SIZE DISCRIMINATION OF TRANSIENT SOUNDS: PERCEPTION AND MODELLING

Niamh O'Meara, Stefan Bleeck

Institute of Sound and Vibration Research, Highfield, Southampton, United Kingdom

Corresponding author: Stefan Bleeck, Institute of Sound and Vibration Research, University Road, Highfield, Southampton, SO17 1BJ, United Kingdom, e-mail: bleeck@gmail.com

Abstract

Humans are able to get an impression of the size of an object by hearing it resonate. While this ability is well described for periodic speech sounds we investigate here the ability to discriminate the size of non-periodic transient impact sounds. Three experiments were performed on normal listeners (n=19) to investigate the importance of the spectral cue in different frequency regions. Recordings from pulse resonance sounds made by a metal ball hitting polystyrene spheres of 5 different sizes were used in the experiments. Recordings were manipulated in order to show that the same cues used in speaker size discrimination are used for transient signals. Results show that the most prominent resonances are the most important cue, but frequencies above 8 kHz also contribute. The results are explained by physiologically inspired model of size discrimination that is based on the Auditory Image Model, and its key part is the Mellin transform. The model can predict which of two objects is bigger. We conclude that similar cues that are used for speaker size discrimination are important for transient sounds.

Keywords: psychophysics • hearing sciences • sound perception

DIFERENCIACIÓN DE TAMAÑOS CON SONIDOS DE CORTA DURACIÓN: PERCEPCIÓN Y MODELACIÓN

Resumen

El hombre es capaz de apreciar el tamaño de un objeto al oír la resonancia que produce. Aunque esta capacidad está bien descrita en publicaciones en lo relativo a los sonidos periódicos del habla, en estudio queremos examinar la capacidad de diferenciar el tamaño con ayuda de sonidos no periódicos de corta duración producidos por golpes. Se realizaron tres experimentos en los que participaron personas que oyen bien (*n*=19) con el objetivo de estudiar el significado de las indicaciones procedentes del espectro acústico en varios ámbitos de frecuencia. En el estudio se utilizó el registro de la resonancia de sonidos procedentes de un solo impulso producido por una bola de metal al golpear bolas de poliestireno vaciadas con cinco diámetros diferentes. Los registros fueron procesados de tal modo que modo que garantizaran que las mismas indicaciones fueran usadas tanto para discriminar el tamaño del hablante como en el caso de los sonidos de corta duración. Los resultados muestran que la indicación más importante son las resonancias más claras, pero también influyen las frecuencias superiores a 8 kHz. Los resultados están explicados mediante un modelo de reconocimiento del tamaño basado en la fisiología, que fue creado en base a un modelo del sistema auditivo y cuyo elemento más importante es la transformada de Mellin. Este modelo puede prever cuál de los dos objetos es mayor. Nuestra conclusión es que indicaciones similares a las utilizadas para reconocer el tamaño del hablante también son importantes para los sonidos de corta duración.

РАЗЛИЧЕНИЕ РАЗМЕРОВ ПРИ КРАТКОВРЕМЕННЫХ ЗВУКАХ: ВОСПРИЯТИЕ И МОДЕЛИРОВАНИЕ

Изложение

Человек может оценить размер объекта, слыша резонанс, который он вызывает. Хотя эта способность хорошо описана в литературе по отношению к периодическим звукам речи, в этом исследовании мы хотели бы исследовать способность различения размера с помощью непериодических, кратковременных звуков, создаваемых путем ударов. Проведены три эксперимента с участием нормально слышащих людей (n=19), чтобы исследовать значение показателей, происходящих из акустического спектра в разных пределах частот. В исследовании использовалась запись резонанса звуков, происходящих с отдельного импульса, вызванного маталлическим шаром, ударяющим в полые стиропорные шары пяти разных диаметров. Записи были переработаны таким образом, чтобы обеспечить использование таких самых показателей как для дискриминации размера говорящего человека, так и в случае кратковременных звуков. Результаты показывают, что самый важный показатель – самые отчетливые резонансы, но влияние имеют также частоты выше 8 кГц. Результаты объяснены с помощью основанной на

физиологии модели различения размера, которая была создана на основании модели слуховой системы, а ее самым важным элементом является преобразование Меллина. Эта модель может предугадать, который из двух объектов больше. Наш итог –подобные указатели, как те, которые были использованы для опознавания размера говорящего человека, являются также важными для кратковременных звуков.

