimulus level increases (Irino and Patterson, 2006). The result is an image of the basilar membrane motion (BMM), seen below in figure 59 using a broadband click as a stimulus.

The next step in AIM is the NAP process which converts the motion of the basilar membrane into neural impulses. This involves half-wave rectification of each filter, which removes the negative values of the oscillating waveforms and effectively makes the filters uni-polar, like the response of the hair-cell which sends the neural pulses in only one direction of its movement. Compression and low-pass filtering is also included in this NAP process. In tAIM, the compression will be included in the dcGC filterbank, and removing the half-wave rectification and low-pass filtering only serves to retain as much information about the signal as possible so as to allow for a size discrimination model eventually. The result of the psychoacoustic experiment earlier regarding the importance of high frequencies in size discrimination of transient signals suggests that the low-pass filtering at this stage would remove the information required.

Finally, the task of analysing transient signals with this model means that there is no longer a need for the averaging process, strobed temporal integration, as was used in *aim-mat*. Since this process averages over each cycle of a periodic signal, it is redundant here. The recognition of repeated pulse-resonances, such as in a vowel sound, is significantly better that a single iteration of the pulse, but it is not necessary, as was shown by Lyon (2010a), where for the discrimination of single and repeated double-damped sinusoids, his subjects performed significantly above chance for all, including the single pulses. While the strobing process on *aim-mat* creates averages of a number of periods in a signal and effectively increases the signal-to-noise ratio, by omitting the process the model become less demanding of

computation, and more useful for analysing environmental sounds as not all sounds that occur in nature are periodic. Removing the strobing mechanism provides a new problem, however, because the process automatically aligns the basilar membrane motion in a way that the measured time intervals all start at the same time at a maximum in the BMM.

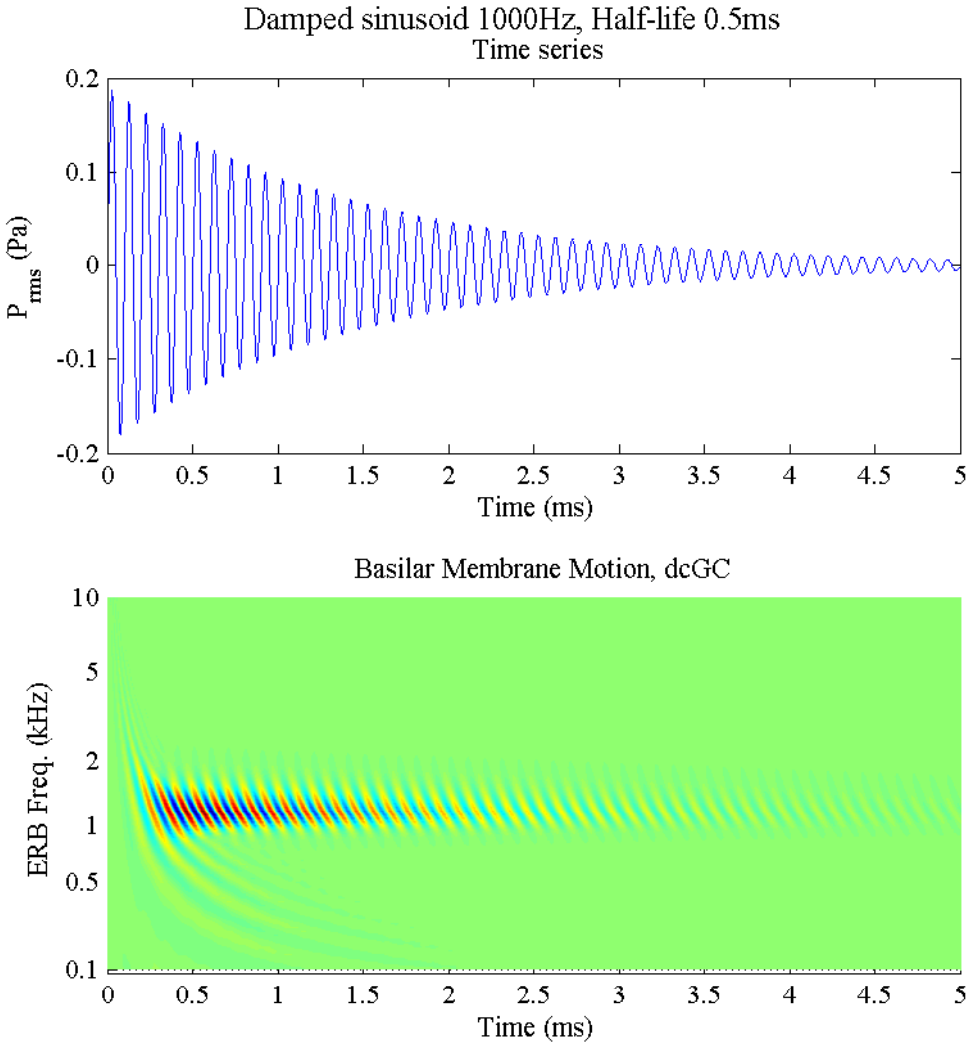

Figure 59: Example of the Basilar Membrane Motion (lower panel) after the dynamic compressive Gammchirp filterbank has been applied to a damped sinusoid of 1 kHz with a half-life of 0.5 ms.

Instead, a method of alignment is used which involves arranging the points of maximum amplitude in each frequency bin so that they occur at the same time

interval point. Other methods of alignment were explored including investigating the group delay in each filter channel or searching for the first point of oscillation of each filter using a click stimulus, but these techniques did little to align the peaks and troughs of each channel in as successful a way as the method used here. The result is a transient version of the Stabilised Auditory Image referred to earlier as proposed by Patterson and colleagues (1995). Here it is simply called the Auditory Image (figure 60).

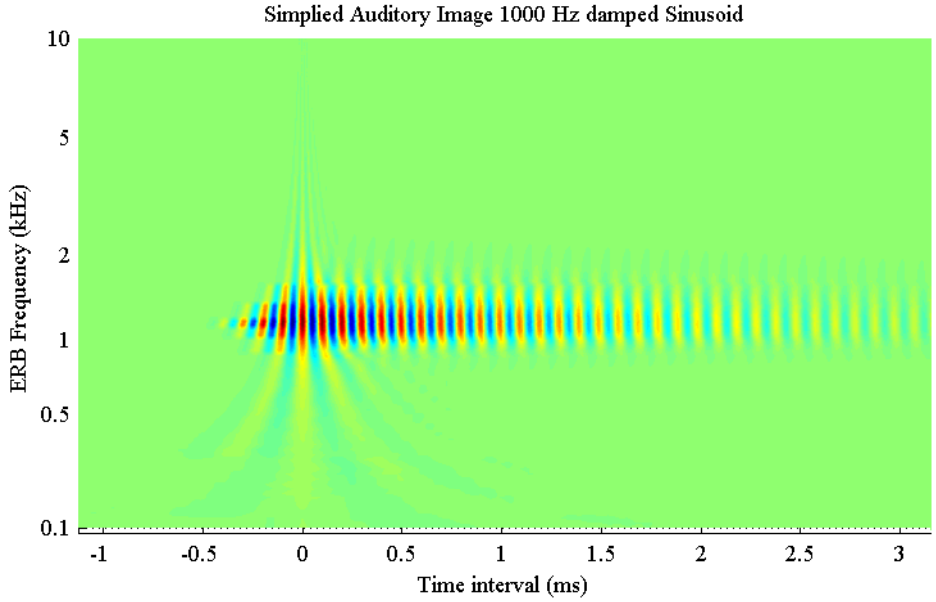

Figure 60: Example of the Auditory Image from tAIM, based on the Stabilised Auditory Image from *aim-mat*. The maximum point in each BMM channel has now been aligned to create the zero-point in the time interval.

The next step is the size normalisation process, and the size-shape image and Mellin image is created using the same method as the original AIM model. The aligned points of the SAI provide the zero-point necessary for the Mellin transform, and each channel is resampled proportional to the centre frequency of that channel. This results in the time axis being converted to a time-interval peak-frequency product. The process effectively stretched out the peaks and troughs of the signal in each channel in order to align them with other channels, and create the vertical lines characteristic of the size shape image. A Fourier transform is taken of each column of the SSI to create the Mellin image. This process produces a pattern that

shows the strengths of the resonances in the SSI. When two signals which have the same resonance relationship are compared, such as those in speech for example, the patterns in the MIs are very similar. In this manner, the MI can normalise for size and thus create a pattern that could be useful for shape identification; however this is not within the scope of this study. The energy bands on the MI show the strength of the relationships between the resonances. The closer and more numerous the MI bands are the closer in frequency the resonances are, until there is only one resonance which creates one large band of energy as is the case in the MI example to follow. Figure 61 shows the SSI and resultant MI of a single damped sinusoid of 1 kHz. A block diagram of tAIM and the size normalisation processing modules is shown in figure 62.

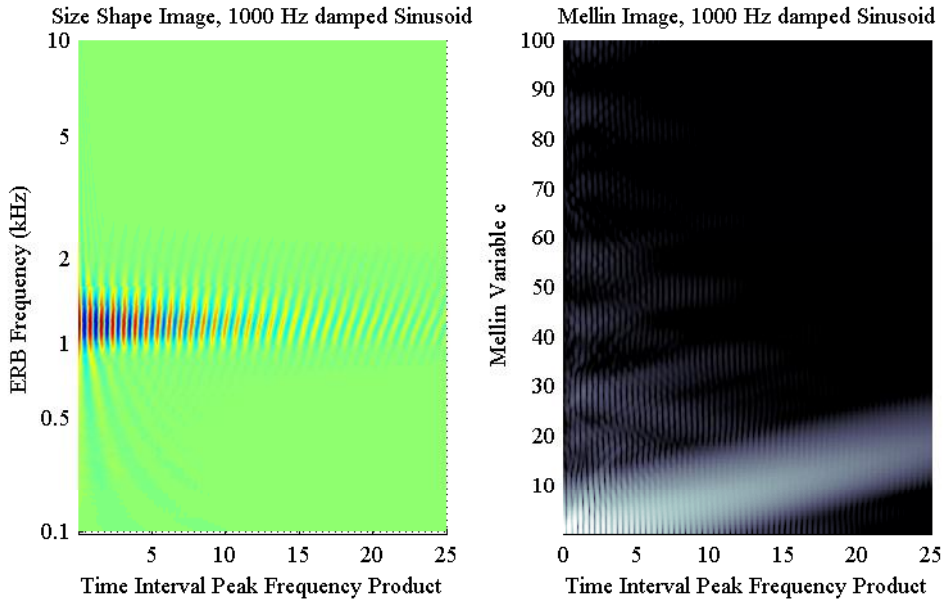

Figure 61: Size Shape Image and Mellin image of the 1000 Hz damped sinusoid. The single resonance in the signal results in an uninterrupted diagonal line in the Mellin Image.

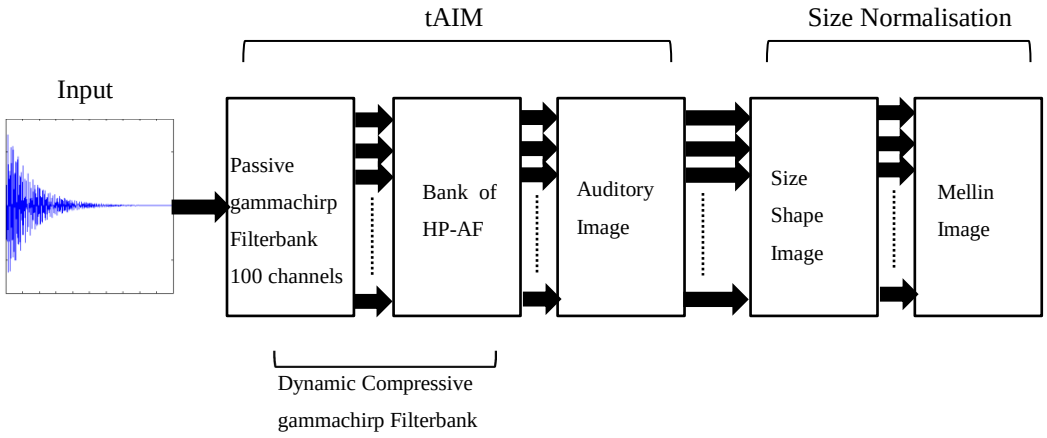

Figure 62: Block diagram of the processing modules of tAIM, followed by the size normalisation computation to produce the Mellin image.

7.2.1 Damped Sinusoids as control sounds

To verify that tAIM is capable of preserving size information, a range of control stimuli was specifically designed to show the effect of changing "size" vs. changing "shape" of an artificial object with known properties. The stimuli that were chosen resemble the stimuli that were used by Irino and Patterson (2002) to fulfil a similar role. But since only a single repeat of each stimulus will be analysed, the stimulus can be simplified considerably. The stimuli of choice are double-damped sinusoid, which have some characteristics of natural vowels: two sine waves, each with a frequency that is designed to loosely resemble a formant; added together to make up one double sinusoid. Both sine waves are exponentially damped by a fixed time constant which is measured by the time that the envelope needs to decay by 50%. This time constant is called half-life. Figure 63 shows an example stimulus and its spectrum.

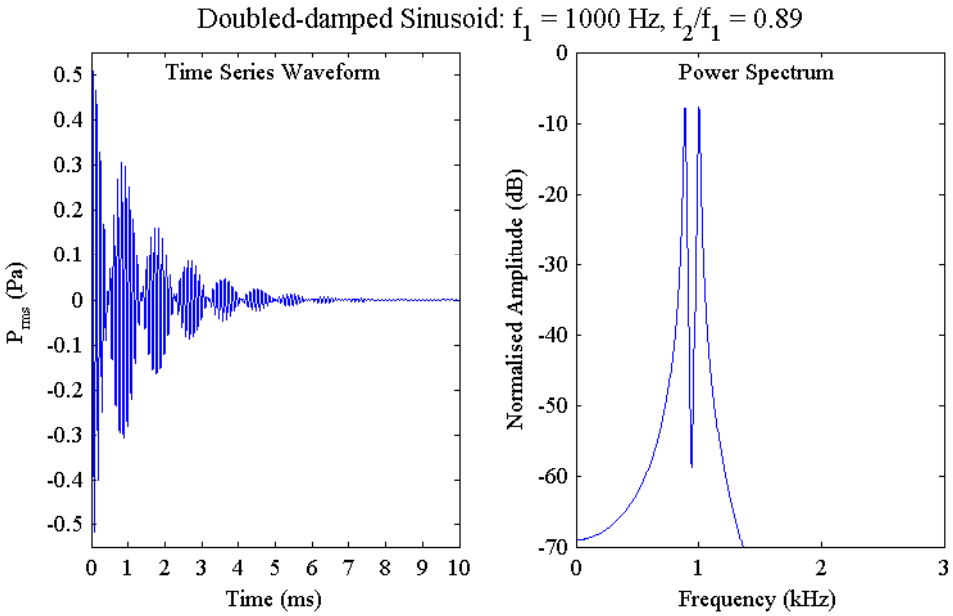

Figure 63: The left panel shows an example double damped sinusoid with a half-life of 1ms. The spectrum in the right panel shows the two frequencies, but does not capture the envelope characteristics.

The double-damped sinusoid simulates crudely a single glottal pulse of a vowel. The signals used here are exponentially decaying sinusoids with a half-life of 1 ms, with the lower sinusoid, $f_{1,}$ increasing in frequency from 1000 to 1400 Hz and the higher, f_2, sinusoid increasing from 2000 to 2800 Hz. Thus, 12 stimuli were generated with different values of f_1 and different combinations of ratios f_2/f_1. Figure 64 below shows the time series waveforms of two examples of signals with the same f_2/f_1 but a different f_1, demonstrating that while the frequencies change the time structure of the envelope remains the same.

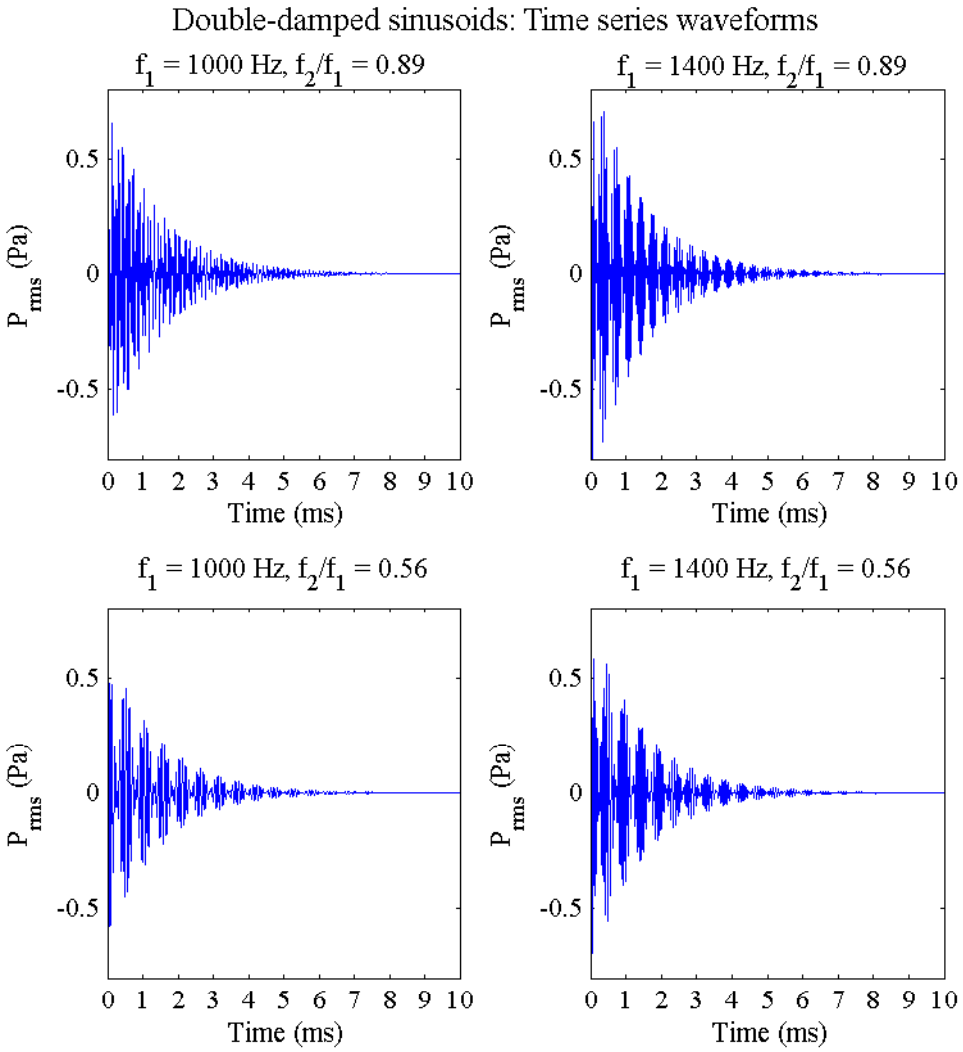

Figure 64: Examples of time series waveforms with different ratios of f_2/f_1, demonstrating the unchanging shape of the envelope despite different f_1's.

Table 11 shows the relationship between the formant frequencies as a ratio between the two sinusoids, and also indicates the frequencies used in the stimuli. These ratios, f_2/f_1, represent changes in spatial frequency between formants and thus simulate a change in vocal tract shape. The change in f_1 simulates a change in VTL, and so as the formant frequency increases the size of the simulated VTL would decrease accordingly.

Musical ratio	Mathematical ratio	f_1	f_2
1 octave = 12 semitones	$f_2 / f_1 = 2$	1000 Hz	2000 Hz
		1200 Hz	2400 Hz
		1400 Hz	2800 Hz
¾ octave = 8 semitones	$f_2 / f_1 = \sqrt[4]{2^3}$	1000 Hz	1682 Hz
		1200 Hz	2018 Hz
		1400 Hz	2355 Hz
½ octave = 6 semitones	$f_2 / f_1 = \sqrt{2}$	1000 Hz	1414 Hz
		1200 Hz	1697 Hz
		1400 Hz	1980 Hz
¼ octave = 4 semitones	$f_2 / f_1 = \sqrt[4]{2}$	1000 Hz	1189 Hz
		1200 Hz	1427 Hz
		1400 Hz	1665 Hz

Table 11: Relationship between damped sinusoids

The stimuli described in table 11 were processed by tAIM, and SAIs were created for each signal. The columns of figure 65 represent a 'size axis', and show the frequency of the lower sinusoid f_1 increasing from left to right by 200 Hz each instance, while retaining the ratio between the sinusoids, i.e. this axis simulates a decrease in the size of a vocal tract for the same vocal tract shape. The rows represent the 'shape axis', and show the ratio between f_2 and f_1 decreasing from top to bottom by quarter of an octave each instance. This shape change is illustrated by a difference between how far apart the formants are in the activity pattern in the SAI. The chosen frequencies of the sinusoids are not intended to represent actual formants of vowels; they were chosen for explanation purposes due their clarity when shown in an SAI.