ROZRÓŻNIANIE ROZMIARÓW PRZY KRÓTKOTRWAŁYCH DŹWIĘKACH: PERCEPCJA I MODELOWANIE

Streszczenie

Człowiek jest w stanie ocenić rozmiar obiektu słysząc rezonans, który on wywołuje. Chociaż zdolność ta jest dobrze opisana w literaturze w odniesieniu do okresowych dźwięków mowy, w tym badaniu chcielibyśmy zbadać zdolność rozróżniania rozmiaru za pomocą nieokresowych, krótkotrwałych dźwięków wytwarzanych przez uderzenie. Przeprowadzone zostały trzy eksperymenty z udziałem osób normalnie słyszących (n=19) w celu zbadania znaczenia wskazówek pochodzących z widma akustycznego w różnych zakresach częstotliwości. W badaniu wykorzystano zapis rezonansu dźwięków pochodzących z pojedynczego impulsu wywołanego przez metalową kulę uderzającą o wydrążone kule styropianowe o pięciu różnych średnicach. Zapisy zostały przetworzone w taki sposób, by zapewnić, że te same wskazówki były używane zarówno do dyskryminacji rozmiaru osoby mówiącej, jaki w przypadku krótkotrwałych dźwięków.. Wyniki pokazują, że najważniejszą wskazówką są najwyraźniejsze rezonanse, ale wpływ mają także częstotliwości powyżej 8 kHz. Wyniki są objaśnione za pomocą opartego na fizjologii modelu rozpoznawania rozmiaru, który został stworzony na podstawie modelu układu słuchowego i którego najważniejszym elementem jest przekształcenie Mellina. Model ten potrafi przewidzieć, który z dwóch obiektów jest większy. Naszym wnioskiem jest, że podobne wskazówki jak te wykorzystywane do rozpoznawania wielkości osoby mówiącej są także ważne dla dźwięków krótkotrwałych.

Background

The human auditory system has arguably evolved to perceive and understand the acoustic environment around us in order to communicate and survive. Acoustic signals radiated from animate or inanimate objects contain information about the object's size, material and shape, and humans are capable of discriminating between two objects when they differ in any one of these properties. These signals can be transient in nature such as a knock on a door, where the action of striking the door causes the door to vibrate and resonate, or periodic as in the case of vowel sounds from human speech. Vowels can also be thought of as a train of transients, or pulse/resonance sounds, where the repetitions of the pulses equate to the vibration of the vocal folds, and the resonances give the information on the shape of the mouth and the length of the vocal tract, and therefore indicating the vowel uttered and the size of the speaker. The repetition of the same signal increases the signal to noise ratio against a single presentation, but even a single impulse carries enough information for correct discrimination [1] to a degree.

The vocal tract length of a human is directly related to the size of the human; 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 differences in vocal-tract length, the same speech uttered by all three humans can be recognised by a listener. It is an ability to ignore the absolute frequencies of the resonances, but hear their relative positions with respect to each other that allow humans to identify which vowel has been uttered [2], thus effectively normalizing for size, and listening to shape. However, the listener can simultaneously extract information about the size of the speaker by learning that the spectral envelope of the vowel formants shift up and down in frequency according to the size and that the spectral relationship between the formants represents the vowels [3]. This supports the hypothesis that some form of scaling transform (or size normalization) is applied to the sounds to remove any vowel confusion that may arise from dealing with speakers of very different sizes.

In order to create images of periodic sounds, the Auditory Image Model (AIM) was originally developed on the basis of spectral information [4] to model pitch perception. AIM is a time-domain filterbank model of the auditory system that analyses complex periodic vowels. The output of AIM is the Stabilised Auditory Image (SAI) – stabilised and static visual representation of a complex periodic signal. The modules of standard AIM (as described in [5]) carry out the tasks shown in Table 1.

Most bio-acoustic communication sounds contain independent size and shape information, and it was suggested that one of the tasks of the auditory system is to perform a segregation of the size and shape information [6]. In this description, a normalisation procedure allows the listener to disregard the size information and keep the size-invariant properties. In order to model this size normalisation, AIM was expanded by [7] to include the Mellin transform that normalised for size and created Mellin Images pertaining to shape only. The Mellin transform is applied to the output of AIM, the stabilised auditory images, and the resultant Mellin Images show the similar pattern for the same vowel spoken by speakers of different sizes. The success of this transform is based on the fact that the relationship between the resonances of the vowels remains constant despite a change in the size of the speaker.

Humans can also estimate the size of objects that do not produce periodic sounds. The human auditory system is capable of discriminating between the sizes of objects when they are struck once, producing a single pulse resonance.

Table 1. Modules in aim-mat, and the equivalent biological process (from [5])

Modules of aim-mat	Auditory System equivalent
Pre-cochlear Processing	Band-pass filtering of outer ear & ear canal
Gammatone filterbank	Spectral analysis of the basilar membrane
Neural Encoding	Half-wave rectification – simulation of the uni-directional movement of the inner hair cell stereocilia Compression – simulates the non-linear input-output function of the hair cell response Low pass filtering – simulates the progressive loss of phase-locking of neuronal action potential with higher frequencies
Strobed Temporal Integration	Averaging over periods of a continuous signal to stabilise and create static stabilised auditory image

Table 2. The dimensions and masses of the recorded polystyrene spheres used in all experiments

	X.Large (XL)	Large (L)	Medium (M)	Small (S)	X.Small (XS)
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

Carello et al. [8] showed that we can reliably estimate the absolute length of rods (from 30 cm to 120 cm) dropped on a linoleum floor. Participants were able to order rods from short to long without any standard of comparison. The authors concluded that the resonance frequencies of the rods due to the different length provided the acoustic cue for length discrimination and not amplitude or signal duration.

Houben et al. [9] performed an experiment where participants listened to recordings of pairs of wooden balls with different diameters (between 22 and 83 mm) rolling over a wooden plate. The recordings were equated in both duration and acoustic energy to remove any signal length and intensity cues, though there still remained some slight temporal information in the amplitude modulation of the real (not perfectly spherical) balls. Participants could identify the larger ball in all pairs except for those with the smallest diameter. The authors suggested that the differences in spectral centroid frequency (SCF) play a larger part in the size discrimination abilities of their participants than temporal cues, but they could not discount the amplitude modulation or other interferences caused by the action of rolling. Humans have also been shown to be able to discriminate the size of wooden balls dropping on plates [10]. The authors showed that the acoustic energy of the signals at low frequencies was the most important cue and listeners only needed to hear the first bounce. This indicated an ability to detect size from a single very short-duration impact sound.

Humans have the ability to discriminate between the sizes of objects that produce transient non-periodic sounds, but it is still not clear which acoustic cues - spectral, temporal or intensity – are most important. Spectral information, notably the resonances, is vital in speaker size discrimination [6], and based on this AIM with the Mellin transform was developed for speaker size normalisation. However, AIM is not capable of creating stabilised auditory images of transient single pulse-resonances because the strobing mechanism in AIM requires at least two periods of a waveform [5].

Here we investigate whether this ability to discriminate size relies on the same spectral cues as in speech discrimination. To do so, real acoustic signals were recorded using polystyrene spheres that were struck with a small metal ball. We report a series of subjective experiments that determine the importance of spectral cues and describe a new model that can discriminate size of transient signals based on the same cues. Using objects of the same shape, we show that the model can discriminate between different sizes, and correctly order the objects by increasing size.

Material and Methods

Three sets of polystyrene Styrofoam spheres of 5 sizes (70 mm, 80 mm, 90 mm, 100 mm and 120 mm diameters) were suspended from the ceiling in an acoustically insulated room, and struck by a pendulum: a metal ball bearing of 10 mm diameter (5.6 g). Polystyrene spheres were chosen because their sounds provide a high ecological validity (i.e. people would usually have experienced the sounds before) and they have a low impedance, thus creating relatively long and loud sounds with a rich harmonic structure. Table 2 indicates the sizes and masses of the spheres used. Density varied in the range from 20.4 to 22.9 kg/m^3. Sounds were recorded through a free field microphone in a sound proof room. The microphone was placed 11 cm radial to the point of impact. All efforts were made to ensure the impact position and force applied was as uniform as possible. Fifteen sets of recordings were made: each sphere was struck 100 times and recorded, making a total of 1500 recordings with which viable averaged signals for each size were created.

Each recording was band pass filtered (Butterworth 4th order) between 100–16,000 Hz. The filtered signals were then edited using a MATLAB script to the same length and aligned to the point of the first negative peak. Within each

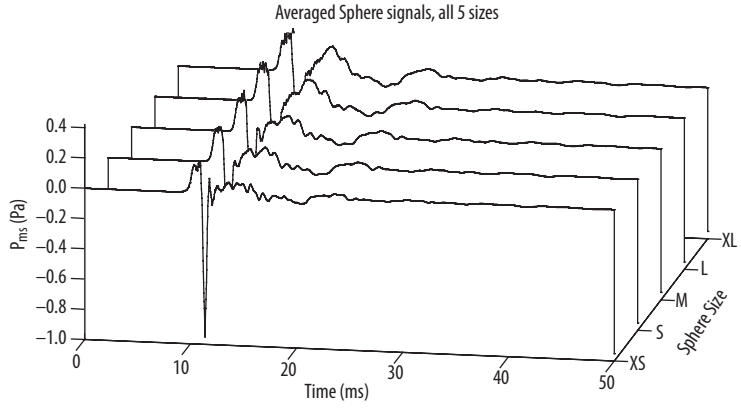

Figure 1. Time series representations of the averaged versions of the 5 sphere signals

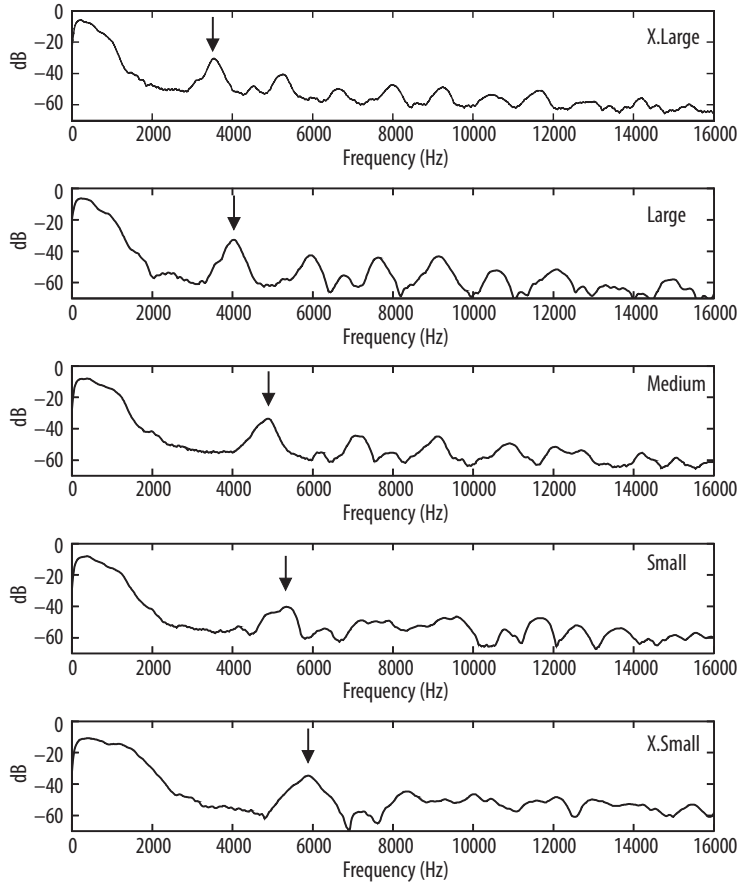

Figure 2. Welch Power Spectral Density (PSD) plots for each spheres. Arrows indicate the position of the resonance F1. F1 increases with decreasing sphere size. F0 is the broad band peak to the left. Higher resonances are visible to the right of F1