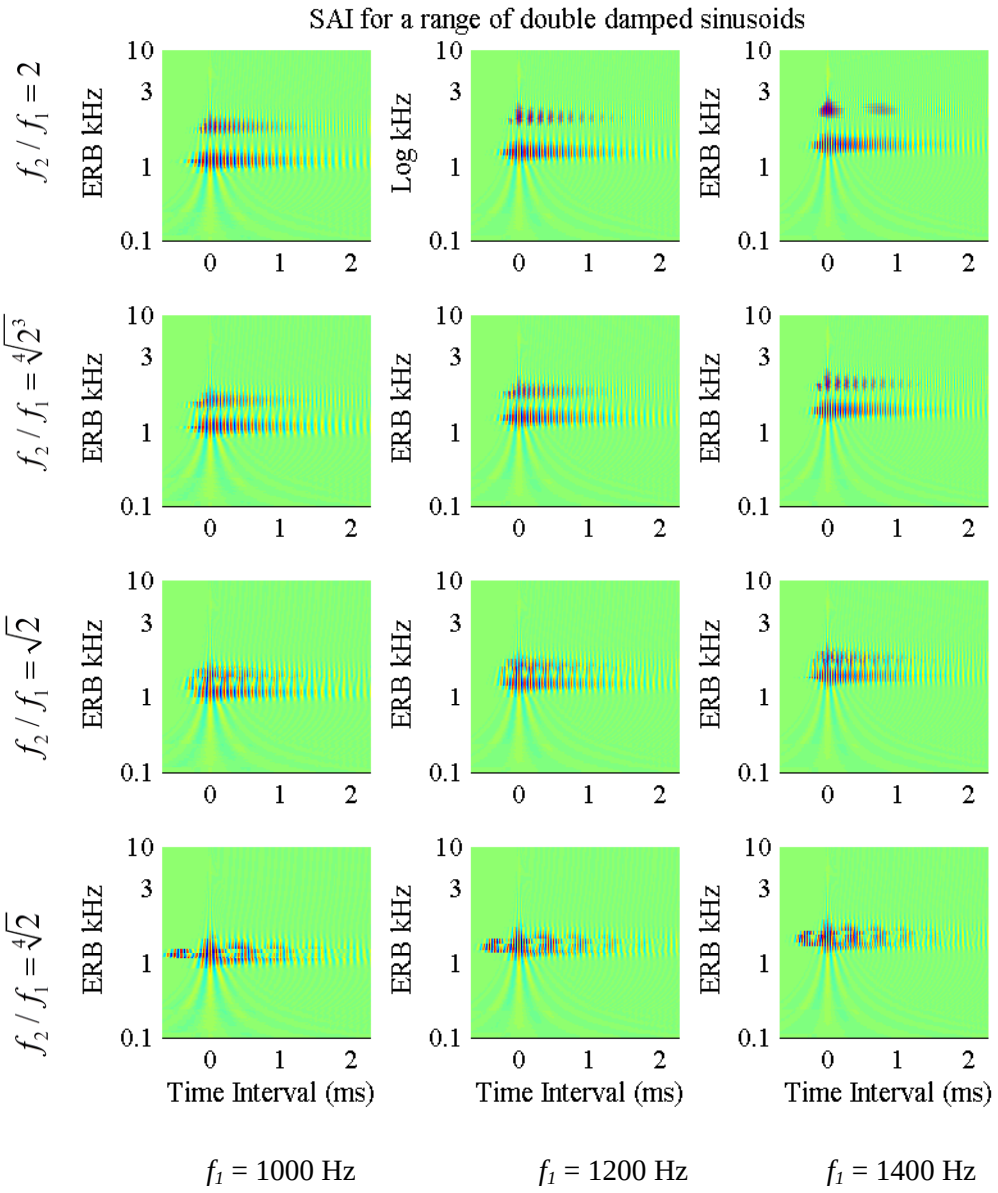

Figure 65: Table of SAIs corresponding to table 9 above. The rows represent shape changes associated with the ratios between f_1 and f_2. The columns represent an increase in f_1, and thus a decrease in size of the simulated object. In all panels, the absolute frequency values of the formants can be seen on a log axis of 200 channels.

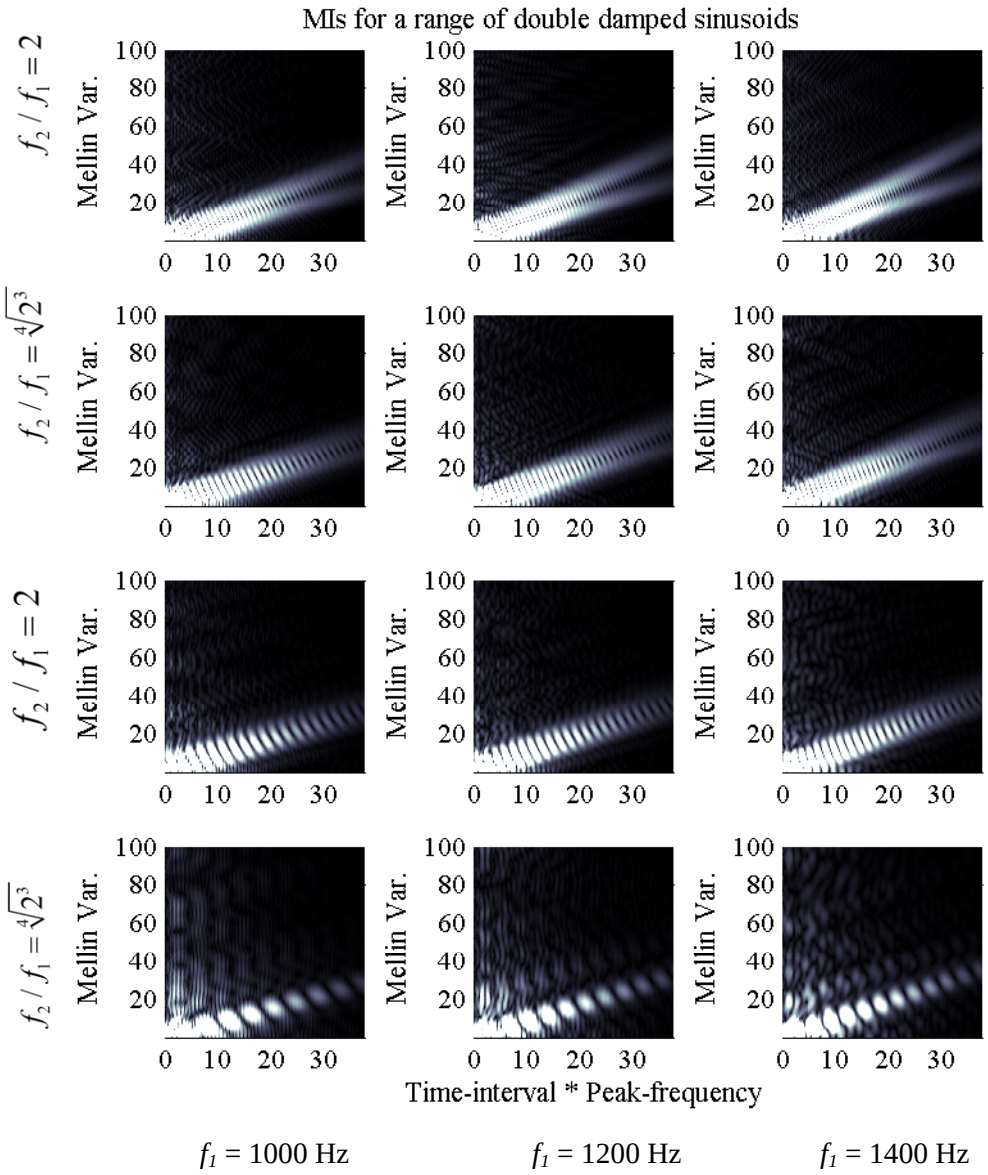

Figure 66: Mellin images of the 12 control sounds. Each panel shows the Mellin power of one stimulus with a particular combination of f_1 and f_2/f_1. Similar to the SAIs in figure 50, the f_2/f_1 changes for each row, and the value of f_1 increases with each column. The change in the absolute frequency values can no longer be seen, and the MI shows patterns corresponding to the spatial frequencies within the SSIs which was calculated from the SAIs in the previous figure. Thus the images have been normalised for size, and as such this figure demonstrates the successful application of the Mellin transform on the output of the tAIM.

The SAIs show how the images change for both size and shape for the chosen stimuli, proving that for precise control stimuli, tAIM can preserve size and shape

information. Figure 66 shows the Mellin images for the same selection of control stimuli, created by performing a Fourier transform along each column of the SSI and plotting the absolute power of the result. Again, the columns represent stimuli that change in size and the rows represent those that change in shape. The change in the f_2/f_1 ratios between the stimuli is clear with different patterns occurring along the shape axis, but there is very little variation along the size axis, with this small variation due to the ERB frequency scale which is more low-frequency weighted than a standard logarithmic scale. The patterns of energy are repeated for each size due to the same difference between f_1 and f_2 being used. As clarified earlier, this difference is transformed into a spatial frequency by the Mellin transform. The shape axis now shows how the spatial frequency changes with each vowel, but disregards the absolute value of the formants, demonstrating how size has been successfully normalised by the transformation using the output of the tAIM.

7.3 Simulated vocal stimuli

The first stimuli for testing the capabilities of tAIM are simulated vowel sounds with frequencies that represent specific vowels used within the British English accent. It has previously been mentioned that there is the possibility of a size normalisation system in play in the auditory system for ideal use with communication sounds. It is known that despite the changes in speaker size and resonant frequencies of vocal tracts of different lengths, humans can distinguish between and identify many different vowels from a range of speaker sizes, as well as some vowels that have been scaled to outside naturally occurring sizes (Smith, 2005).

The double-damped sinusoid signals used earlier were scaled accurately to create ideal signals. To test tAIM's ability to analyse less precise stimuli, a range of simulated vowels were created according to data put together by Borden and colleagues (Borden et al., 2003) as collected by Peterson and Barney (1954). Table 12 below provides the first and second formant frequencies from three common vowels obtained by averaging over the recordings of the vowels spoken by men,

women and children. For ease of identification, the vowels are bound by the letters 'b' and 'd' to create a word for each. The words are 'bead', 'bud', and 'bawd', and the corresponding IPA (International Phonetic Alphabet) symbol for the vowels are shown in the adjacent column. Figure 67 shows the vowel formant frequencies on a log scale and it can be clearly seen that from the man to the child, the formant frequencies for each vowel increase in value.

Word	IPA	Formants	Men	Women	Children
Bead	/i/	F1	270	310	370
		F2	2290	2790	3200
Bud	/ʌ/	F1	640	760	850
		F2	1190	1400	1590
Bawd	/ɔ/	F1	570	590	680
		F2	840	920	1060

Table 12: A sample of the results obtained in a study by Peterson and Barney (1954) as displayed in Borden et al (2003). Three words and the corresponding IPA vowels are shown above, and the averages of the F1 and F2 formants for each vowel sound for groups of men, women and children.

Word	IPA	Ratio	Men	Women	Children
Bead	/i/	F2/F1	8.48	9.0	8.65
Bud	/ʌ/	F2/F1	1.86	1.84	1.87
Bawd	/ɔ/	F2/F1	1.47	1.56	1.56

Table 13: This table shows the similarities between the ratios of F1 and F2 for the different vowels spoken by men, women and children.

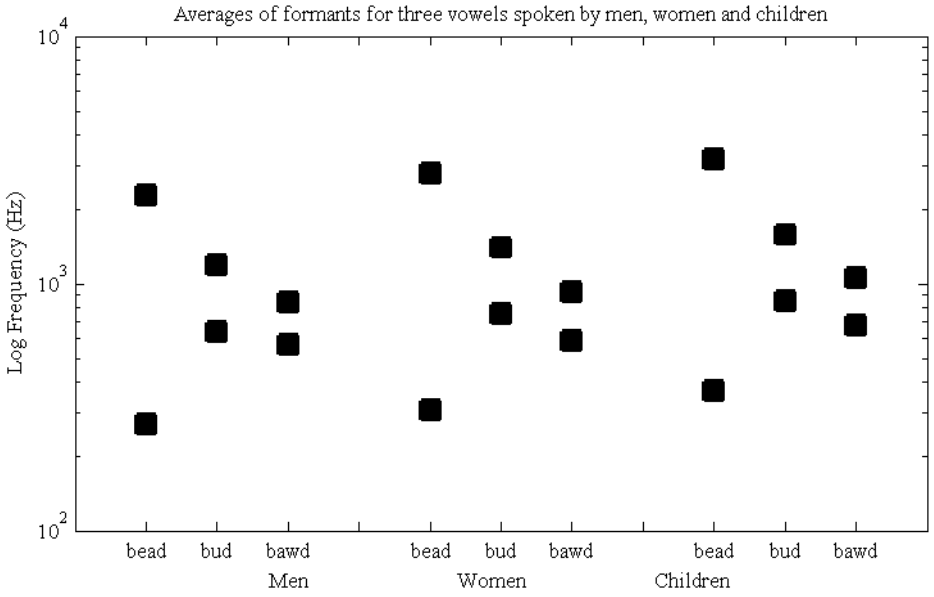

Figure 67 shows the F1 and F2 values from table 12 above in order to display how the relationship between the formants changes very little despite the difference in the size of the speaker. The averages for the men have lower frequency values than those of the children on account of them having a larger VTL.

From the tables and figure above, it can clearly be seen that the resonances shift as a unit depending on the size of the speaker. The ratio between the resonances is restricted to a very small range: for the vowel /i/ in 'bead', f_1 ranges from being between 11.1 – 11.7 % of f_2. The first formant f_1 in the vowel /ʌ/ in 'bud' falls between 53.5 and 54.3 % of f_2. Finally, the vowel /ɔ/ in 'bawd' is the most variable but is restricted to between 64.1 and 67.8 % of f_2. The relationship between the vowels and the speaker size is linear, as seen in figure 68. The vowels are intelligible regardless of the frequencies of the formants. Turner et al (2009) discuss the data provided by Peterson and Barney (1952) with reference to the relationship between VTL and GPR as the speaker matures. It is the ratio between the formants that defines them, and which allows the vowels to be scaled up or down to change the size of the speaker, as was performed in an experiment by Smith and colleagues (2005). It is the ratio that defines their identity, but there is a degree of variability with the ratios (Borden et al,. 2003).

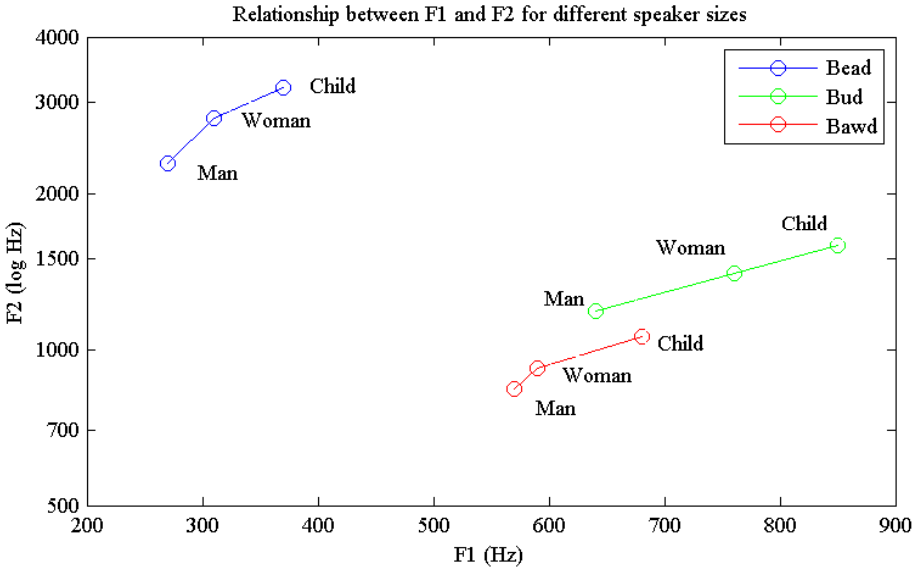

Figure 68: The relationship between the formants and the speaker size is roughly linear on a log F2 scale, with a small degree of variation.

An example of the simulated continuous vowel sound 'aw' was shown in figure 2 earlier, to show how the time series waveforms can differ despite the same vowel being spoken. The glottal pulse rate is of a higher frequency in that of the female voice, and thus the length of the resonance is shorter, but there are small similarities in the resonances between each glottal pulse. Single pulse-resonance versions of each vowel in figure 68 were created in MATLAB, and were sent through tAIM. Each signal had a half-life of 50 ms and was set to 70 dB SPL, and the parameters of tAIM were as before; the dcGC filterbank was assigned 100 channels to filter from 100 Hz and 10 kHz. Figure 69 below displays the SAIs for each of the nine signals, three vowel sounds and three vocal tract lengths. The rows show the change in ratio between the resonance according to the vowel spoken, and the columns show the resonances shifting upwards as the speakers decrease in size. The same layout is followed with the Mellin images in figure 70, and the columns show that size normalisation has been successful. There is almost no variation between the patterns for the MIs of the men, women and children for the same vowels. Table 14 shows the calculated SCF values for each of the vowels indicating that as the length of the vocal tract decreases the SCF increases, as expected.

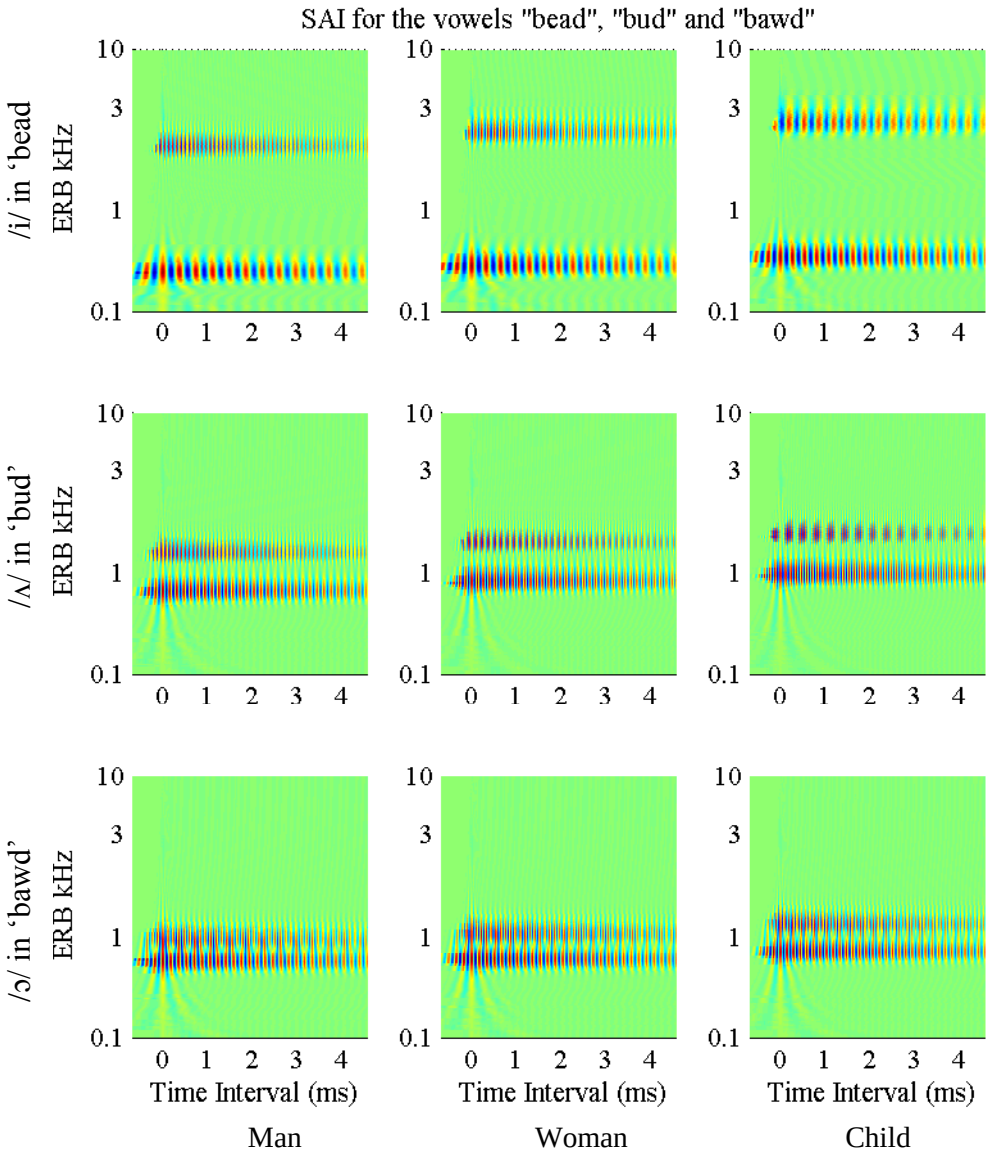

Figure 69: Simplified Auditory Images of the vowels in 'bead', 'bud' and 'bawd', as simulated from results of those spoken by men, women and children as displayed in Borden et al (2003). The resonances show how the vocal tract shape changes with each vowel, and the columns show how the resonances shift upwards as the vocal tract length decreases in the speakers from man to woman to child.