set of recordings, the signals were screened automatically for uniformity, and those that had RMS values greater than two standard deviations from the mean of their set were removed. Remaining signals were averaged to create one signal per size that all had the same RMS and length. Each signal was cut off after 50 ms, but almost all energy is between 10 and 30 ms. There was no perceptible contribution of the sound of the striking object due to the large difference in acoustic impedance. Figure 1 shows the resulting averaged time series of all 5 signals and Figure 2 shows the respective spectral density plots. The waveforms look similar but the spectra reveal differences mainly in the position of the resonant peaks. The first broad frequency peak between 0 and 2000 Hz is very similar in all signals and is called F0 in this paper. The resonances are visible as subsequent higher frequency peaks. These are very different between spheres and specifically the first peak (called F1 in this paper, indicated by black arrows in the figure) shows an increase with decreasing sphere size. Further resonances are called F2, etc.

Three experiments were conducted in order to determine the importance of the spectral cue in discriminating the size of the described spheres, and to determine the most informative region of the spectrum. In experiment 1 we

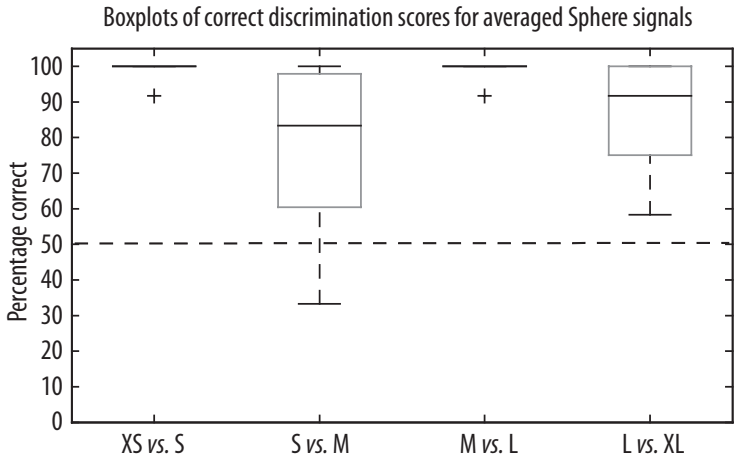

Figure 3. Box plots of results for basic size discrimination task. The line inside each box indicates the median, the boxes show the interquartile range and the whiskers represent 1.5 IQR. Crosses are outliers. The dotted line at 50% is the chance level for correct discrimination in this experiment

investigated if the recordings are valid sounds by asking whether participants could tell which of two spheres sounds larger. Experiment 2 investigated if spectral cues are important for size discrimination used warped signals. Finally in experiment 3 the signals were filtered in various ways to determine which region of the spectrum is most informative for size discrimination. High-pass, low-pass and band-stop filtering was used, with extra focus on the importance of the strongest resonance.

Testing procedure

Thirty-two participants between the ages of 18 and 35 years (average 26 years) recruited from the University of Southampton (24 males) completed all three experiments. All reported normal hearing. The testing was split into five 30 minute sessions, one session a week. Level of stimulation was set to 75 dB(A) as calibrated by an artificial ear for a 50 ms signal. Because of the impulsive nature of the signals this was a comfortable listening level. At the start of the first session the method for the experiment was explained and participants were given time to familiarise themselves with the polystyrene spheres (for example by bouncing or striking them). After familiarisation all reported that they could easily hear the size differences. During the experiments the spheres were visible to the subjects but they were not allowed to touch them. Apart from this familiarisation there were no training sessions for any experiments. No feedback of correct results was given at any stage.

Of the 32 participants, 13 were excluded for the analysis later because their results were inconsistent leaving 19 individuals. Consistency was measured by a reliability index calculated as follows: each time a repeated presentation (only unfiltered and unscaled sounds were counted) gave the same response the index was increased by one; each time a different answer was given it was reduced by one. A person guessing would score 0. Only participants that scored a normalised score of at least 0.6 on average were considered for further analysis. Note that our definition of reliability does not indicate correctness, only consistency. Reliability was highly correlated with musical ability (self-reported): 40% of the non-musical participants were reliable compared to 68.2% of those that were musical. Note that the criterion of 0.6 is rather strict. We deliberately chose a high reliability threshold criterion because we aimed to investigate the influence of specific cues and not general population ability levels and we chose to exclude all but the most reliable participants. All but 4 of our 19 consistent participants performed significantly above chance level on average. All experiments were carried out under the approval of the Human Experimentation Safety and Ethics Committee, Institute of Sound and Vibration Research, University of Southampton.

All recorded data was analysed statistically using SPSS. Homogeneity of variance was tested using Levene's test and normality with Shapiro-Wilk tests. One-sample or paired sample t-tests were used for normally distributed data, Wilcoxon Signed Rank tests otherwise. Repeated-measures ANOVAs were carried out using a Greenhouse-Geisser correction and significant differences between sets were tested using a Bonferroni Post-hoc test. All reported significances are on a 5% level.

Results

Experiment 1: Size discrimination of polystyrene spheres

A size discrimination task was carried out by presenting pair-wise recordings to test if participants could tell the difference between the sizes of the spheres from the sound they emitted when struck. Each presentation consisted of five identical signals, 200 ms apart. Each signal was 50 ms long. In a 2-alternative forced-choice (2AFC) paradigm, only neighbouring pairs were compared to each other, i.e. XS *vs.* S, S *vs.* M, M *vs.* L, and L *vs.* XL, and the question asked to participants was "Which object sounds bigger?" One test set involved signals taken from the batch of averaged signals, and the other contained signals taken at random from the library of 1500 recordings to test how robust size discrimination was with original (not averaged) sounds.

Results showed that discrimination was significantly better with averaged signals than with the original un-averaged. Therefore averaged signals were used in subsequent experiments. On average, participants could correctly identify the bigger object at 90% (SD ±7.9) (Figure 3). All comparison

pairs are significantly above chance level. Participants indicated in informal feedback that the task was easy. XS *vs.* S and M *vs.* L were answered correctly by all participants apart from one, shown as an outlier. S *vs.* M, and L *vs.* XL pairs are significantly lower than the other comparisons, probably because of differences in the SCF as discussed below.

Experiment 2: Size discrimination of scaled signals

From the results of experiment 1, we are confident that size discrimination of the polystyrene spheres was possible. In order to design a subsequent model of size discrimination we investigated the importance of the spectral cue in experiment 2. The rationale behind this experiment was to attempt to confuse the listeners by adjusting the spectral content of each signal in several ways so as to adjust its perceived size. If this manipulation was successful, and participants could be fooled into perceiving sizes differently, we would establish the importance of the resonance frequency cues. If, however, this manipulation is not successful, other cues (for example temporal modulation or relative intensities) must be taken into consideration.

The simplest method of changing the perceived size is to scale the signals in frequency or, which is equivalent, increase or decrease the playback sample rate (PSR). This method is motivated by more complex speech vocoders that change perceived speaker size such as STRAIGHT (Kawahara et al, 1999). Preliminary attempts at using this method (data not shown) proved unsuccessful: participants (n=5) reported that the signals sounded confusing, because they heard two or more harmonics that contradicted moving in different directions, i.e. one increased in pitch with increasing PSR while the other decreased. This made the decision to choose which sound came from the bigger object impossible. Therefore further attempts in this direction were halted and a more suitable method of altering the size perception was developed.

An improved method of scaling is based on the frequency of the first resonance (F1) in each size. This is motivated by the fact that it has the most energy apart from the F0 and is distinctively different in each sphere. Spectral analysis of the averaged signals show, apart from F0, that F1 contains at least 15 dB more power than all other resonant peaks (see Figure 2). To modify the signal of one sphere to sound like another, the spectrum was scaled according to the ratio between its own F1 and that of the sphere of desired size. That way, each signal was scaled to its neighbours, for example S to M and labelled as MS (see Figure 4). Other than simply changing PSR, signals that are manipulated in this way do not have confusing harmonics. In experiment 2, signals were only scaled to their neighbouring sizes so as to avoid any distortion due to large scaling factors, i.e. Medium was scaled to both Large and Small, but X.Large was only scaled to Large.

The stimuli pairs used in this experiment could have either different F1 values (test 1) or the same F1 value (test 2). The rationale for choosing pairs with different but manipulated F1 was to find out if participants could be fooled by an F1 different to its original value; the rationale for choosing pairs with the same F1 is to find out if participants have to resort to guessing because of the spectral similarities.

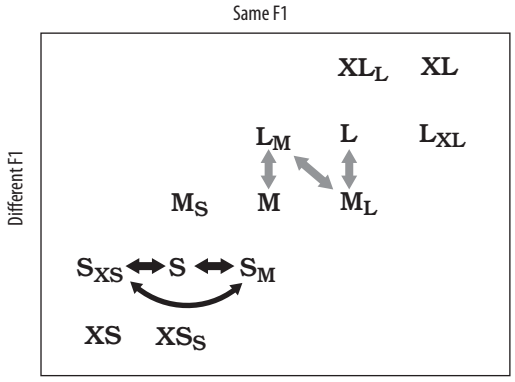

Figure 4. Schematic matrix indicating the signals used in the two tests of experiment 2. According to their physical size, signals on the main diagonal are labelled (XL, L, M, S, XS). Off diagonal signals are labelled as their presumed perceived sizes after scaling: e.g. LM, indicates that the signal L had been created by scaling the signal M to have the same F1 as L. Signals in each row have the same F1, and signals in each columns have different F1 values. The arrows indicate the examples outlined in the text

Stimuli pairs are illustrated in Figure 4. The following types of pairs were presented to the participants: pairs with different F1 (as part of test 1) are illustrated as hollow lines in figure 4. Such pairs could be an unscaled signal compared to a version of itself that is scaled for either a lower F1 (such as M *vs.* LM) or a higher F1 (L *vs.* ML). It can also be a signal scaled for a lower F1 compared to a signal with a higher F1 (LM *vs.* ML). Pairs with the same F1 values (test 2) are illustrated with solid arrows in Figure 4. These pairs consist of a signal compared with either a signal scaled upwards to match its F1 value (such as S *vs.* SXS) or a signal scaled downwards to match its F1 value (S *vs.* SM). Also a pair could consist of two signals scaled to have the same F1 (SXS *vs.* SM). In total there were 23 pairs in this experiment.