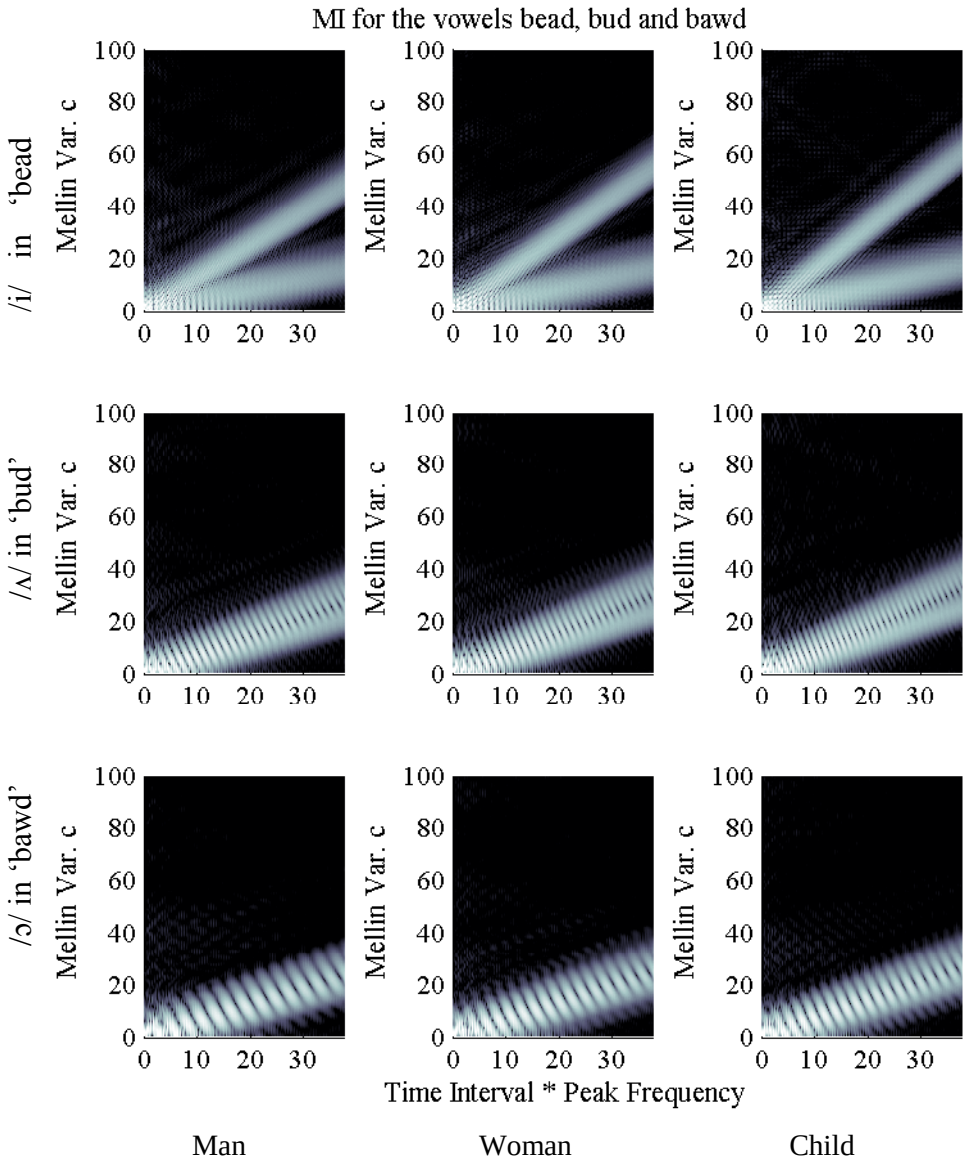

Figure 70: Mellin images of the simulated vowels from 'bead', 'bud' and 'bawd' as spoken by men, women and children. The rows show the change in shape due to the different vowels, and the columns represent the vowels as spoken by speakers of different size. The limited variation among the columns proves the success of the size normalisation of the output from tAIM.

SCF of vowels	Man	Woman	Child
/bead/	1644.7 Hz	1940.9 Hz	2185.3 Hz
/bud/	1207.4 Hz	1391.1 Hz	1543 Hz
/bawd/	974.5 Hz	1030.8 Hz	1162.6 Hz

Table 14 Spectral Centroid Frequencies of the vowels uttered by men, women and children. It shows that as the size of the person, i.e. the length of the vocal tract decreases, the SCF increases in value.

7.4 Simulated underwater sounds

The tAIM was then used to process simulated underwater scattered pulses used previously in the study by (Fox et al., 2007). The pulses were generated to simulate those reflected/scattered by infinitely long water-filled or air-filled cylindrical shells submerged in water, in a manner not unlike how underwater mammals echolocate. The incident pulse was a 2-cycle cosine weighted 1 kHz pulse of unit amplitude with the majority of its energy showing up in the 0-2.5 kHz region of an amplitude spectrum. The pulse was positioned normal to the cylinder and calculated at a fixed distance of 200 m from the centre of the cylinder. These stimuli were not used in a psychoacoustic test of size discrimination to measure how humans could fare at telling them apart; they were generated in order to test the model (Fox et al., 2007). However, it must be noted that when listened to by the author, the signals representing cylinders with radii under 1m became very difficult to discriminate between, and almost impossible for the smallest. Perhaps this is due to a limit in size discrimination abilities of the listeners, but that is not our question. All the other sizes had transient pitches that allowed for easy size discrimination and so only a selection of these signals used will be presented as stimuli for this account of tAIM.

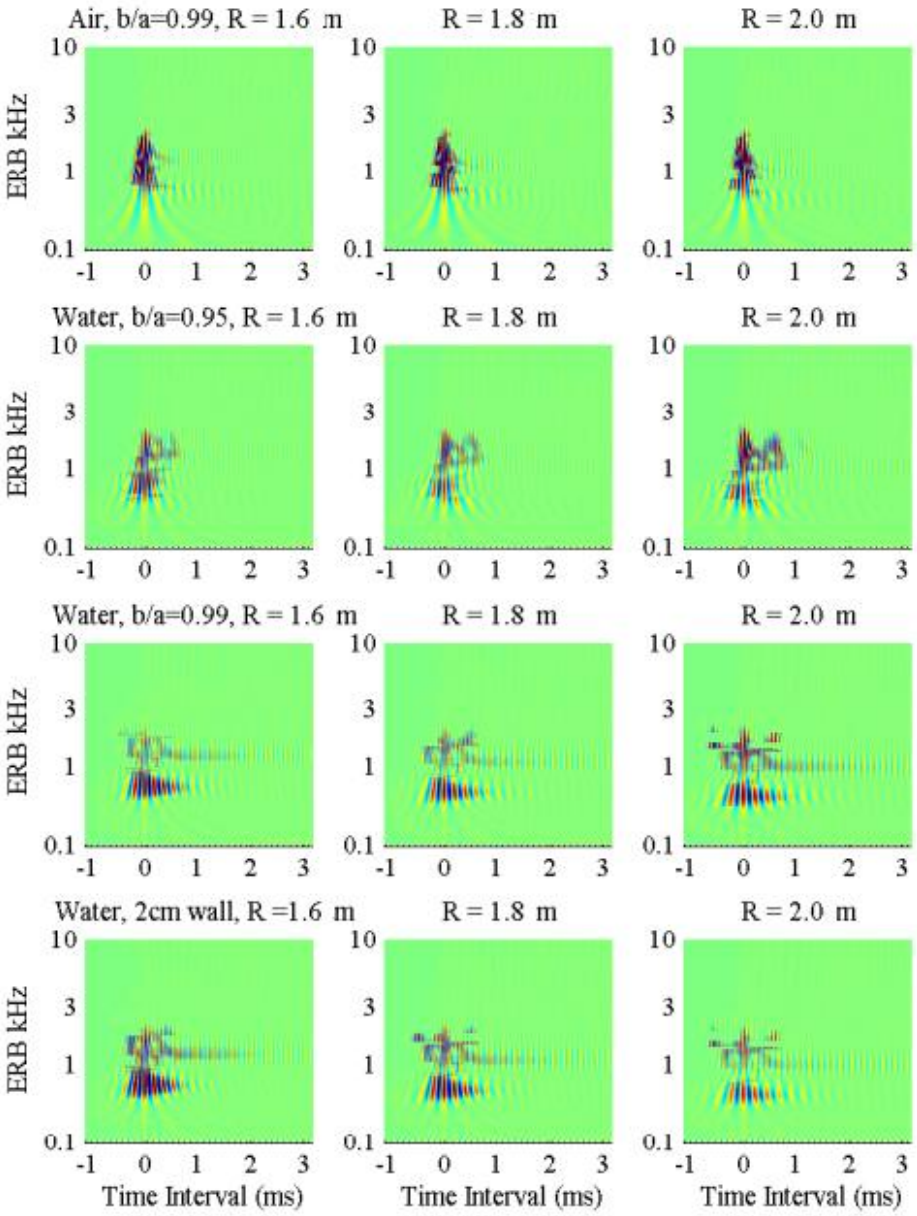

Figure 71: The Simplified Auditory Images of a range of different simulated scattered signals from underwater cylinders. *b/a* is the ratio between the inner and outer wall thereby giving an indication of the thickness of the cylinder wall. The radius refers to the outer wall radius of the cylinders. The rows represent the different types of cylinders: air-filled cylinder with a *b/a* of 0.99, water-filled with a *b/a* of 0.95, water-filled with a *b/a* of 0.99, and finally a water-filled cylinder with a constant wall thickness of 2 cm. The columns show how the resonances in the SAIs decrease in frequency as the outer radius increases.

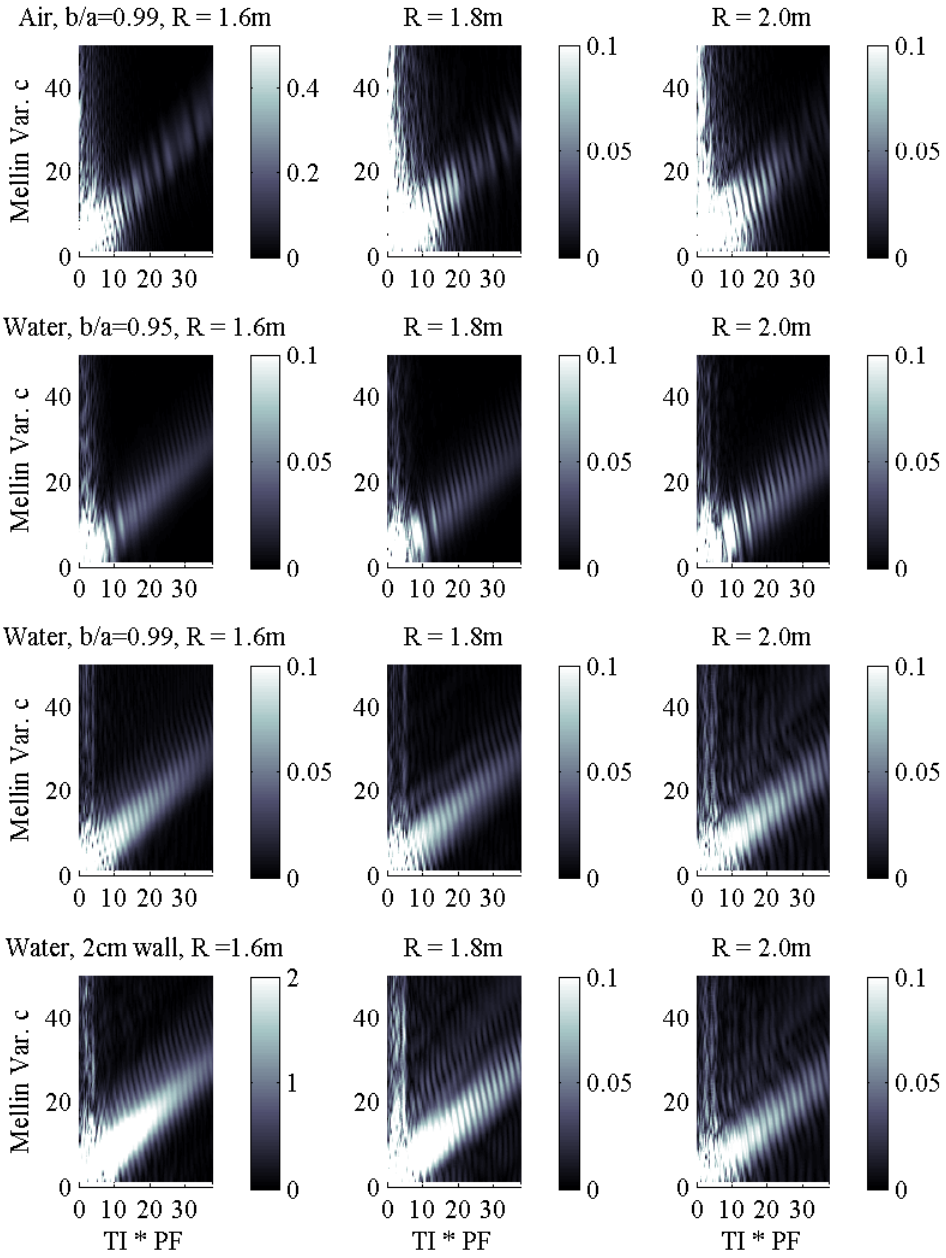

Figure 72: The Mellin images corresponding to the AIs of figure 70. The cylinders with the larger radius show more effect of size normalisation due to the strength of the resonances, and so they are the cylinders shown here. There are distinct similarities between the patterns created by the Mellin transform along the rows, for cylinders of the same type.

Figures 71 and 72 show the SAIs and the Mellin images respectively of scattered pulses from different types of cylinders with outer wall radii of, from left to right, 1.6, 1.8 and 2.0 metres. The first set of stimuli was produced from an air-filled cylinder with a very thin wall, where the ratio of inner radius of the wall to outer radius was b/a=0.99. The other three sets were water filled, the first two with ratios of inner to outer wall radii of b/a=0.95, and b/a=0.99, and the final with a fixed wall thickness of 2cm, but the radii were varied (i.e. b=a-0.02).

Similar to the increasing resonances with decreasing vocal tract size, the SAI shows the resonances of the scattered pulses increasing in frequency as the outer radius of the cylinders decrease, for all cylinder types. The length of the resonances on the horizontal axis decreases with decreasing radius size for all cylinders, which could explain the limited discrimination abilities of the authors for the smaller cylinders. The range of frequencies excited by the cylinder resonances is different for each, with the water-filled cylinders of b/a=0.99 and constant wall thickness of 2 cm showing the widest frequency range and the most cumulative energy. The air-filled cylinder has the least energy of all the cylinders, an unexpected result considering the large impedance at the interface which would generate more scattering from the cylinder wall. This may be due to resonances occurring outside the frequency range considered, or the alignments methods employed by tAIM.

Normalising for size of cylinder, the Mellin images now show information about the shape of the cylinders. Similarities between the horizontal bands of energy in the power images can be seen for almost all of the cylinder types. For example, five uniformly placed bands of energy are visible in the power images of the air-filled cylinders. Likewise, there are similar curved bands in the MIs for the larger of the water-filled b/a=0.99 and 2cm wall cylinders. These similarities become less distinct in the smallest cylinders (radius 0.8 m), where the bands blur together for the cylinders filled with air and the water-filled cylinders of b/a=0.99.

The simulations of the cylinders do not show resonances that are as long as those in the vowels created above. The Mellin transform has been successfully applied to vowels (Irino and Patterson, 2002) which had a glottal pulse period of 10 ms, and were shown to have resonances which rang for at least half that length.

The vowels that were simulated here had a half-life of 1 ms and so were almost 5 ms long in their resulting SAIs. The SAIs below show that after alignment, the resonances for the cylinders do not ring for longer than 2.5 ms. There are difficulties with some of the cylinders with smaller radii, but those with resonances that last longer create the clearest Mellin images. There could be several reasons for this difficulty with the Mellin transform here, but they will be discussed later in this chapter.

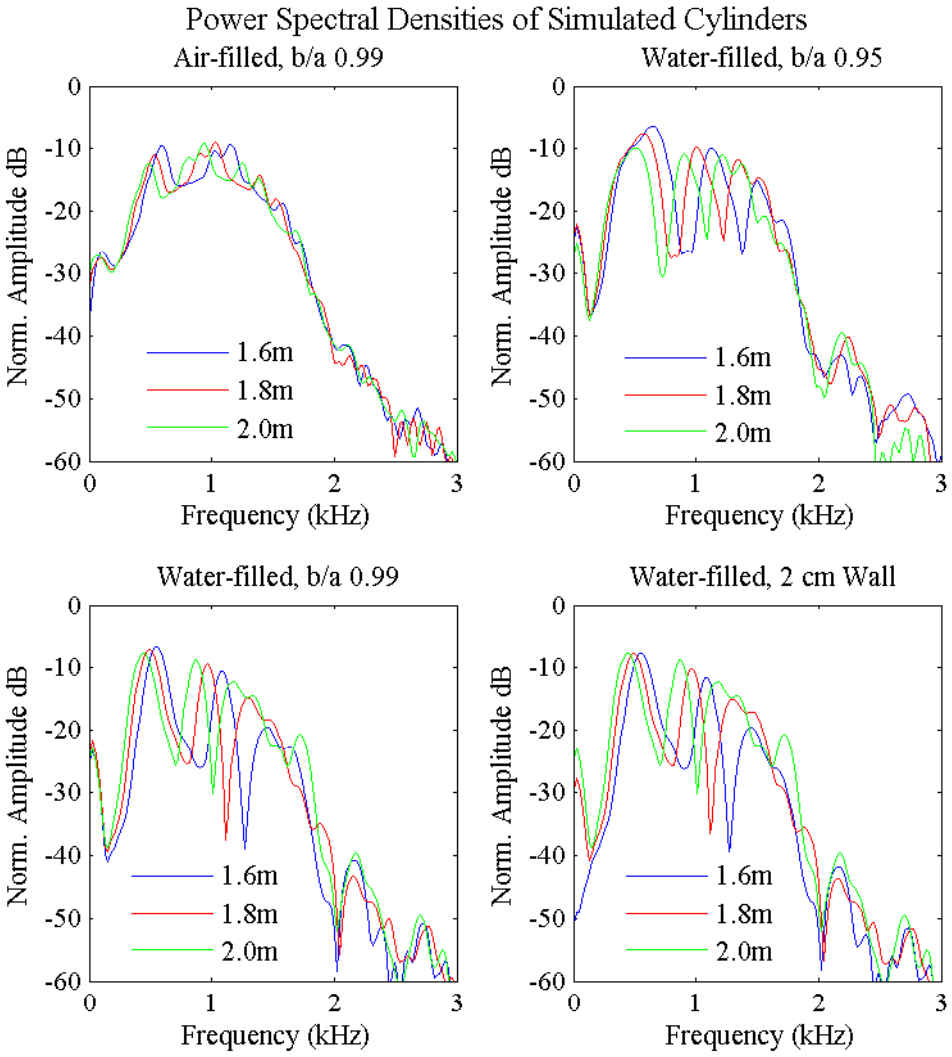

Figure 73: The Welch power spectral densities of the cylinders to be processed by tAIM. Each panel represents one 'shape' of cylinder, i.e. the wall thickness. The green line in each panel represents the spectrum of the cylinder with the biggest radius, and for each shape, the peaks are lowest in the spectrum.