Signals were scaled in MATLAB using the ratios between the F1s for each sphere size as a scaling factor, and the new signals were created by sample-by-sample interpolation. The increase factors ranged from 1.08–1.14, and the decrease factors ranged from 0.82–0.92. The resultant signals were the same length and had the same RMS energy as the originals. The higher resonances F2 and F3 also aligned well. As in experiment 1, the participants were asked the question, "Which sound comes from the bigger object?" using the 2AFC method.

We hypothesised for the first test (where stimuli pairs have different F1) that participants would choose the signals with the lower F1 as the larger signal, regardless of whether or not it was a scaled signal. Accordingly, we hypothesised that in test 2 (where stimuli pairs have the same F1 and thus assuming the scaling method sufficiently removed the spectral cue); participants would not be able to reliably decide and had to guess.

Results show that in test 1 (different F1) participants consistently and significantly chose the signal with the lower

Table 3. The filter types applied to each signal in experiment 3. (c/o = cut-off frequency of filter). Specific values of each F1 and F2 are shown in Table 4. Position of the other resonances can be seen in Figure 2

	Abbreviation	Type of signals in the set		Resonances present
Unfiltered	Unf	Averaged unfiltered signals		F0, F1, F2, F3...
Band-stop Filtered	BS no F1	No F1	Between 5% below and above F1	F0, –, F2, F3...
Low-pass filtered	LP with F1	With F1	c/o 5% above F1	F0
	LP no F1	No F1	c/o 5% below F1	F0, F1
High-pass Filtered	HP with F1	With F1	c/o 5% below F1	F1, F2, F3...
	HP no F1	No F1	c/o 5% above F1	F2, F3...

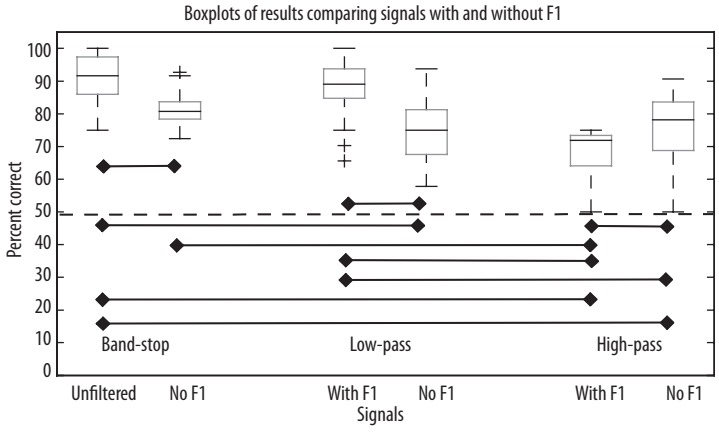

Figure 5. Results from experiment 3. Presence and absence of F1 affect the size discrimination. The horizontal brackets in the figure indicate significant differences between the means. The dashed line represents chance level (50%). All results are significantly higher than chance

F1 as the larger sound whether or not it was scaled (84.5% *vs.* 15.8%, SD ±12.0). This result confirms our hypothesis. We conclude that the spectral cue given mainly by F1 is important for size discrimination. In test 2 (where pairs had the same F1), contrary to expectation, participants chose the 'wrong' answer consistently: signals that were scaled up from smaller were chosen over unscaled (65%, SD ±9.0); unscaled were chosen over those scaled down from bigger (71%, SD ±9.4) and those scaled from smaller was chosen over those scaled from bigger (65.8%, SD ±4.2). In all three cases, results are significantly higher than the hypothesised chance level of 50%.We conclude from this result that F1 is not the only cue that can be used for size discrimination. Visual inspection of the spectra shows that F2 and F3 were well aligned and that the frequency region of interest is probably above 8 kHz.

In order to explore in detail which frequency range is important for size discrimination, differently filtered signals were investigated in experiment 3.

Experiment 3: Size discrimination of filtered signals

Experiment 3 was carried out to investigate the importance of different frequency regions for size discrimination. Here, the same signals that were used in experiment 1 were filtered using low-pass (LP), high-pass (HP) or band-stop filters (BS) (4th order Butterworth), carefully choosing cut-off frequencies to either keep or remove specific parts of the spectrum containing resonances. Based on the results of experiment 2 we know that F1 is important, but it is not the only frequency cue that can be used for size discrimination. In order to investigate this further, here we created signals that were characterised by muting certain spectral regions and tested if size discrimination was still possible. We hypothesised that F1 is the most important cue, but that in the absence of F1 other cues can be used. In order to investigate this we designed signals that were identical apart from F1 and we expected listeners to perform better when F1 was present. Secondly, we hypothesised that if F1 is absent, high frequency information above 8 kHz can be used for size discrimination.

Six kinds of signals were generated from each original sphere size: 'low pass filtered' isolated the effect of F1 in the absence of high frequencies; 'high pass filtered' isolated the effect of F1 in the absence of F0; 'band-stop filtered' investigated the effect of removing F1 entirely. Table 3 shows the different types of filters and the cut-off/band-stop frequencies used.

Results of experiment 3 are shown in Figure 5. Participants performed significantly above chance in all 6 conditions.