Figure 73 shows the Welch PSD plots for the cylinder signals processed by tAIM. In each panel, the green line corresponds to the largest signal. When listening to the signals, the author was satisfied with their very clear differences in transient pitch. However, the PSDs appear to be very similar and surprisingly, there seems to be more energy in the higher frequencies than in the lower frequencies for the larger signals, especially for the water-filled cylinders in the two lower panels. Table 15 shows the SCF calculations from the cylinders and confirms this; the SCF values are contrary to expectations and to the results shown earlier for the spheres. The values increase with increasing size, instead of decreasing.

The link between SCF and timbre or brightness of a sound was discussed in section 2.3, and how it has been proven to be independent from pitch in periodic signals. It is possible, that due to the simulation method, and the fact that the SCF values are relatively similar, that the signals are all sufficiently 'bright' in timbre due to their high frequency content that SCF has little relevance when it comes to transient pitch and size discrimination. But it is also possible that the relationship between SCF and transient pitch of non-periodic signals are also independent. This requires further investigation if this is the case, to be done as possible future research, as the table clearly shows a result contrary to expectations from signals that sound so clearly different.

SCFs of Cylinders	**1.6 m radius**	**1.8 m radius**	**2.0 m radius**
Air b/a 0.99	1005.8 Hz	1011.8 Hz	1010.6 Hz
Water b/a 0.95	961.9 Hz	989.5 Hz	1012.4 Hz
Water b/a 0.99	914.4 Hz	940.6 Hz	976.1 Hz
Water, 2cm wall	921.9 Hz	958.3 Hz	976.1 Hz

Table 15 The Spectral Centroid Frequencies of the cylinders. In all cases, the general trend is an increasing SCF for increasing radius, contrary to the SCF values of the spheres and the general understanding of the relationship of SCF to size. This is most likely due to the methods in which the cylinders were simulated.

7.5 Recorded Polystyrene Spheres

The final set of signals to be analysed by tAIM is the set of recorded polystyrene spheres used earlier in the size discrimination experiment. The signals are real, in-air sounds which have been shown to be easily told apart in a simple size discrimination task. All sphere signals have two clear resonances in each of their spectra, as can be seen in the spectral density plot below in figure 74. These resonances shift upwards in frequency in a linear fashion as the size of the sphere decreases. Despite all these factors the signals have proven difficult to scale up or down in a manner similar to the simulated vowel, as shown in experiment 1 (section 4.2). The study gathered five people and asked them to adjust the playback sample-rate of the spheres in order to force them to be the same transient pitch as the reference sphere. The results showed there to be a lot of difficulty in the task, with most participants reporting diverging resonances during scaling which added an element of confusion to the task. It was also concluded that the method of scaling the signals was not ideal.

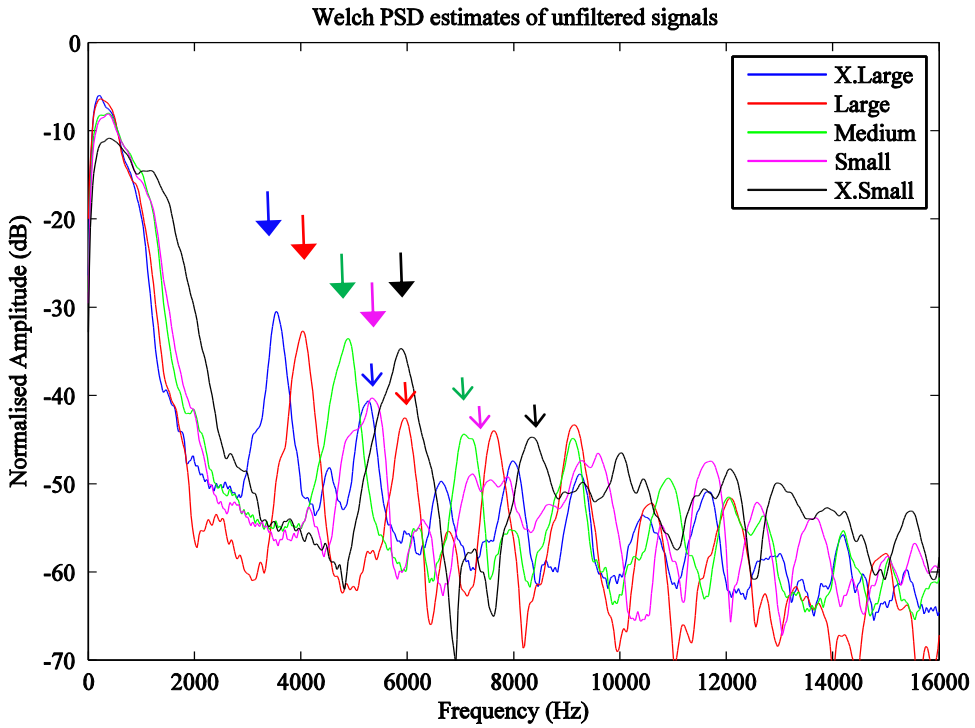

Figure 74: Welch Power Spectral Density (PSD) plots of each of the five averaged signals. Each coloured line represents the PSDs for the five different sized spheres.

The coloured arrows point to the first (🡻) and second (🡻) resonant peaks for each PSD.

The AIs and Mellin images of the polystyrene spheres can be seen below in figure 75. The AIs, with ERB frequency from 100 Hz-12 kHz on the y-axis, show how the resonances shift upwards with decreasing sphere size. The first mass of energy is the energy under 1 kHz in the spectrum, and makes up most of the energy in the auditory images. The important formant, F1 (shown with a 🡻 in figure 74) is just visible at the top of each auditory image, marked with an arrow. Limits were placed on the colours of the AIs so that F1 could be seen, as the magnitude of F1 compared to the wide-band energy of F0 was too low to be seen. Compression in the gammachirp filtering was expected to counter this, but instead the limits were placed as a post-filtering image compression in order for the F1s to be made visible. The Mellin images are displayed in the right-hand column of figure 75, and while all seem to be similar in shape with consistent diagonal bands of energy crossing each other. However, due to the low magnitude of F1 the pattern detail and size is limited. A small anomaly in the alignment method has occurred, causing a small jump in the alignment of the points of maximum energy. This is a point to look into in the future as an improvement of the method, but the anomaly does not present any problems for the purpose of this study.

The difference in magnitude between F0 and F1 is shown more clearly in the time-zero extractions from the SSI images in figure 75. The frequency scale is ERB, and compared to the PSD plots in figure 74, there seems to be very little change in the relative magnitudes of F0 (100 – 1000 Hz wide energy band) and F1, contrary to what is expected of a dynamic compressive filter. It is also clear that the resolution of the high frequencies is not as clear, which is expected due to the widening of the filters on the basilar membrane, however it is not representative of the results shown in the psychophysics experiment when testing the size discrimination of heavily high-pass filtered signals. Also, F2 is not evident in either figure 75, or 76, despite the values being lower than the filterbank's upper limit of 12 kHz. This is an issue with the filterbank, and could not be avoided in this study.

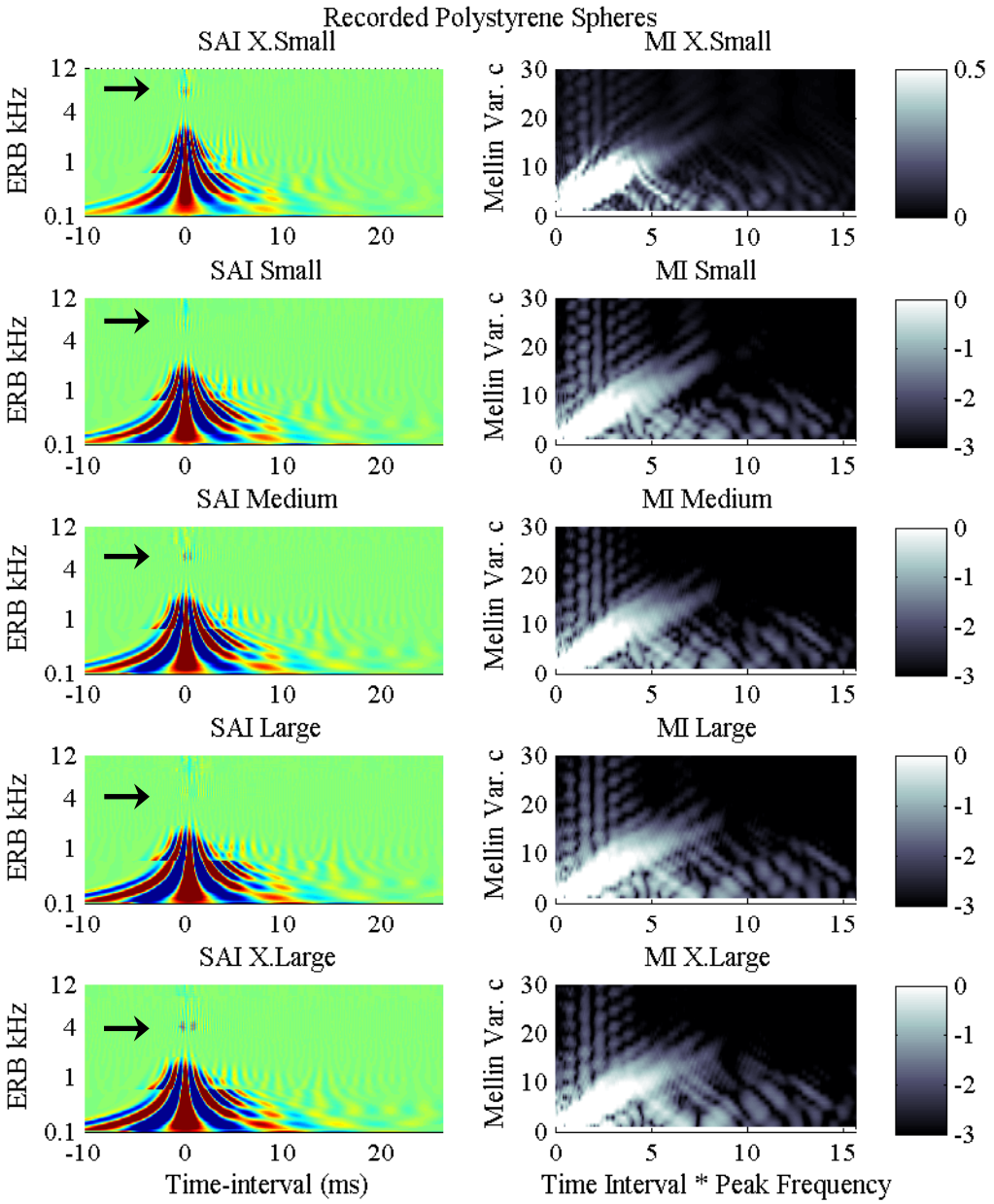

Figure 75: SAIs and MIs of the recorded polystyrene spheres. The large band of energy from 0.1-2 kHz is from the wide band of low frequencies, F0, a result of the impact on the sphere. The horizontal line in the F0 is a small anomaly due to the alignment method, but this does not affect the important F1 and F2 resonances. In these images, there is a limit placed on the energy range so that F1 is visible; marked with an arrow near the top of each SAI. Without the energy limits the F1s would not be visible. The downward shift in F1s with increase in sphere size is visible in the SAI, and there are similarities in the MIs. F2 is not visible.

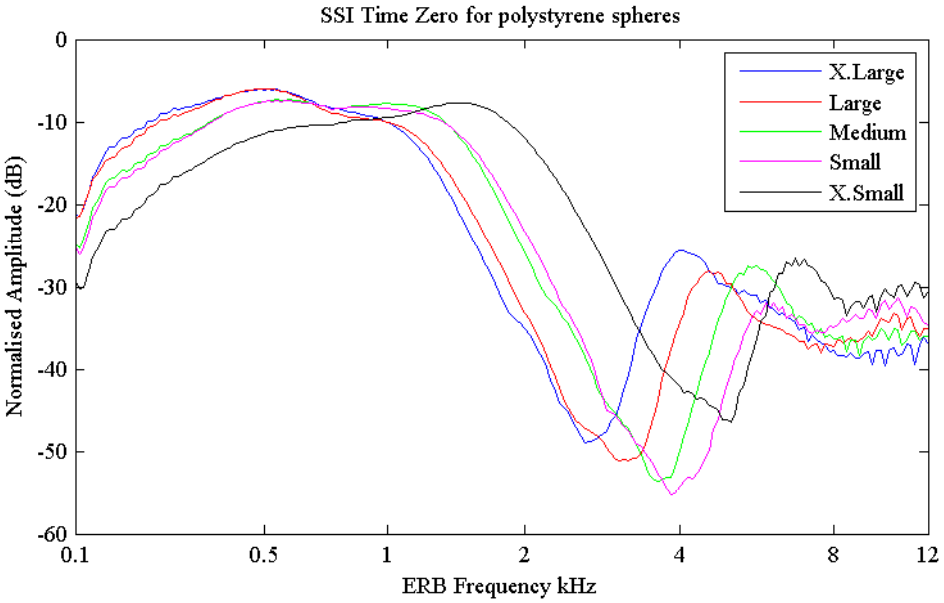

Figure 76: Time Zero extraction from the SAIs, showing the magnitude of the F1s next to the low frequency bands. It appears that the compression of the dcGC has done little to boost the higher frequencies; there is almost 20 dB difference between the low frequency band and F1. Figure 72 shows the same difference using PSDs of the signals. The absolute measure of frequency is not accurate, since the F1 values are ~1 kHz higher than the PSD estimates, and F2 is not shown here despite an upper limit of 12 kHz.

tAIM has shown to be capable of normalising for the size of a range of simulated and recorded signals, using the Mellin transform applied to the aligned output of the dynamic compressive gammachirp filterbank. In a similar manner to the Mellin size normalisation of vowels using AIM, the transform analyses the relationships between the resonances in order to normalise for the position of the resonances in the frequency spectrum. In this way, the assumption of the Mellin transform for vowel sounds and thus in this case with tAIM and transient signals is that size is dependent on a shift in the spectrum, and so far the signals that have been analysed have shown that this is the case for their different sizes.

8. Size discrimination using tAIM

It is now apparent that tAIM can replicate the size normalisation capabilities of AIM by creating shape specific Mellin images of transient signals. However, to the author's knowledge, there has never before been an attempt at using Mellin images to perform size discrimination. The purpose of this research is to analyse the output of an auditory model to do carry out a size discrimination task based on psychophysical results. The model is based on the assumption that size information depends solely on the position of the spectral envelope, and that the envelope moves proportional to the size of the object. The psychophysics experiments carried out in this study have already shown that manipulating the spectrum in this manner influences the listener's perception of size, and also that size information may be found more easily in certain regions on the spectrum but that the information can be extracted from other regions. Thus, the results of the experiment will determine which aspects of the output of tAIM should be analysed along with alternative cues that would be otherwise unavailable if the analysis was to be carried out on a PSD of the signal alone.

A short revision of the results reminds the reader that the most important cue was found to be the differences between the most prominent resonances. The SCF was found to be related to this, where the greater the difference in SCF of the spheres, the higher the mean scores for correct discrimination. Scaling the signals to have the same F1 in order to uncover any alternative cues that did not relate to frequency information fooled the listeners into choosing the signal with the lower SCF as the larger object, an unexpected result if the signal was correctly scaled, and a fact which led to the conclusion that although the resonances seemed to be linear in relationship, a linear scaling method was not sufficient. Scaling the signals to

sound larger or smaller than their comparison showed that transient pitch was the most important cue, where the signal with the lower SCF was chosen as the larger signal and F1 was the most prominent resonance. Filtering the signals showed how fine-tuned the auditory system is to limited spectral and temporal cues. Signals that were high-pass filtered to either include F1 or had a cut-off above it still contained enough information for above chance results in correct discrimination. And signals without F1 performed significantly better than signals with F1, showing that size discrimination is still possible using only the high frequency content. The testing of the size discrimination of the HPF and BSF signals produced the most conclusive results showing that it is the difference between the comparison signals' most prominent resonance that provides the best cue for size discrimination.

tAIM was created to be both computationally inexpensive and to retain as much of the spectral information as possible after using the dcGC filterbank. The dcGC filter contains a high-pass asymmetric which compresses the loud low frequencies in order to boost the important higher frequencies, which in terms of the evolution of speech contains the vital information pertaining to vowel identification, and has also been shown to contain the most important cues for transient size discrimination from the psychophysics experiments carried out in this study.

Size normalisation from tAIM was achieved by performing a Mellin transform. The resultant Mellin images were the magnitudes of the ratios between the bands of energy present in the SSIs. Size discrimination from tAIM was proposed due to the cues discovered in this study. The most important cues found were the differences between the prominent resonances, and related to that, the SCF of a signal's spectrum. A method of automatic size discrimination using SCF value comparisons was considered. However, SCF has previously been shown to be independent of pitch for periodic signals (Plomp, 1970), and due to the discovery that the SCFs of the simulated cylinder signals used in chapter 7 were contrary to the expected relationship between SCF and the sizes of objects, it is possible that the SCF of transient signals is also independent of transient pitch and therefore an unreliable cue upon which to base a model. Therefore, the relative positions of two signals'

most prominent resonance are used as a tool for size discrimination from Mellin images from tAIM.

In chapter 7, the time-zero of the SSI was shown to be comparable, although with differences, to PSD estimates of the sphere signals. The Mellin transform retains the position of T-0, and creates an image containing magnitudes relevant to the positions and strengths of the resonances within the signal. Therefore, size discrimination from tAIM can be achieved by analysis of the Mellin phase of T-0. Since T-0 is also the point of highest energy in the auditory image, it therefore the most useful information regarding size and shape of the analysed object. The transform has effectively performed an FFT on that particular column of the auditory image. The resultant phase of Mellin T-0, henceforth referred to as Mellin phase, contains the size information of the signal. The Mellin phase corresponds to where on the SSI T-0 the resonances occur, and therefore the position of the resonances in the spectrum. The relationship between the Mellin phases of two signals hence shows the differences between the resonances; this is the spectral cue that contains size.

A crude example of this is shown below in figure 77 with the comparison of two double-damped sinusoids. The 'larger' signal contains resonances at 1200 and 1800 Hz, and the smaller at 1400 and 2100 Hz. The top panel shows T-0 from the SSI, and the two resonances for each of the signals are seen as peaks in the ERB spectrum. The centre panel is the T-0 from the Mellin image and shows the size normalisation that has occurred due to the ratios between the resonances being equal; F2/F1 = 1.5. Finally, the bottom panel is the Mellin phase. The average rate of change of the Mellin phase (dy/df) is calculated for each signal and used in comparison with each other in order to discriminate for size. The value nearest to zero corresponds to the 'big' signal, representing the fact that the resonances in the SSI occur closer to 100 Hz, and thus are lower in frequency. For size 1 the Mellin phase value is -0.6, and for size 2 it is -0.64, indicating that size 1 contains resonances from a larger object.

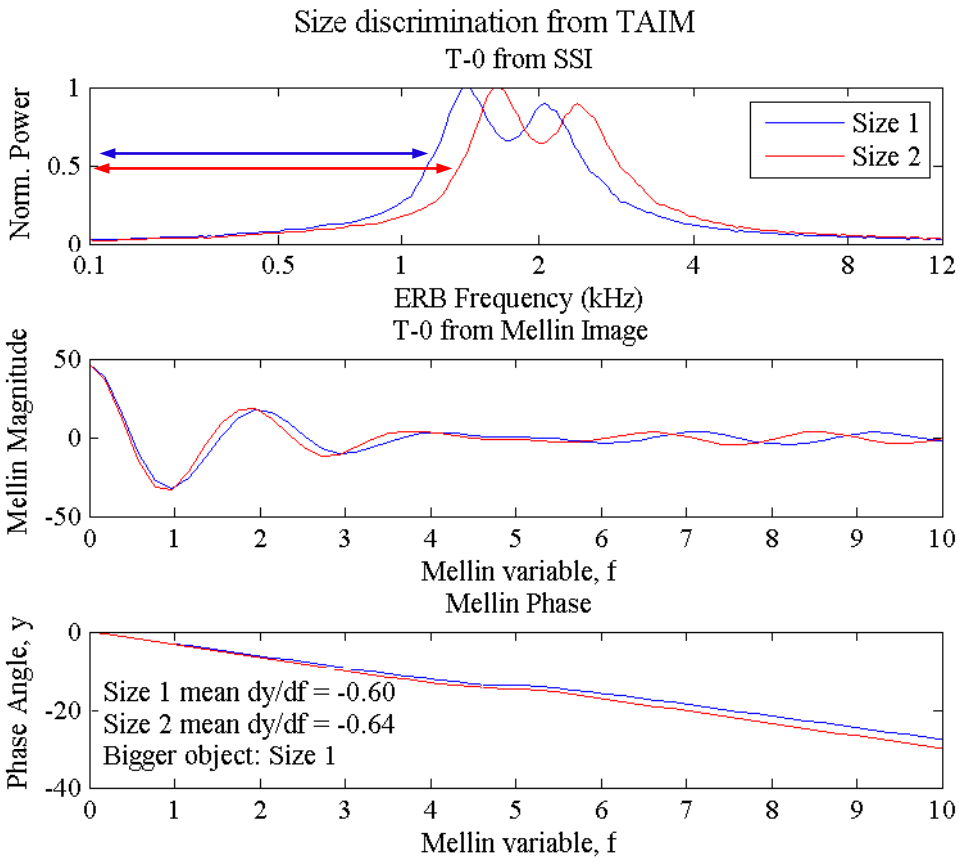

Figure 77: Size discrimination from tAIM using two double-damped sinusoids to simulate two objects of different size. The Mellin phase is extracted from the Mellin image at T-0, and the average rate of change (*dy/df*) for each is compared to conclude which object is bigger. This corresponds to where the resonances are positioned along the ERB axis after the output of the dcGC filterbank has been aligned – the blue and red arrows in the top panel.

8.1 Size discrimination of simulated and real stimuli

Now that there is a method of size discrimination from tAIM using the Mellin transform, the rest of the signals can be analysed. This section shows the results of size discrimination from tAIM of simulated vowels (from chapter 7), polystyrene sphere recordings from the experiments and recordings of other polystyrene shapes, moulded FIMO clay balls, and finally the simulated cylinders signals also from chapter 7. The new stimuli are discussed in due course.

Table 16 shows the vowels used and the corresponding formants as spoken by a man, woman and child. Figure 78 shows the successful size discrimination of each of these vowels for each of the speakers, where the Mellin phase *dy/df* is closer to zero for the bigger object.

Properties of simulated vowels from different speaker sizes

Word	IPA	Formants	Men	Women	Children
Bead	/i/	F1	270	310	370
		F2	2290	2790	3200
Bud	/ʌ/	F1	640	760	850
		F2	1190	1400	1590
Bawd	/ɔ/	F1	570	590	680
		F2	840	920	1060

Table 16: Three words and the corresponding IPA vowels are shown above, and the averages of the F1 and F2 formants for each vowel sound for groups of men, women and children. These values are taken from a sample of the results obtained in a study by Peterson and Barney (1954) as displayed in Borden et al (2002, 4th ed., p92).

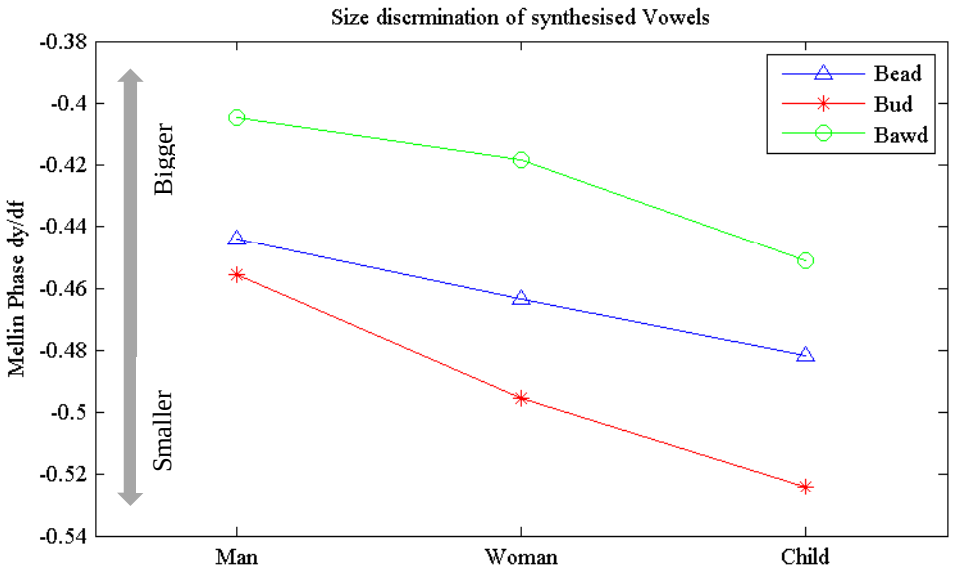

Figure 78: Size discrimination of simulated vowels as spoken by a man, woman and child. The higher the Mellin phase *dy/df* value in the graph, i.e. the closer the value to zero, the bigger the sound source. This image shows the correct size discrimination of the speakers.

The polystyrene spheres used for the psychophysics experiment in chapter 4 were processed by tAIM and the size discrimination results from the model are shown below in figure 79. A reminder of the dimensions and masses of the spheres are shown in table 17. The results of tAIM's size discrimination is correct, although the difference between the Mellin phase *dy/df* values of X.Large and Large, and also between Medium and Small are very small. The model has correctly discriminated between them, but the reasons for the almost equal *dy/df* values will be discussed later in limitations section.

Properties of polystyrene spheres

	X.Large	Large	Medium	Small	X.Small
Diameter	120 mm	100 mm	90 mm	80 mm	70 mm
Mass	17.95 g	11.36 g	8.22 g	5.61 g	3.40 g

Table 17: The dimensions and the masses of the real recorded polystyrene spheres used in the experiment in chapter 4.

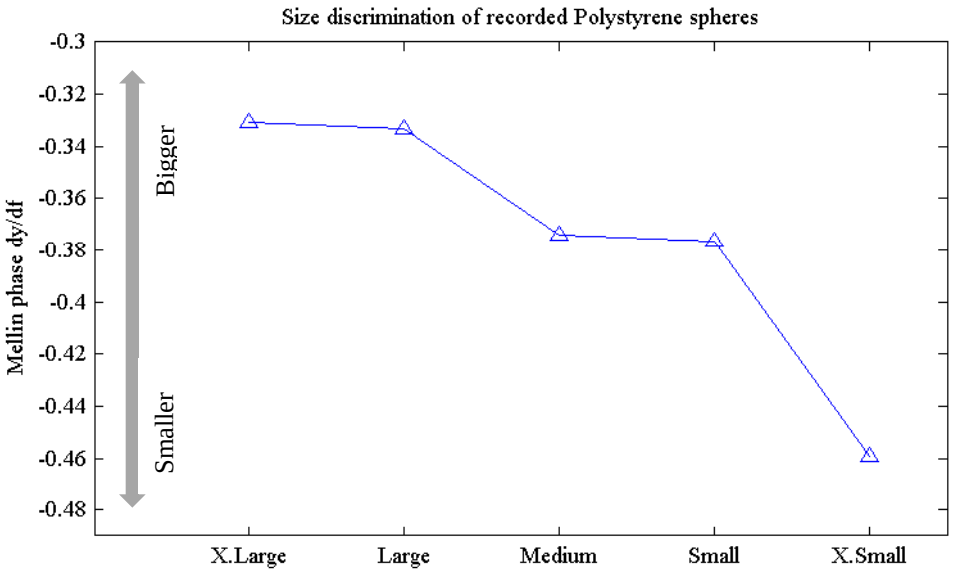

Figure 79: Size discrimination of real recorded Polystyrene spheres using tAIM. The figure shows correct discrimination of the spheres, however the difference in *dy/df* values for X.Large and Large and also between the Medium and Small spheres are very small.

Figure 79 shows the tAIM size discrimination results from the full sphere signals, and it is seen to be successful with a *dy/df* value moving further from zero with decreasing size. The Mellin phase here is related to the differences between the starting points of the wide-band energy of F0, which is not a resonance in the typical sense but more the result of the impact of the striking ball; it is the difference between the resonances such as F1 in signals that was shown to be the most important cue. However, the shape of the curve of figure 79 is very similar to the SCF values of the signals, on a reversed frequency axis. Figure 80 shows the SCF values on such an axis, for comparison. For the spheres, SCF values are a good indication of size, and could be used as a cue. It is a coincidence, therefore, that the Mellin phase values for size discrimination show similarities to the SCF values, as evidence for other signals suggest otherwise.

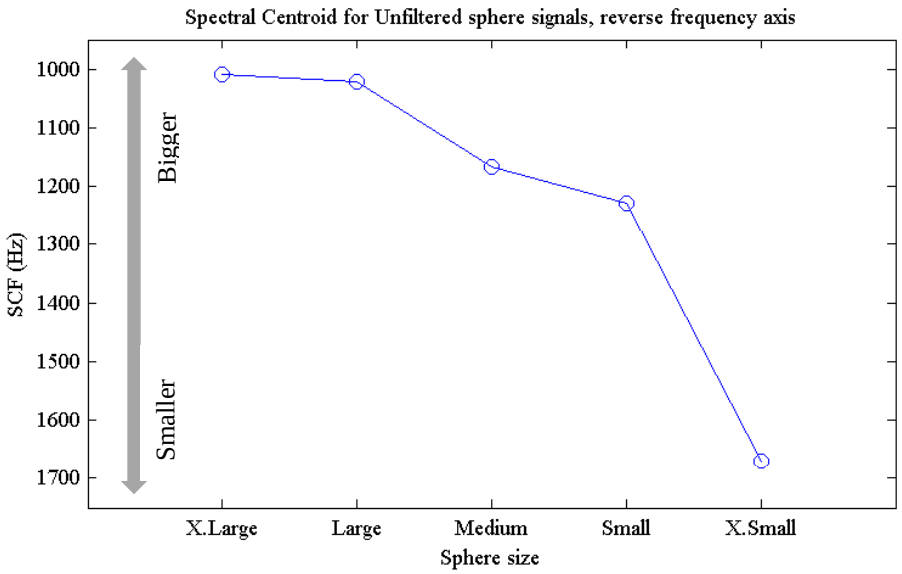

Figure 80: SCFs for unfiltered sphere signals plotted on a reverse frequency axis, in order to show similarities in shape with the Mellin phase values in figure 77.

The most important cue for size discrimination as found through psychophysics experiment was the difference between the resonances. Figure 81 below shows the tAIM size discrimination results for HPF sphere recordings, filtered with a cut-off

5 % below F1; this filters out only the wide-band F0 and retains the important F1s. The figure shows an increase in the differences between the *dy/df* values for different object sizes, which is interpreted as an improvement in the discrimination of size from tAIM.

Figure 82 shows this in terms of the correlation between the Mellin phase *dy/df* value and the absolute size of the spheres. The top panel shows Pearson's correlation coefficient, r, to be 0.86, but after the signals are filtered, this improves to r = 0.98. The small difference between the Small (80mm) and X.Small (70mm) sphere *dy/df* values, and consequently the deviation of their values from the line of best fit below in figure 82, could be due to the similar positions along the SSI at T-0 that the energy for the F1s of these signals begins to appear (see figure 76). The reason for the difference in shape of the SSI T-0 for the Small and X.Small sphere signals is possibly due to the recording method, or perhaps an unknown issue with the construction of the spheres. Nevertheless, the differences between the values of F1 was used as a cue for the participants of experiment 2, and consequently tAIM discriminated for size between the spheres by using the differences between F1.

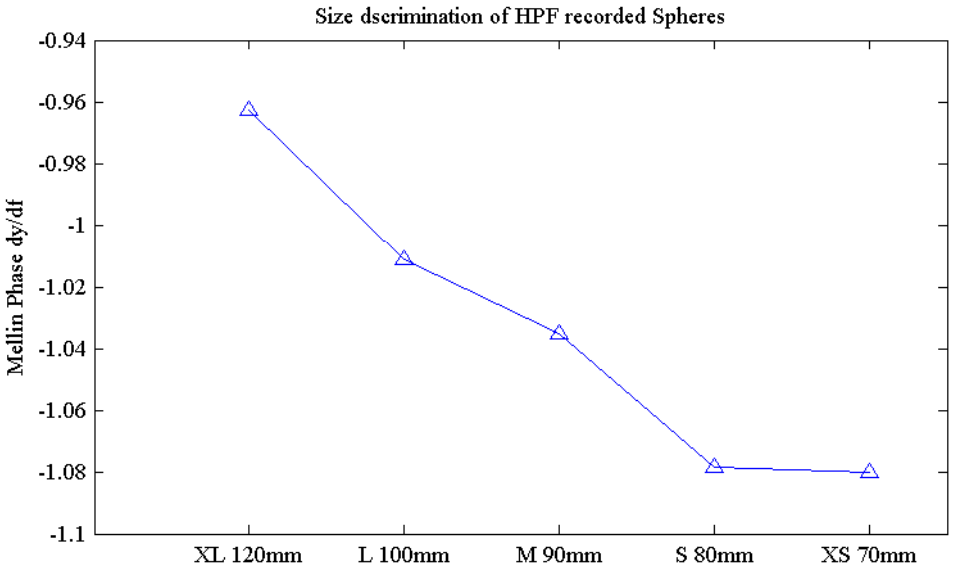

Figure 81: HPF with F1 polystyrene sphere signals, analysed by tAIM for size discrimination. Compared to the size discrimination of both the full signals and the LPF signals, the differences between *dy/df* values are greater apart from between the smallest spheres.

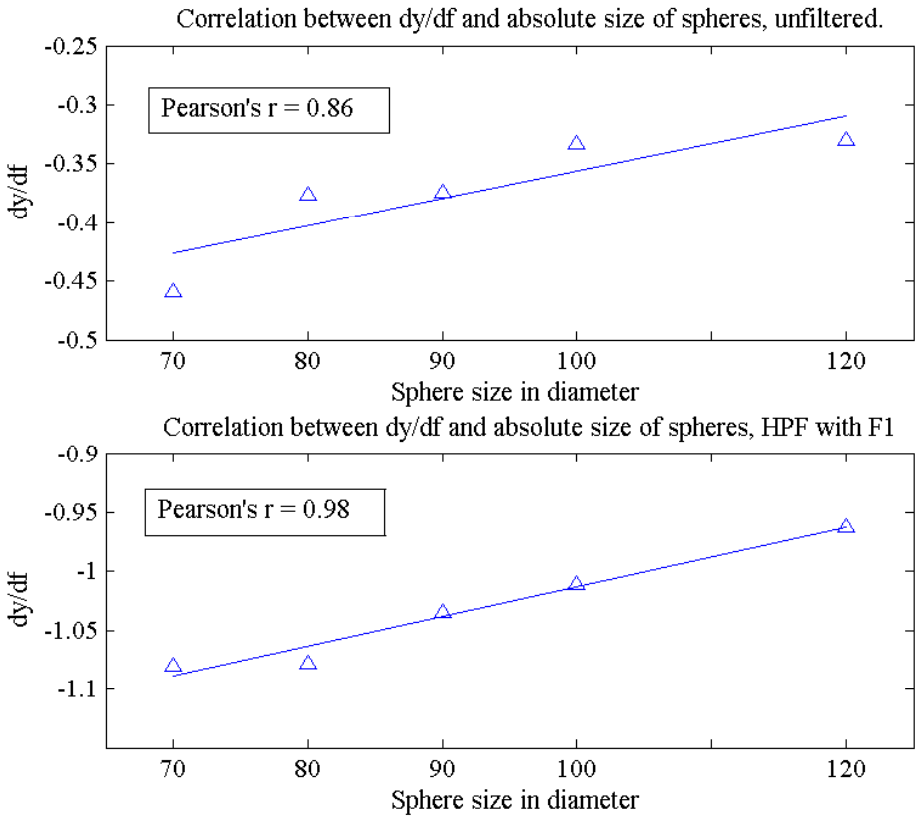

Figure 82: Correlation between Mellin phase values of, in the top panel, unfiltered signals, and in the lower panel, HPF with F1 signals with the absolute size of the spheres. Pearson's correlation coefficient, r, increases from r = 0.86 to 0.98, indicating an increase in strength of the correlation between *dy/df* and size after F0 has been filtered out.

Recordings of other polystyrene shapes were provided by a size/shape experiment carried out by Clarke (2009), and these were processed by tAIM next. Three sizes of cone, egg, and heart shaped objects and two of cube shaped objects were recorded in a similar manner as described in section 4.1.1. Prior to analysis by tAIM, the signals were band-pass filtered with cut-offs at 100 Hz and 16 kHz in order to minimise noise, and normalised to RMS = 1. Averages of three recordings for each size were taken to be the signal for that particular size. Averages of recordings of spheres constructed from FIMO polymer were also analysed. The FIMO sphere sound source was acquired by using a Newton's cradle style of apparatus, colliding two spheres of the same mass which were hanging from a

wooden frame at 20 cm apart. Both spheres were pulled to one metre from their rest position in opposite directions, and were then allowed to fall freely. The resultant signal was the collision between the two spheres.

Properties of recorded polystyrene shapes and FIMO spheres

Size/Shape	***Properties***	***Cone***	***Cube***	***Egg***	***Heart***	***FIMO sphere***
Large	Weight (g)	75.4	3.2	7.7	8.4	550
	Max Dimension (cm)	31.3	6.0	10	11.2	11
Medium	Weight (g)	26.3	1.9	5.2	4.6	520
	Max Dimension (cm)	26.1	4.0	8.4	8.3	10
Small	Weight (g)	13.5	n/a	2.3	2.0	515
	Max Dimension (cm)	20.4	n/a	5.6	5.5	9.5

Table 18: Real recorded stimuli to further test the size discrimination abilities of tAIM The cones, cubes, eggs and hearts were the same material as the polystyrene spheres described in chapter 4, and the mass of each shape and size are displayed in this table. There are three sizes of each shape except the cube. Thanks to Dr. Stefan Bleeck for providing the details of the polystyrene objects.

Table 18 shows the shapes and masses of the objects recorded. The cones had a much larger mass than the other polystyrene shapes. The FIMO spheres were the heaviest above all due to their dense polymer composition. The number of recordings made of the objects below was not conducive to creating clear spectra of the signals, shown in figure 83. Resonances, for example F1 and F2 which were visible from the spectra of the polystyrene spheres, are not as evident in the spectra of the objects below, and similarities between spectra of different sizes are not evident. This could be a result of the recording method, or the low number of signals used to create the averages; further recordings may solve the problem. The SSI T-0 extractions are shown in figure 84, and though the shape for each size is still clearly different, there is however a clear difference in the position of the spectra in most cases, and so for this reason the unfiltered signals were used for discrimination by the model.

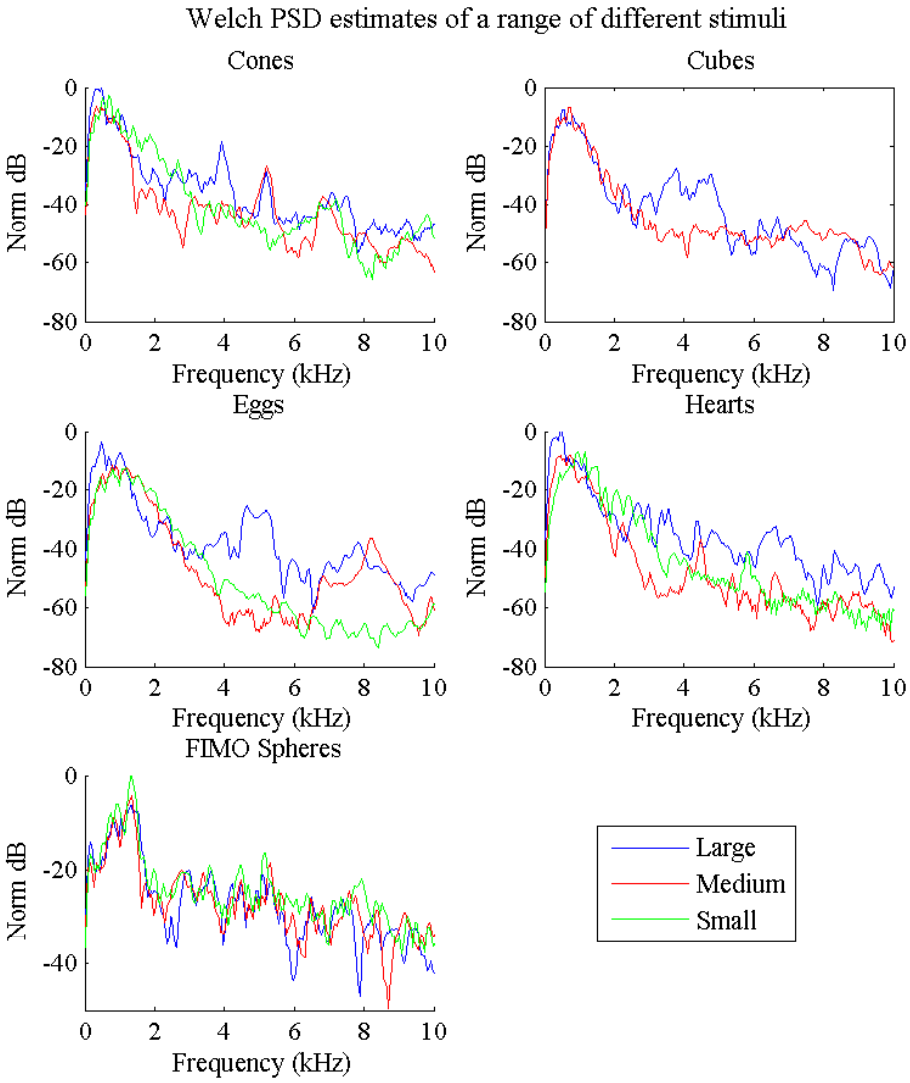

Figure 83: Welch PSD estimates for the polystyrene cones, cubes, eggs and hearts, and the polymer FIMO spheres. The spectra do not show resonances for these polystyrene objects that are as clear as the polystyrene spheres, possibly due to the recording method, or the low number of signals out of which to create an average signal.