All filtered signals produced significantly lower mean scores than unfiltered (Unf) signals (90.8%, SD ±7.9) apart from LPwithF1 (87.9%, SD ±9.4). Removing F1 alone significantly decreased performance from Unf to BSnoF1 (81.8%, SD ±5.4). This confirms that F1 is indeed the most important frequency cue when estimating

Table 4. Left: frequencies of the first 3 resonances of all 5 spheres. Right: differences between F1s, F2s and F3s. The rightmost column shows the average differences

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

Table 5. Properties of objects used to test tAIM. The cones, eggs and hearts were made from the same material as the polystyrene spheres. FIMO spheres are much heavier but are in the same range of size

Size/Shape	Properties	Cone	Egg	Heart	FIMO sphere
Large	Weight (g)	75.4	7.7	8.4	550
	Maximum geometric dimension (cm)	31.3	10	11.2	11
Medium	Weight (g)	26.3	5.2	4.6	520
	Max dimension (cm)	26.1	8.4	8.3	10
Small	Weight (g)	13.5	2.3	2.0	515
	Max dimension (cm)	20.4	5.6	5.5	9.5

the size of objects in realistic environments. However, while in the cases of 'band stop' and 'low pass' removing F1 decreases performance, removing F1 in the 'high pass' condition surprisingly increases performance significantly (HPwithF167.8%, SD ±7.7; HPnoF1 75.8%, SD ±11.5). This suggests that high frequencies provided more information than we assumed based for example on the range of natural speech sounds.

Analysis of the differences provides a further cue to explain this improvement in the high pass filtered sounds: Table 4 shows that the average differences between F3 values (840 Hz) is larger than the differences between F2 values (775 Hz), which are larger than the differences between F1s (581.5 Hz). It could be that the lower F1 resonance masks the higher ones due to upward spread of masking. So although the task was easy in both cases, the removal of F1 facilitated an improvement in scores due to larger differences between F2.

Transient Auditory Image Model – tAIM

In order to explain the observed results we propose here a model to analyse and compare transient signals, and to carry out automatic size discrimination. The model is based on the Auditory Image Model (AIM) and was originally described by [4]. The version used as the basis of our model is aim-mat, the MATLAB implementation of AIM [5].

The original AIM analyses periodic vowel signals by way of spectral analysis and pattern stabilisation. AIM produces 'stabilised auditory images' that captures the fine structure of repeated sounds. In order to normalise for size [2] added the Mellin transform of the model to produce the same outputs for speakers of different size. Due to the strobing mechanism that is used in the image stabilisation process, AIM only works for stimuli that contain at least 2 periods and produces no outputs for transient signals that are not periodic.

Table 5. Comparison of the modules of AIM and the modules of the newly created tAIM

aim-mat	tAIM
Pre-cochlear Processing	–
Gammatone filterbank 100–6400 Hz	dcGC filterbank 100-10k Hz
Neural Encoding	–
Strobed Temporal Integration	Alignment of Maxima
Stabilised Auditory Image	Simplified Auditory Image
Pattern Normalisation	Pattern Normalisation
–	Mellin Phase analysis for Size comparison

Here we present a modification of AIM for the analysis of non-periodic, transient signals called transient Auditory Image Model (tAIM) that also performs pattern analysis for size comparison.

The modules of tAIM were simplified compared to aim-mat in order to reduce the computational expense of signal analysis, but also to minimise the processing and filtering while retaining spectral analysis and producing a comparable (however not stabilised) auditory image. Table 5 compares the modules of aim-mat and tAIM.

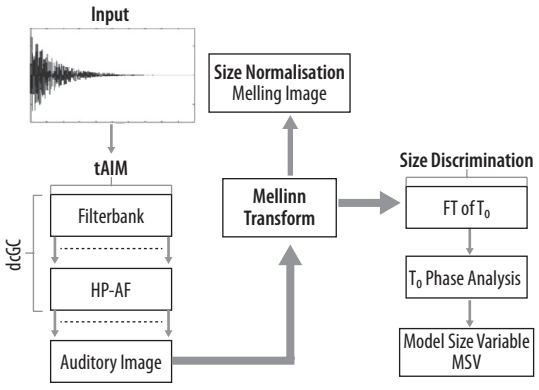

Figure 6. Flow-chart showing the processing blocks of tAIM and the automatic size discrimination modules. For a detailed description see text. FT = Fourier Transform

Figure 6 shows the flow diagram of tAIM. The parameters in the original aim-mat limited spectral analysis of the vowels to between 100 and 6400 Hz in order to simulate pre-cochlear processing in the important speech range. However, we demonstrated above that the frequency range above 8 kHz is important for size discrimination. Thus, in tAIM we use a filterbank of 100 channels spaced from 100 Hz to 10 kHz on an ERB scale – this creates an image showing the motion of the basilar membrane (BMM). In contrast to aim-mat, a dynamic and compressive gammachirp auditory filter (dcGC) is used here as a simulation of the basilar membrane. The dcGC 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 [11]. We chose this filter for tAIM because compression is simulated more realistically in the dcGC filterbank compared to the passive gammatone filter, and also removing the half-wave rectification and low-pass filtering serves to retain as much information about the signal as possible for the eventual automatic size discrimination process.

The strobing process included in aim-mat was not required here with a non-periodic stimulus. Instead, a method of alignment is used which involves arranging the points of maximum amplitude in each frequency channel so that they occur at the same time interval point, T0. The result is a transient version of the stabilised auditory image called here the transient Auditory Image (tAI).