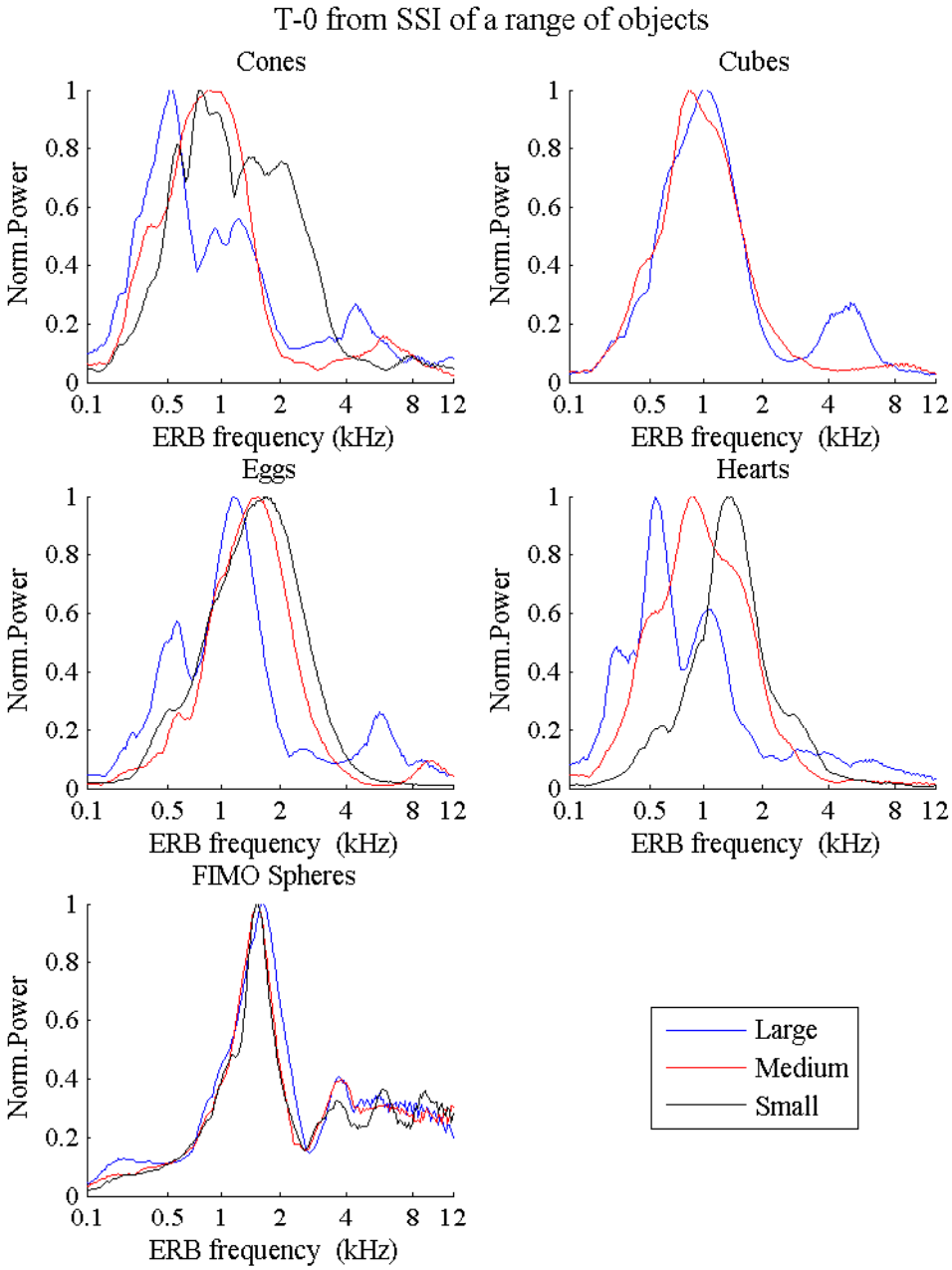

Figure 84: The T-0 section of the SSIs for a range of polystyrene shapes and FIMO spheres. This image shows that only the Large and Medium eggs, the hearts and the sphere signals show consistencies in their spectral envelope shapes and thus explaining why tAIM was capable of discriminating between them.

Figure 85 contains the size discrimination results obtained from tAIM, and with the exception of the two cube signals and the medium and large cones, the model correctly discriminated for size. These exceptions are due to limitations in the model, which will be discussed in section 8.2. Figure 86 shows the correlation between absolute size of object and the Mellin phase *dy/df* value extracted by tAIM. The Pearson's correlation coefficients, r, show strong positive correlations between size and *dy/df*: Cones, r = 0.85, Eggs, r = 0.81, Hearts, r = 0.95, FIMO Spheres, r = 0.98. The value for the cube signals was not calculated due to its negative size discrimination result.

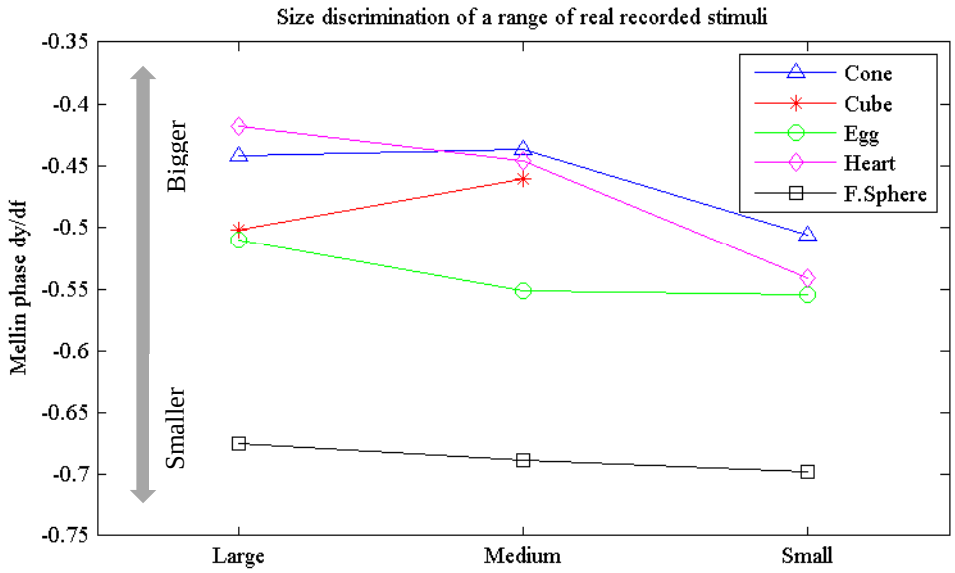

Figure 85: Size discrimination of the shapes described in the table above. For the most part, size discrimination of the objects was successful, except for the cubes, and between the large and medium cones. For all other sizes, the Mellin phase *dy/df* value correctly discriminates for the size of the object.

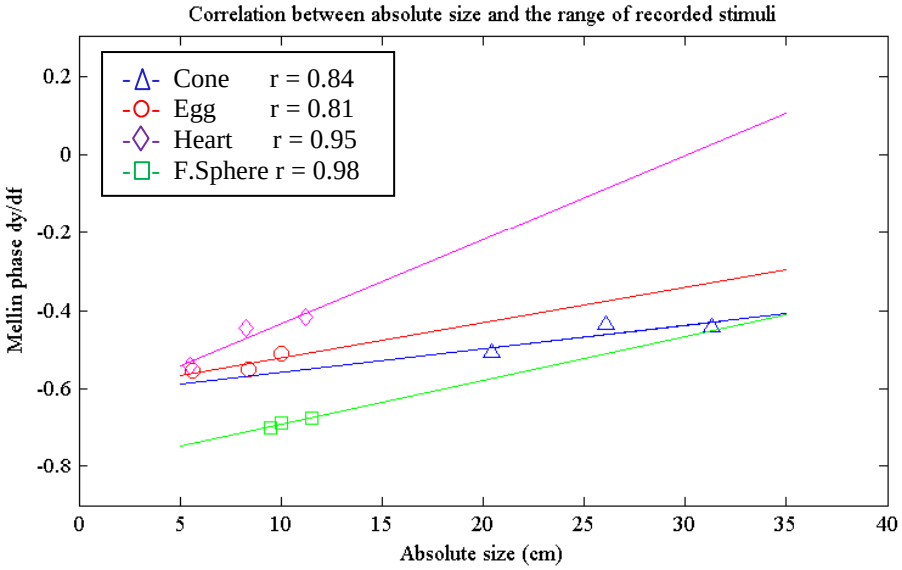

Figure 86: Correlation between absolute size of the recorded objects and the Mellin phase *dy/df*. The Pearson's correlation coefficients, r, are shown to be close to 1, indicating stong correlation between the *dy/df* extraction and absolute size of the sound source, especially for the polystyrene hearts and Fimo spheres. The cubes were removed from this plot due to the negative result for size discrimination.

The final set of signals to be analysed were the simulated scattered signals from underwater cylinders. Four sizes from four different types of simulated cylinder signals were processed by tAIM; scattered signals from Air-filled cylinders with b/a = 0.99, Water-filled with b/a = 0.95, Water-filled with b/a = 0.99 and Water-filled cylinders with a constant wall thickness of 2 cm. The cylinder signals with diameters of 2 m, 1.8 m, 1.6 m, and 1.4 m were processed and the resulting Mellin phases *dy/df* values are plotted below in figure 87. In this case, tAIM was unable to correctly discriminate for size for any of the signals processed. In order to establish the cause of this considerable failure in size discrimination, as well as the incorrect result for the cubes and two of the cone recordings above, the stages of the size discrimination process within tAIM will be analysed in the next section.

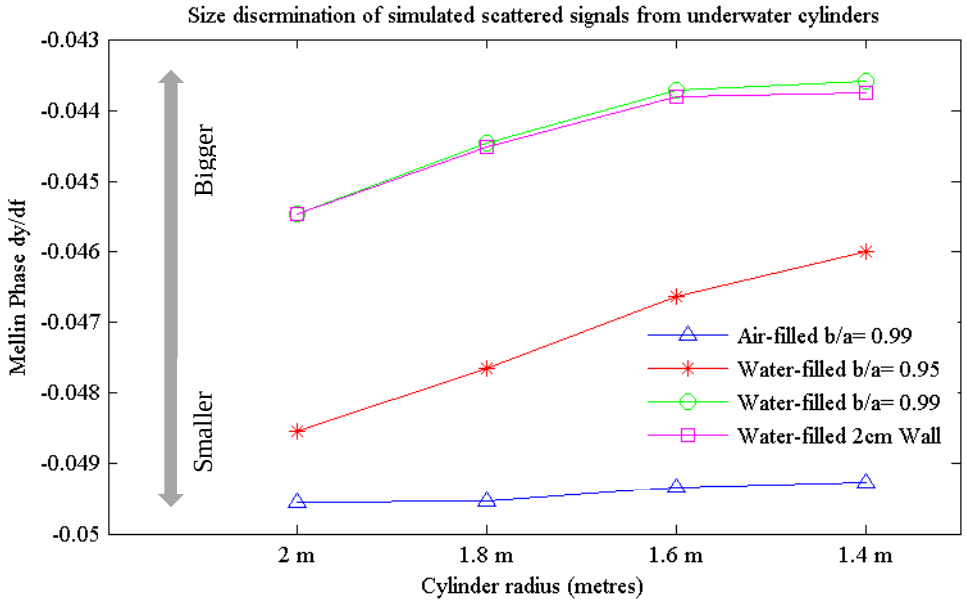

Figure 87: Size discrimination results from four different sizes of the four different types of underwater cylinders. These simulated signals were processed by tAIM and the result in all cases is incorrect size discrimination, as can be seen by the cylinders with the smaller diameters showing *dy/df* values closer to zero.

8.2 Limitations of tAIM

It was already shown that an improvement in the size discrimination performance of tAIM was found after HPF the sphere signals in order to highlight the important F1 resonances of the signals. The limitations of the Mellin transform lie in the assumption that size is dependent solely on the position of the spectral envelope, and that the envelope shifts along the spectrum according to the size of the sound source. It was shown earlier in section 4.2.2 how F1 and F2 shift upwards in the frequency spectrum as the size of the polystyrene spheres increases. Also, in section 7.4, the PSDs of the simulated scattered signals from underwater cylinders show a shift upwards with decreasing cylinder diameter. The literature has shown that vowel resonances retain the same ratio despite the length of the vocal tract from where they emanate (Smith et al., 2005). tAIM was capable of creating size-normalised images for the simulated vowels, scattered cylinder signals and recorded polystyrene spheres. However, size discrimination was not achieved for all

the signals, and so the individual processes of the tAIM size discrimination will now be analysed in order to find out why.

The double-damped sinusoids and the simulated vowels were formed in such a way that the relative amplitude between the resonances was similar. The resulting SAIs and MIs in chapter 7 showed clear resonances and spatial frequency patterns. The recorded spheres were also normalised for size, due to the almost consistent ratios between F1 and F2 (table 4a, section 4.2.3). The simulated cylinders, however, had resonances that were closer in frequency, and in the case of the air-filled cylinders, there were almost no resonant peaks in the spectrum (see figure 73), the information was contained within a range of ~1300 Hz. The resultant MIs from the cylinders were a single diagonal band of energy, compared with the separate bands from the vowels which merged together to bands which appeared to interfere with each other. This suggests that the Mellin transform only works well when there are resonances which are clearly distinct.

Figure 88 contains the T-0 extraction from the SSIs for the cylinders. The frequency scale is ERB, and this represents the spectral analysis carried out by the dcGC filterbank. With the exception of the air-filled cylinders, the signals seem to have been split into two distinct resonances. The air-filled cylinder signals appear to have no difference in the position on the ERB scale, which would account for the Mellin phase dy/df being incorrect in discriminating for size. The first of the water-filled cylinder signals, b/a = 0.95, shows narrow F0 bands which are low in magnitude for larger signals. F1 for all sizes appear to have the same frequency value, with larger magnitudes for large sizes. This would explain the high SCF for the larger signals, but does not allow for size discrimination using tAIM as the processing requires differences between the F1s in order to discriminate. The lower two panels show very similar basilar membrane responses for the cylinder signals, where F0, also narrow but high in magnitude and equal for all sizes, shifts upwards slightly with each decreasing size of cylinder and F1 occurs at the same position for each size, but with bigger magnitudes for bigger cylinders.

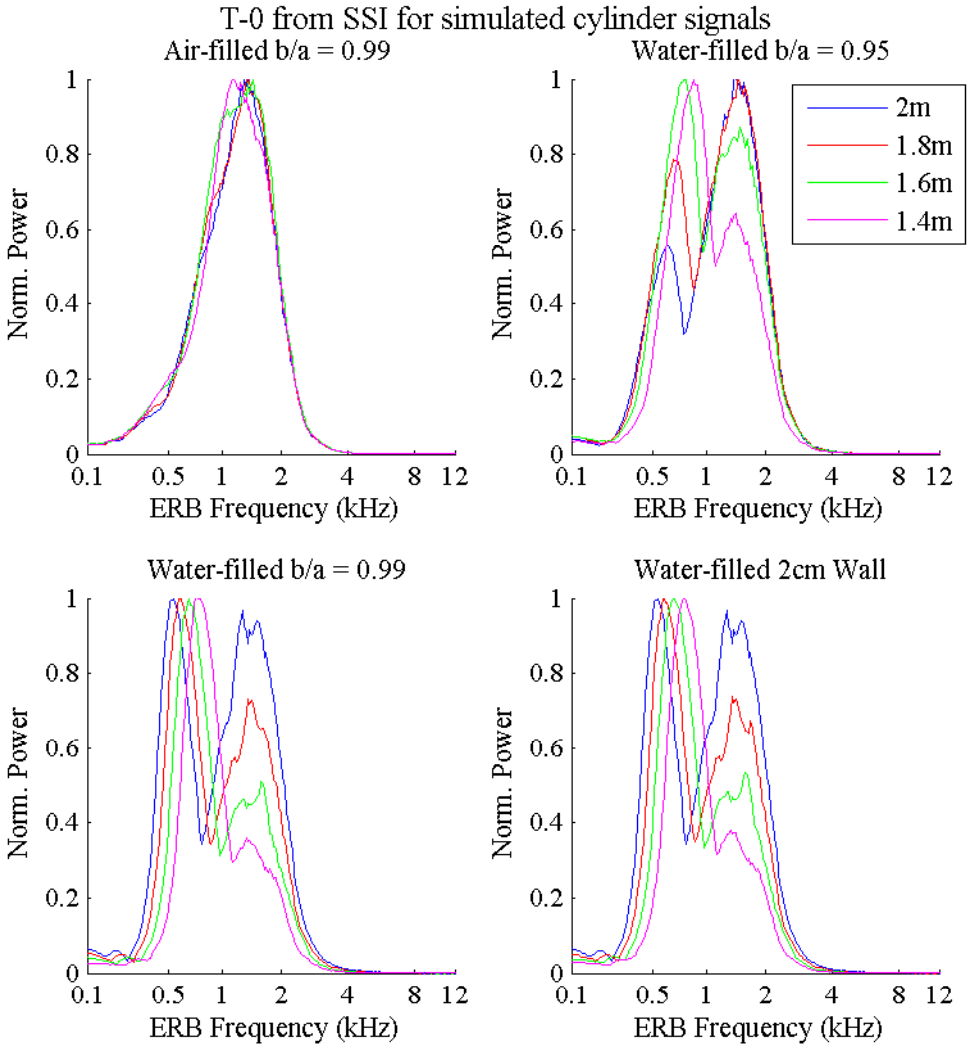

Figure 88: Each panel shows the T-0 section of the SSI for four diameter sizes of each cylinder type. This shows the spectral analysis carried out on the signals by the dcGC. The air-filled cylinders show very little difference between the sizes. The water-filled 0.95 cylinders show no difference in the relative positions of the resonances, though there is a magnitude difference between the individual F0 resonances. The two lower panels have almost identical SSI spectra, with F0 showing values proportional to the diameter size, but F1 showing the same value but with different magnitudes for difference sizes.

The size discrimination method used in tAIM calculates the distance from the lowest ERB value to where the first resonance appears by measuring the phase of the Mellin transform. This method should have worked on the signals shown in the lower two panels of figure 89 below. It might be suggested that the distance

between the resonances is not enough for a phase difference to be detected. Increasing the resolution by increasing the number of filters in the dcGC filterbank was not enough to cause improvement. Secondly, for the Mellin transform to normalise for size, the ratios between the resonances must be the same or have a very small degree of variation. Since the first resonances shift according to size, yet only the magnitude of the second resonances differ, the ratios are different, and thus the Mellin normalisation was not successful, thereby rendering the size discrimination of the signals impossible.

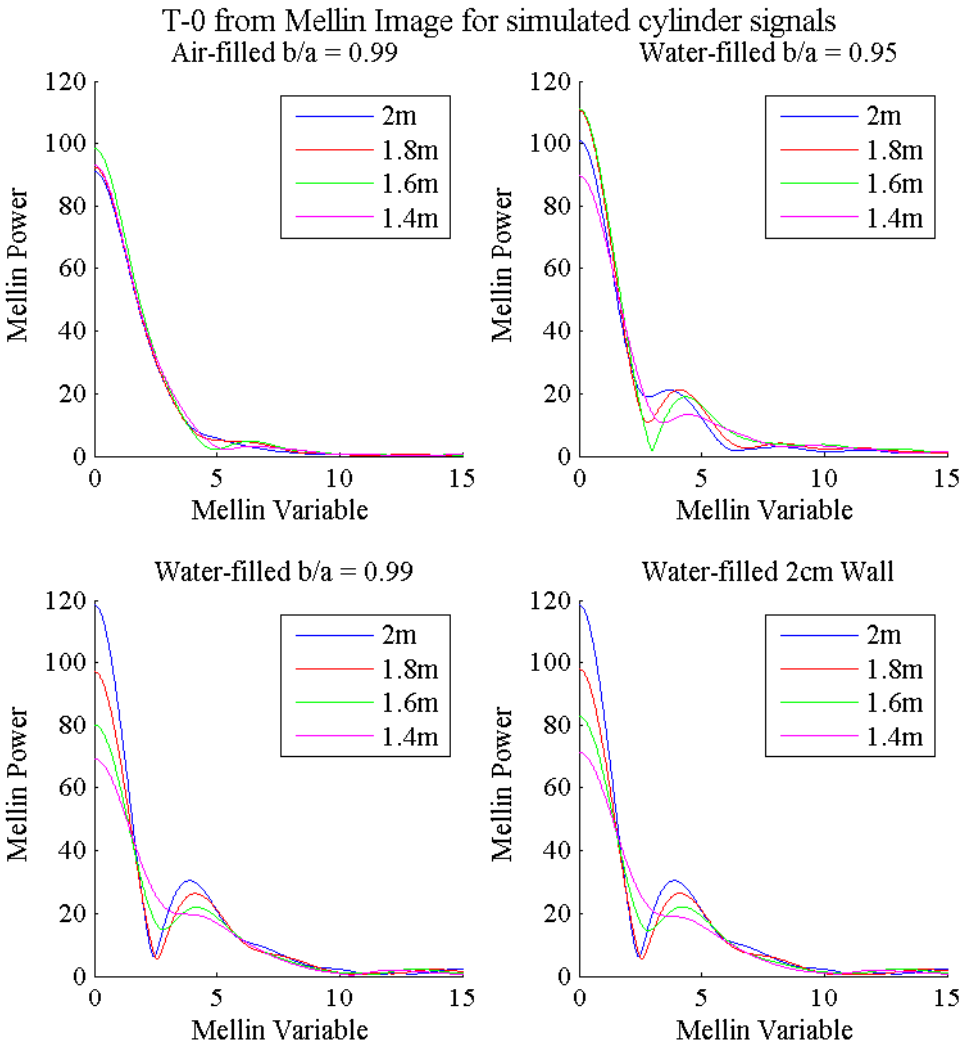

Figure 89: T-0 extraction from the Mellin images for four sizes of each cylinder type. This image shows how, with the exception of the air-filled cylinder signals, the size normalisation of the cylinders was not as successful as originally

perceived in chapter 6. The peaks in the spatial frequencies are similar but they become shallower with less distinction for smaller cylinders.

From the psychophysics experiment in chapter 4, participants performed above chance for size discrimination when comparing the signals that had been low-pass filtered with a cut-off frequency of below F1. It has been shown that the relative positions of the resonances influence size discrimination abilities (section 5.3.2), and so if this theory is applied along with the abilities when comparing LPF signals, the results of tAIM for the simulated cylinder signals should improve. Figure 90 corresponds to this theory, with tAIM displaying successful size discrimination of the LPF cylinders, except for the air-filled. This was expected, due to the shape of the T-0 SSI, and how the model compared signals for size discrimination. The correlation between absolute size and *dy/df* values are then plotted in figure 91 for all cylinders except the air-filled type. The Pearson's correlation coefficients show strong positive correlation between size of the water-filled cylinders and *dy/df* values: b/a 0.95, r = 0.993, b/a 0.99, r = 0.994, and 2cm wall, r = 0.996.

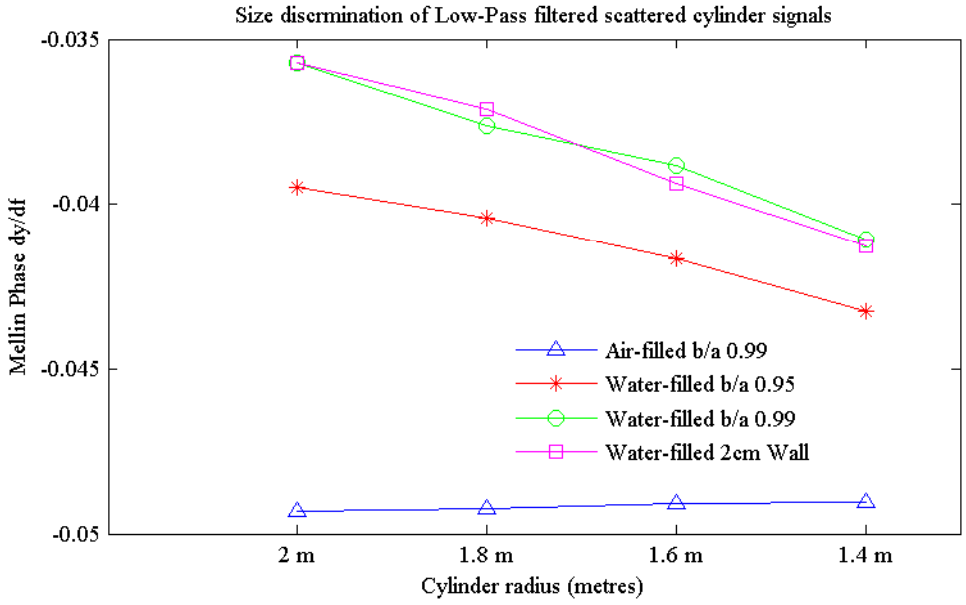

Figure 90: The size discrimination of the simulated underwater scattered signals which have had a low-pass filter applied to the signals, with cut-off frequencies at 5% below F1. For all cylinders expect air-filled, tAIM has successfully discriminated for size.

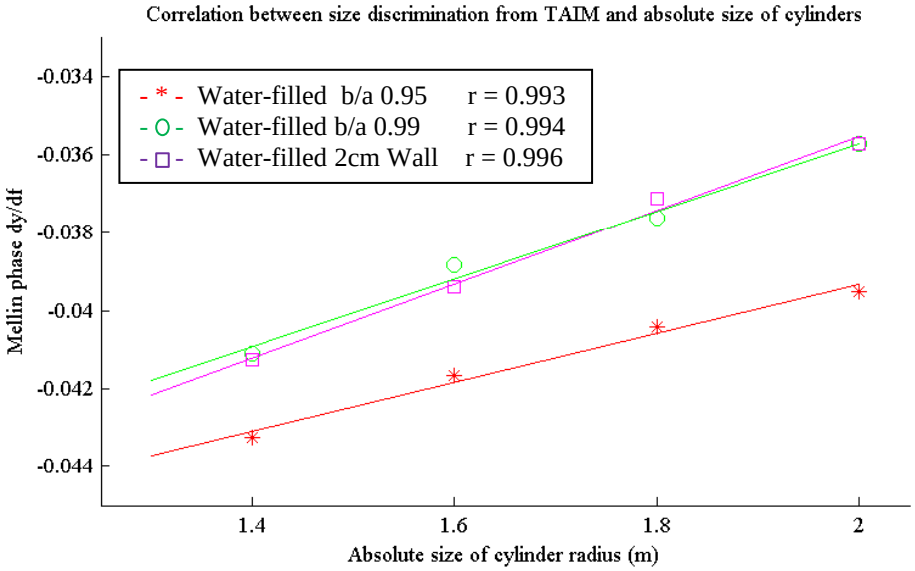

Figure 91: Correlation and Pearson's correlation coefficients for the relationship between absolute size of the cylinders and the extracted *dy/df* value of size discrimination from tAIM. Air-filled cylinders were not included due to the incorrect size discrimination result. All other values show a strong positive correlation with size.

A variety of recordings of other polystyrene shapes and one set of spheres constructed from the polymer FIMO were also processed by tAIM for size discrimination. The results showed that for the egg-shaped and heart-shaped objects and the FIMO spheres, size discrimination was achieved by tAIM. However, the large and medium cones and cubes were not correctly discriminated for size. The T-0 SSI values will show the frequency analysis carried out on the shapes by the dcGC filterbank, and whether or not the spectral envelope for the objects were similar. Figure 84 showed that not all the shapes had envelopes that are consistent with changing size. The Mellin T-0 extractions shown in figure 92 correspond with this theory, showing that only the FIMO spheres were successfully normalised for size, and all the other objects show Mellin T-0 extractions appear to be from different objects, or of objects that have not been scaled successfully. This suggests, again, the possibility of transient signals that cannot be scaled linearly, although better recordings would need to be acquired of these objects in order to better support this theory.

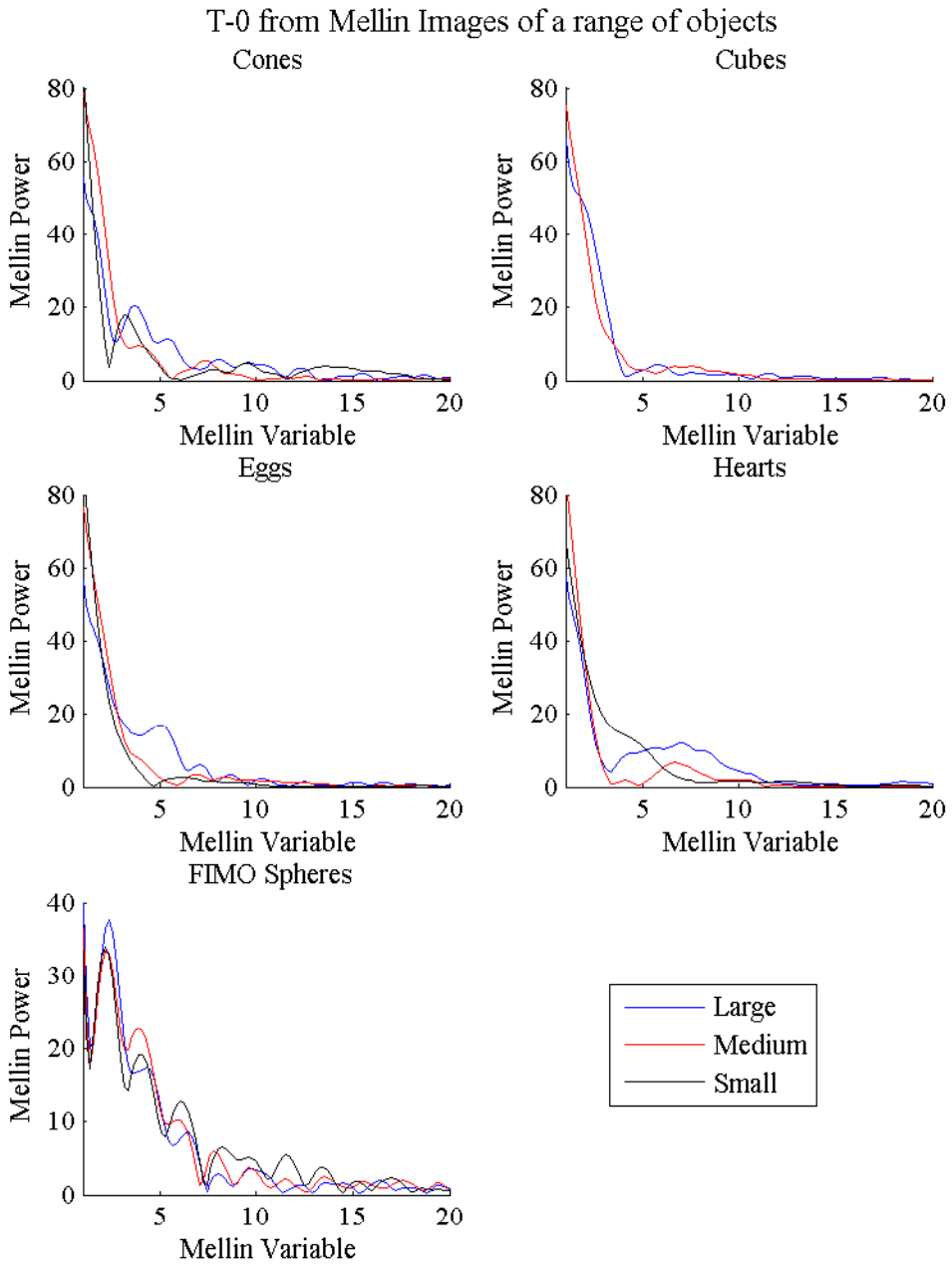

Figure 92: T-0 Mellin Image section for the different polystyrene shapes and FIMO spheres. Only the Cubes and the FIMO spheres show consistency in their spatial frequencies; all other shapes have not been normalised for size.

8.3 Conclusions of the modelling of size discrimination

The conclusions to be drawn from this are two-fold. It appears that size normalisation is not possible for all types of signals, although the recording methods of the signals were not well documented and there may have been a degree of inconsistency within the signals. The spheres, both polystyrene and FIMO, showed clear SSI spectral envelopes that shifted with size, and tAIM successfully discriminated for size. Applying a high-pass filter to the signal in order to expose the proven important F1s improved size discrimination by tAIM. This proved that size discrimination using tAIM is reliable only when there are clear distinctions between resonances. This is further indicated by the cylinder signals. The dcGC spectral envelope showed small differences between F0, but F1 values were the same but with different magnitudes corresponding to size.

It seems that there is a limit to the 'just noticeable difference' of size using tAIM, which is especially obvious after analysing the spectra of the simulated cylinder signals. The spectra are seen to be slightly lower for larger sizes, but without low-pass filtering to remove the similar upper resonances that appears in the dcGC filterbank, tAIM did not detect this. It is possible that the resolution of the dcGC filterbank is not as fine as expected, but it is also possible that the Mellin transform is not very effective on transient signals, especially since it has been shown that despite linear relationships between size and resonant frequency values, linear scaling was not successful. In any case, tAIM showed success in size discrimination for simulated vowels from men, women and children, recorded polystyrene spheres, LPF simulated cylinders signals and a range of other recorded stimuli.

9. Conclusions

9.1 Summary and Considerations

The purpose of this study was to:

1 - Investigate the importance of spectral cues in human auditory size discrimination of struck objects.

2 – Investigate how an auditory model using a dynamic compressive gammachirp filterbank might be adapted perform size discrimination on the transient signals of stuck objects.

Other studies been carried out on the extent of human and mammal size discrimination abilities, from the range of the abilities to how the presentation method affected discrimination abilities, and all had suggested cues that were used in these tasks but none had specifically targeted spectral cues. None had attempted to manipulate signals in order to test the importance of the resonances in transient signals which have been shown to be the information carrier in vowel sounds.

In this study a number of psychophysics experiments were carried out with 40 participants using signals recorded from polystyrene spheres of different sizes that had been struck with a small metal ball-bearing. The signals were processed in order to remove, as much as possible, any temporal or intensity cues that could provide size information in order to isolate the spectral cues. They were then presented to participants in three ways: unfiltered in order to determine the basic size discrimination abilities; scaled by resampling the signals proportional to the ratio between F1s in order to determine whether it was possible to make a signal appear as though it was a different size; and finally filtered by high-pass, low-bass or band-stop filters to remove different sections of the spectrum and test the importance of these areas. The results proved the importance of spectral differences, with the

signals with a greater difference between SCF showing fewer errors in size discrimination; participants were successfully led to believe a signal was a different size through the manipulation of its spectral envelope; and most interestingly, the presence of F1 reduced the number of errors for the band-stop signals and the low-pass filtered signals. The high-pass filtered signals resulted in better results when F1 was filtered out due to the larger differences between the sphere's F2s. Following the discovery that the SCF of the simulated cylinders used in chapter 7 increased with increasing size instead of the opposite, it was finally deduced that the most important cue in the size discrimination of transient signals is the difference between the comparison signals' most prominent resonance.

The importance of the spectral cues in the size discrimination of the spheres, and the result that the sizes could be altered by shifting the spectral envelope which was similar to the STRAIGHT study on vowels (Kawahara et al, 1999), motivated the design of an auditory model to create auditory images of transient signals. tAIM was built based on the well-known AIM (Patterson et al, 1995), with few processes and achieved the same result: auditory images that were capable of being normalised to create Mellin images of the same objects of different sizes. Its success is highly dependent on the signals presented for analysis, they must have spectral envelopes that are very similar, with obvious resonances that shift in accordance to the size of the object. tAIM was then extended to carry out size discrimination of signals, a task that was successful with objects whose spectral envelopes had these shifts. In some cases, filtering the signals improved the output of the tAIM size discrimination.

A number of points were not considered in this study. A short study of the discrimination of simulated vowel sounds used to test tAIM should have been carried out, as well as the simulated cylinder signals, to back up the experience of the author with these signals. While the literature has plenty of proof of the ability of listeners to hear speaker size, this should have been carried out to strengthen the validity of these signals as a method of testing tAIM. The simulated cylinder signals present an interesting problem in this thesis, showing spectral centroid frequencies that decrease with decreasing size, contrary to other signals. Although the author

was satisfied in how different the signals sounded, a short test with a few participants would have strengthened that argument. Future work on the importance of SCF is needed as the signal's apparent sizes do not correspond to the calculated SCF.

The importance of F1 in scaling was investigated due to the strength of this resonance in the spectra of the sphere signals. When scaling signals to sound larger in order to match the size of a comparison signal, this resulted in the scaled signal being chosen as the bigger size, or conversely with signals scaled to be smaller, the unscaled signal was chosen as the larger signal. The scaling method, while the spectra showed a degree of success, was not perceptually successful, resulting in SCF values too low for the signals scaled to sound larger, and SCF values too high for signals scaled to sound smaller. The results for the filtered signals showed that removing F1 using a band-stop filter showed a significant improvement in discrimination abilities, and this was attributed to the differences between the F2s of neighbouring signals. Perhaps a test of using the ratios between F2s might have shown better results or been more successful as a scaling method. Or, even though the scaling method used was inspired by STRAIGHT, perhaps a method a lot more similar to it could be used by interpreting the wide band of F0 energy as similar to glottal pulse energy, then the spectral envelope to be stretched and compressed would consist of all the other resonant peaks in the spectrum.

If time had allowed, recordings of different shapes would have been made by the author. While thanks is given to the students who provided the recordings used here to test tAIM, they were limited in number and inconsistent in spectra possibly due to a noisy background. The extra recordings could have been averaged in the same way as the sphere recordings and could have shown more positive results when processed by tAIM. They could also have been used to carry out a number of smaller size discrimination experiments on the different shapes, and possibly expand this study into an investigation of shape as well as size discrimination of transient signals.

Finally, the adaptation of the Auditory Image Model here was carried out for the consideration of transient signals instead of periodic. The method of creating an

auditory image of periodic vowels in AIM ensures the alignment of all the maxima correctly. The method employed in tAIM is simplified due to the signals' transient nature, but the author would have preferred to have had the opportunity to create a more adaptive alignment method, though it is unclear whether or not this would improve the results of tAIM.

Finally, a summary of the key contributions of this work:

- Spectral cues have been found to be crucial in a size discrimination task, with the difference between prominent resonances in comparison signals being the most important cue.
- Size information in the frequency spectrum is not limited to one area of the spectrum, and that transient signals that have been heavily high-pass filtered signals can still be discriminated for size, suggesting that there is possible less low-pass filtering in the auditory filter than estimated by current models.
- An auditory model for the purpose of creating auditory images of transient signals, tAIM, has been created. This model can normalise for size of transient signals using the Mellin transform, and extract Mellin phase dy/df values that can provide information towards the discrimination between objects of two different sizes.

9.2 Future Work

The robustness of the human auditory system to hear size despite the signals being presented heavily filtered compared with the difficulty of tAIM to analyse some transient signals presents a few possible research questions.

Does the dcGC filter apply enough compression? The results of the psychophysics showed that with that addition of F1 to the LPF signals, fewer errors in discrimination were made. The F1s for the spheres ranged from 3531 – 5857 Hz. These are above the frequencies wherein most of the vowel and speaker information lies in speech sounds. Furthermore, the removal of F1 in high-pass

filtered spheres improved performance due to the differences between F2s for the spheres was larger than the differences for F1, even though both types of high-pass filtered signals showed scores significantly above chance. The auditory images produced by tAIM were aligned versions of the basilar membrane motion created by the dcGC filterbank. Without limits applied to the images in figure 69, neither F1 nor F2 were visible. The extraction of the T-0 from the SSI compared with the PSDs of the spheres shows that there was very little compression applied to the signals with the filterbank. Considering the importance of F1 in the results of the experiment, it seems that further investigation into the compression of the basilar membrane is needed. Also, the AIM contains low-pass filtering after the BMM process, which would involve further removal of high frequencies. This reduction in high frequency content seems excessive considering the results of the experiment in this study, and warrants further investigation.

The simulated underwater scattered signals from cylinders were processed by tAIM for size discrimination by the model. The model was unsuccessful when the signals were unfiltered. The PSDs of the cylinders showed that the cylinders had spectral envelopes that appeared to be similar. However, when the signals were analysed by tAIM, the T-0 extraction showed poor resolution, and for the air-filled cylinders, the energy in the ERB spectrum appeared to occur in the same frequencies for the different sizes. Anecdotal evidence, however, suggests that the sizes are discriminable and so the spectra are different. The T-0 extraction for the other signals also suggests that the resolution of the dcGC filterbank may be lower than the results of the psychophysics suggests, especially for the higher frequencies.

Above all, any continuation of this research should begin with acquiring more recordings of the polystyrene shapes for spectral analysis and to test tAIM. The inconsistencies in the spectra of the cones and cubes suggest that the recordings may have been compromised by human error or poor recording design, but also there is the possibility that some shapes do not produce spectral envelopes that shift with frequency like spheres and cylinders. A database of recordings from different shapes and sizes would allow for further research into size and shape discrimination, and would also allow for the improvement of tAIM to make it more

robust, with the possibility of extending it to produce estimates of relative size differences between different types of objects.

Finally, this study has resolved that which has been suggested by previous research, that it is indeed the spectral content of a transient signal which contains the most important cues for size discrimination, and has pinpointed this exactly in the difference between the resonances. tAIM has proven to be capable of discriminating for size of simulated and recorded objects, with few limitations. This research can go towards further improving the modelling of the auditory system for signals other than periodic, and also toward improving technology in the field of object recognition and identification for both hearing implant applications but also in the area of military defence. Above all, tAIM has provided the groundwork for further work into the future of automated size discrimination models based on the biological principles of human hearing.

References

Anderson, J.D. and B.F. Fisher. (eds) (2005). "Moving Image Theory – Ecological Considerations". Southern Illinois University Press.

Arias, C. and O.A. Ramos (1997). "Psychoacoustic tests for the study of human echolocation ability." Applied Acoustics **51**(4): 399-419

Au, W. W. L. and D. A. Pawloski (1992). "Cylinder wall thickness difference discrimination by an echolocating Atlantic bottlenose dolphin." Computational Physiology **170**: 41-47.

Au, Whitlow W. L. (1997). "Echolocation in dolphins with a dolphin-bat comparison". Bioacoustics: The International Journal of Animal Sound and its Recording, **8**:1-2, 137-162

Au, W. W. L., A. N. Popper, et al.,(eds) (2000). "Hearing by Whales and Dolphins". Springer Handbook of Auditory Research.

Bergman, A.S. (1994). "Auditory Scene Analysis: The perceptual organisation of sound". MIT Press: Cambridge, MA.

Bleeck, S., T. Ives, and R.D. Patterson (2004). "Aim-mat: The Auditory Image Model in MATLAB." Acustica **90**: 781-787.

Borden, G. J., K. S. Harris, and L.J. Raphael (2003). "Speech Science Primer: Physiology, Acoustics, and Perception of Speech". Lippincott, Williams & Wilkins.

Carello, C., J.B. Wagman and M.T. Turvey (2005). "Acoustic Specification of Object Properties". In: Anderson, J.D. and B.F. Fisher (eds), Moving Image Theory – Ecological Considerations. Southern Illinois University Press. p79-104.

Carello, C., K. L. Anderson, and A.J. Kunkler-Peck (1998). "Perception of object length by sound." Psychological Science **9**(3): 211-214.

Chittka,L and A. Brockmann (2005). "Perception Space – The Final Frontier". Available at http://www.soundonsound.com/sos/mar11/articles/how-the-ear-works.htm

Clarke, E. (2009). "Investigating human auditory perception of 3D polystyrene object size and shape parameters". Masters Dissertation, ISVR, University of Southampton.

Darwin, C.J. (2005). "Pitch and Auditory Grouping". In: Plack, C.J., A.J. Oxenham and R.R. Fay (eds), Pitch: Neural Coding and Perception". Springer Handbook of Auditory Research, p278-305

de Boer, E. and H. R. de Jongh (1977). "On cochlear encoding: Potentialities and limitations of the reverse-correlation technique." Journal of Acoustical Society of America **63**(1): 115-135.

De Boer, E. and P, Kyuper (1968). "Triggered Correlation". IEEE Trans, Biomed Eng. BME-15, 169-179

de Boer, E. and A. L. Nuttall (1997). "The mechanical waveform of the basilar membrane. I. Frequency modulations ("glides") in impulse responses and cross-correlation functions." Journal of Acoustical Society of America **101**(6): 3583-3592.

De Sena, A. and D. Rocchesso (2005). "A study on using the Mellin transform for vowel recognition". Sound and Music Computing '05 Salerno, Italy.

DeLong, C. M., W. W. L. Au, et al. (2007). "Echo features used by human listeners to discriminate among objects that vary in material or wall thickness: Implications for echolocating dolphins." Journal of Acoustical Society of America **121**(1): 605-617.

Elragi, A. F. (2006) "Selected Engineering Properties and Applications of EPS Geofoam - Introduction." Softoria Group. Avaliable at: http://www.softoria.com/institute/geofoam/material.html. Accessed 26 Sep. 2011.

Evans, W.E. and B.A.Powell (1967). "Discrimination of different metallic plates by an echolocating delphinid". In R.G.Busnel (ed) Animal Sonar Systems: Biology and Bionics. Laboratoire de Physiologie Acoustique, Jour en Josas, France.

Fitch, W.T. and J.Giedd (1999). "Morphology and development of the human vocal tract: a study using magnetic resonance imaging." Journal of Acoustical Society of America **106**: 1511-1522.

Fletcher, H. (1940). "Auditory Patterns". Reviews of Modern Physics **40**

Fox, P., S. Bleeck, P.R. White, T.G. Leighton and V.F. Humphrey (2007). "Initial results on size discrimination of similar underwater objects using a human hearing model". Proceedings of the Institute of Acoustics **29**(6).

Gasser, M. (2009). "How language works. 3.3 Vowels". Available at: http://www.indiana.edu/~hlw/PhonUnits/vowels.html . Accessed 6 June 2012.

Gaver, W.W. (1993). "What in the world do we hear? An ecological approach to auditory event perception". Ecological Psychology. **5**(1): 1-29.

Gelfand, S.A. (2007). "Hearing: An Introduction to Psychological and Physiological Acoustics". 4th Ed. Informa Healthcare USA, Inc.

Giordano, B.L. and S. McAdams (2006). "Material identification of real impact sounds: Effects of size variation in stell, glass, wood and plexiglass plates". Journal of the Acoustical Society of America **119**(2)L 1171-1181.

Glasberg, B. R. and B. C. J. Moore (1990). "Derivation of auditory filter shapes from notched noise data." Hearing Research **47**: 103-138.

Gordon, M.S. and L, Jarquin (2000). "Echolocating distance by moving and stationary listeners". Ecological Psychology **12**(3): 181-206.

Grassi, M. (2002). "Recognising the size of objects from sounds with manipulated acoustical parameters." In J. A. Da Silva, E. H. Matsushima, & P. Ribeiro-Filho (Eds.), Fechner Day 2002: Proceedings of the International Society for Psychophysics: 392-397. Rio de Janeiro, Brasil.

Grassi. M. (2005). "Do we hear size or sound? Balls dropped on plates." Perception and Psychophysics **67**(2): 274-284

Grinnell, A.D. (1995). "Hearing in bats: An overview". In A.N.Popper and R.R.Fay (eds), Hearing in Bats, Springer-Verlag

Greenwood, D.D. (1961)."Critical bandwidth and the frequency coordinates of the basilar membrane." Journal of the Acoustical Society of America **33**(10) 1344-1357

Houben, M. M. J., A. Kohlrausch, and D.J. Hermes (2004). "Perception of the size and speed of rolling balls by sound." Speech Communication **43**: 331-345.

Houtgast, T. (1977). "Auditory-filter characteristics derived from direct-masking and pulsation-threshold data with a rippled-noise masker". Journal of Acoustical Society of America **62**(2): 409-415

Irino, T. (1995). "An optimal auditory filter." IEEE ASSP Workshope on Application of Signal Processing to Audio and Acoustics: 198-201.

Irino, T. and R. D. Patterson (1997). "A time-domain, level-dependent auditory filter: The gammachirp." Journal of Acoustical Society of America **101**(1): 412-419.

Irino, T. and R. D. Patterson (1999). "Stabilised wavelet Mellin transform: an auditory strategy for normalising sound-source size". Proceedings Eurospeech '99.

Irino, T. and R. D. Patterson (2002). "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabiliised wavelet-Mellin transform." Speech Communication **36**: 181-203.

Irino, T. and R. D. Patterson (2006). "A dynamic, compressive gammachirp auditory filterbank." IEEE Transactions on Audio, Speech, and Language Processing **14**(6): 2222-2232.

Johannesma, P.I.M. (1972). "The pre-response stimulus ensemble of neurons in the cochlear nucleus". Proceedings of the Symposium on Hearing Theory, p 58-69, IPO, Eindhoven, The Netherlands.

Kawahara H, Katayose H, de Cheveigné A, Patterson R D (1999) "Fixed point analysis of frequency to instan-taneous frequency mapping for accurate estimation of F0 and periodicity", In Proc. EUROSPEECH'99,ESCA.**6**, 2781–2784

Kellogg, W.N. (1962). "Sonar System of the Blind", Science **137** 399-404

Lufti, R.A. and R.D. Patterson (1984). "On the growth of asymmetry with stimulus intensity". Journal of Acoustical Society of America **76**: 739-745.

Lyon, R.F. and C.A. Mead (1988). "An analog electronic cochlea". IEEE Trans Acoustic, Speech and Signal Processing **36**(7): 1119-1134

Lyon, T. (2010a). "Investigating the effect of formant spacing and duration on the frequency discrimination ability of vowel like stimuli". MSc Dissertation, ISVR, University of Southampton.

Lyon, Richard F. (2010). "Machine Hearing: An Emerging Field". IEEE Signal Processing Magazine [135]

Lufti, R.A. (2008). "Human sound source identification". In Yost, W.A., Popper, A.N. & Fay, R.R. (eds) Auditory Perception of Sound Sources Springer, pp.13~42

McClellan, M.E. and A.M. Small (1966). "Time separation pitch associated with noise pulses". Journal of the Acoustical Society of America **40**, pp. 570–582.

McGrath, R., T. Waldmann, and M. Fernström (1999). "Listening to rooms and objects". AES 16TH International conference on Spatial Sound Reproduction.

Meddis, R. (1986). "Simulation of mechanical to neural transduction in the auditory receptor". Journal of the Acoustical Society of America **79**(3): 702-711.

Micheyl, C., K. Delhommeau, X. Perrot and A.J. Oxenham (2006). "Influence of musical and psychoacoustical training on pitch discrmination." Hearing Research **219**: 36-47.

Moore, B.C.J. (2003). "An Introduction to the Psychology of Hearing". 5th Ed. Academic Press

Moore, B.C.J. and B.R.Glasberg (1983). "Suggested formulae fo calculating auditory-filter bandwidths and excitation patterns." Journal of the Acoustical Society of America **74**(3): 750-753

Nakahara, F., A. Takeura and T. Koido (1997). "Target discrimination by an echolocating finless porpoise, Neophocaena Phocadnoides". Marine Mammal Science **13**(4): 639-649.

Patterson, R.D. (1974). "Auditory Filter Shape". Journal of the Acoustical Society of America **55**(4): 802-809

Patterson, R.D. (1976). "Auditory filter shapes derived with noise stimuli." Journal of the Acoustical Society of America **59**(3): 640-654Patterson, R. D., M. H. Allerhand, and C. Giguere (1995). "Time-domain modelling of peripheral auditory processing: A modular architecture and a software platform." Journal of Acoustical Society of America **98**(4): 1890-1894.

Patterson, R.D. and B.C.J. Moore (1986). "Auditory filters and excitation patterns as representations of frequency resolution". In Frequency Selectivity in Hearing, edited by B.C.J. Moore. Academic, London.

Patterson, R. D. and I. Nimmo-Smith (1980). "Off-frequency listening and auditory-filter asymmetry." Journal of Acoustical Society of America **67**(1): 229-245.

Patterson, R., I. Nimmo-Smith, J. Holdsworth and P. Rice (1987). "An efficient auditory filterbank based on the Gammatone function". Appl. Psychol. Unit, Cambridge University.

Patterson, R. D., K. Robinson, J. Holdsworth, D. McKeown, C. Zhang and M. Allerhand (1992). "Complex sounds and auditory images." Auditory physiology and perception, Proc. 9th International Sympsonium on Hearing: 429-446.

Patterson, R.D., M.H.Allerhand and C.Giguere (1995). "Time-domain modelling of peripheral auditory processing: A modular architecture and a softward platform." Journal of the Acoustical Society of America **98**(4): 1890-1894

.Patterson, R.D., D.R.R. Smith, R. van Dither and T.C. Walters (2007). "Size information in the production and perception of communication sounds". In: Yost, W.A., A.N. Popper and R.R. Fay (eds), Auditory Perception of Sound Sources. Springer Handbook of Auditory Research. **29**: 43-75.

Peterson, G.E., (1952). "The information-bearing elements of speech". Journal of Acoustical Society of America **24**: 629-637. In Gelfand, 2007, 4ed.

Peterson, G.E. and H.L.Barney (1954). "Control methods used in a study of the identification of vowels." Journal of the Acoustical Society of America **24**: 175-184

Plomp, R. (1970). "Timbre as a multidimensional atribute of complex tones". In Plomp, R. & G.F. Smoorenburg (eds), Frequency analysis and periodicity detection in hearing. Leiden: Sijthoff, p397-414

Riesz, R.R. (1928). "Differential intensity sensitivity of the ear for pure tones". Physical Review **31**: 867-875

Rice, C.E. (1967). "Human Echo Perception". Science **155** (3763): 656-664.

Rosen, S., Baker, R. J., Kramer, S. (1992). "Characterizing changes in auditory filter bandwidth as a function of level". Auditory Physiology and Perception **83**: 171-177.

Rosen, S. and R.J. Baker (1994). "Acoustic reflexes in the measurement of auditory filters at high-levels in normal listeners". Audiology **33**(1), 37-46

Russell, Roger (2008). "Listening and Hearing". Available at http://www.roger-russell.com/hearing/hearing.htm (Accessed 13 February 2011)

Schafer, T.H., R.S.Gales. C.A.Shewmaker and P.O.Thompson (1950). "The frequency selectivity of the ear as determined by masking experiments." Journal of the Acoustical Society of America **22**(4): 490-496

Schofield, D. (1985). "Visualisations of speech based on a model of the peripheral auditory system. " NPL Report DITC 62/85

Schubert, E., J. Wolfe and A. Tarnopolsky (2004). "Spectral centroid and timbre in complex, multiple instrumental textures." Proceedings of the 8th International Conference on Music Perception & Cognition: 654–7.

Simon, R., M.W. Holderied and O. von Helverson (2006). "Size discrimination of hollow hemispheres by echolocation in a nectar feeding bat." The Journal of Experimental Biology **209**: 3599-3609

Smith, D. R. R., R. D. Patterson, et al. (2005). "The processing and perception of size information in speech sounds." Journal of Acoustical Society of America **117**(1): 305-318.

Smith, D. R. R., R. D. Patterson, et al. (2005). "The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex and age". Journal of Acoustical Society of America **118**(5): 3177-3186.

Supa, M., M. Cotzin, and K.M. Dallenbach (1994). " 'Facial Vision': The perception of obstacles by the blind". The American Journal of Psychology **57**(2): 133-183.

Turner, R.E., T.C. Walters, J.J.M. Monaghan & R.D. Patterson (2009). "A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data". Journal of Acoustical Society of America **125**(4): 2374-2386.

Unoki, M. and T. Irino (2006). "Comparison of the roex and gammachirp filters as representations of the auditory filter." Journal of Acoustical Society of America **120**(3): 1474-1492.

van den Doel, K. and D.K. Pai (1998). "The sounds of physical shapes". Presence **7**(4): 382-395.

Vanderveer, N.J. (1979). “Ecological acoustics: Human perception of environmental sounds”. Doctoral dissertation, Cornell University. In Giordano & McAdams (2006).

Viemeister, N.F. and S.P. Bacon (1988). “Intensity discrimination, increment detection, and magnitude estimation for 1-kHz tones”. Journal of Acoustical Society of America **84**: 172-178

von Helversen, D. (2004). "Object classification by echolocation in nectar feeding bats: size-independent generalisation of shape." Computational Physiology **190**: 515-521.

Weir, C.C, W. Jesteadt & D.M. Green (1977). “Frequency discrimination as a function of frequency and sensation level”. Journal of Acoustical Society of America **61**(1): 178-184.

Wright, H.N. (1964). “Background masking for tones in narrow-band noise.” Journal of the Acoustical Society of America **36**: 2217-2221

Appendix

Questionnaire

Questionnaire for Size Discrimination Experiment

All answers and results will be kept confidential.

First name: _____

Surname: _____

M/F: _____

Date of Birth: _____

Handedness: Left or Right ? (please circle as appropriate)

How is your hearing?

Have you been exposed to loud noises, either prolonged or otherwise, in the last 48 hours? Please explain.

_____

_____

_____

Do you or your immediate family have any known hearing problems?

_____

_____

_____

Do you have trouble understanding speech in loud background noise?

_____

Do you suffer from troublesome tinnitus?

_____

Are you currently on any medication or suffering from a headcold?

_____

Musical background

Have you had musical training or do you play a musical instrument? Please specify.

_____

_____

For how long have you played?

_____

Do you have perfect pitch?

_____

Have you ever taken part pitch discrimination experiments before? Please elaborate.

_____

_____

_____

After testing

How would you describe what you were listening for when you were answering the questions “Which sound is bigger?”

__ Length difference __ Frequency/pitch __ Timbre

__ Loudness difference __ Other, please explain

_____

_____

_____

If two signals sounded the same, how did you decide which sounded bigger?

_____

_____

_____

Calculation of noise exposure for each participant for day of testing.

Guide to Experimentation involving Human Subjects

by

Human Experimentation Safety and Ethics Committee

ISVR Technical Memorandum No. 808, October 1996

Daily exposure is calculated below:

For a single stimulus:

The fractional dose, *f*, is given by: $f = t/8$ antilog [0.1 (L-76)]

where L is the A-weighted sound level of the stimulus and t is the exposure duration, in hours.

Bruel and Kjaer Sound Level Meter and an artificial ear calculated the level of each individual 50 ms signal to be a maximum of 75 dB (A).

A signal recording was 50 ms in length, or 0.000013888 Hrs.

$$f = (0.000013888/8) \times 10^{\wedge}(0.1 \times (75 - 76)) = 1.3790 \times 10^{-6}$$

For more than one stimulus:

The total dose, F, is given by : $F = f1 + f2 + \ldots + fn$, where $f1$ to fn are calculated as above.

A single recording of a sphere signal was 50ms, repeated 5 times in pairs, with 40 pairs per session, and the option to repeat (avg = 2). There were no more than 4 sessions in any one session on any one day.

No. of single stimulus = 40 x 5 x 2 x 2 x 4 = 3200 stimuli in any one day.

$$F = 1.3790 \times 10^{-6} \times 3200 = 0.0044$$

Daily exposure dB (A) = 10 log (F) + 76 dB (from the table below) = **52.43 dB**

(Total time exposure in a day = 3200 x 50 ms = 2.6667 minutes)

Table 1 Sound levels above which an experiment is defined as UNUSUAL

Total 'on-time' during any 24 hour period	Sound level which defines an 'UNUSUAL' experiment
8 h	76 dB(A)
4 h	79 dB(A)
2 h	82 dB(A)
60 min	85 dB(A)
30 min	88 dB(A)
15 min	91 dB(A)
12 min	92 dB(A)
6 min	95 dB(A)
3 min	98 dB(A)
90 s	101 dB(A)
45 s	104 dB(A)
36 s	105 dB(A)
22.5 s	107 dB(A)
11.2 s	110 dB(A)
5.6 s	113 dB(A)
2.8 s	116 dB(A)
1.4 s	119 dB(A)
any duration	120 dB(A)