The next step in tAIM is to carry out a Mellin transforms described in detail by Irino & Patterson [6]. Based on the results of this transformation either a size normalisation or size discrimination can be carried out. This is illustrated in Figure 6 as a branch of the Mellin transform box and is shown in order to highlight how tAIM differs from AIM. While AIM uses the Mellin 'spatial frequency' in order to do size normalization, tAIM concentrates on the complementary Mellin 'spatial phase' information that contains the size information. We have called this the 'model size variable' (MSV) and will investigate if it correlates with the physical size of objects.

In order extract the relevant size information a Fourier transform (FT) was performed on the T0 time column. This is done at this specific point because T0 is the point of highest energy in each channel and thus contains the most useful information regarding size and shape of the analysed object. The frequency part of the FT was discarded and the phase part (which we call Mellin phase) contains the relevant size information. The output of the model is the Mellin phases and this can be used for automated size discrimination when compared to the Mellin phase of another objects of the same shape and material. The results of the psychophysical experiments above indicate that F1 is the most important cue. This could have been simulated in the model by peak-picking and comparing the respective resonances. However, since the results above also demonstrate that size discrimination is possible without F1 based on a combination of higher frequencies, a different approach was chosen that reflects the overall phase structure rather than single resonances. Therefore we chose to calculate the average rate of change of the Mellin phase. Smaller values indicate larger objects because smaller objects produce a more fluctuating Mellin phase due to more and higher resonance frequencies. The calculation of an average is also more robust than pinpointing specific resonance frequencies.

Testing tAIM

The model described above was used to process a variety of stimuli in order to test its validity. Signals ranged from simple synthesised sounds like multiple damped sinusoidal sounds and vowels to recordings of real sounds. Recorded sounds were the spheres mentioned above, spheres made from modelling clay and other shapes made from polystyrene. Mellin phases of all objects were calculated by processing the wave files through the model. The results of the output of tAIM for the synthesised sounds demonstrated that it works as expected (data not shown) for the idealized synthesised sounds. In the following we therefore show the more interesting results of real recorded signals of the polystyrene objects (spheres and other shapes).

The calculated MSVs of the recorded polystyrene spheres are shown in Figure 7. The top panel shows MSVs from original signals, the bottom panel shows the results from signals when F0 is removed. The data demonstrate that MSVs of spheres are ordered in the same way as the size of the original spheres. Pearson's correlation coefficient is 0.86 for the full signals and 0.98 for the filtered signals.

The differences between paired MSVs are not constant. The differences XS *vs.* S and M *vs.* L are bigger than S *vs.* M and L *vs.* XL. These differences correspond well with the observed psychophysical results of direct comparison shown in Figure 3: the smaller the difference in MSV, the larger the number of errors in comparison.

The bottom panel shows the results of high pass filtered sounds where F0 is removed. We expect the model to perform well for these signals because the MSV in these cases is mainly due to the position of the F1 resonances. As visible in Figure 2, the F0 resonances are all much broader than the higher resonances and roughly cover the same frequency range. The F0 is physically mostly the product

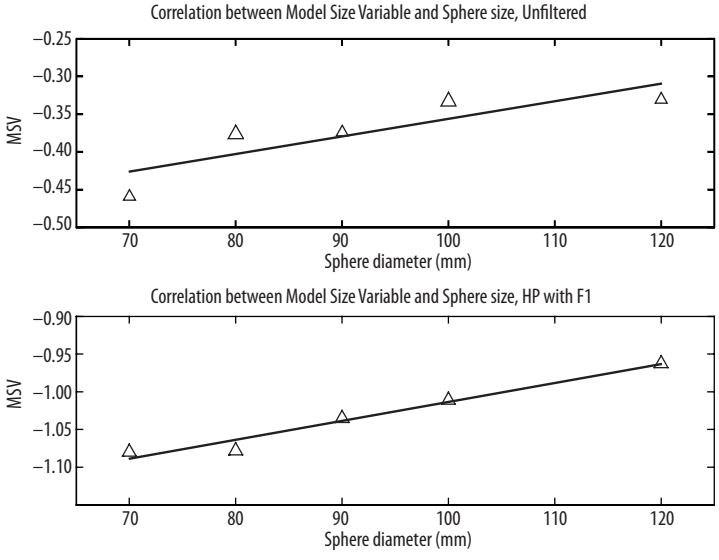

Figure 7. 'Model size variable' of polystyrene spheres measured by tAIM. Top panel: unfiltered signals, bottom panel: filtered signals with F0 removed

of the broadband response of the impact of the striking ball, not a resonance of the sphere, and it thus carries little size information. However, because of its high energy it dominates the calculation of the MSV in the model. By removing it, the model concentrates more on the energies around F1 and higher resonances that contain more size information. This can be seen in Figure 7 in the bottom panel: when removing F0 the correlation is improved. Although the differences in some cases are small, bigger objects always have larger MSV.

To test tAIM further, we also explored if it can predict relative sizes between other objects. We chose a range of polystyrene shapes that were readily available from a craft shop. Shapes tested were cones, eggs and hearts and each came in three different sizes. We also tested a range of spheres constructed from FIMO polymer modelling clay. Properties of all objects are shown in Table 5.

It is impossible to describe the relative geometric size of different shaped objects in one number. To compare sizes relative between object families, we only display the 'maximum dimension' in Table 5 which is the longest physical length of every respective object: for the cone and the egg this is the height, for the heart this is the difference between bottom and highest point. For each family of shapes, the relationship between dimensions is a perfect descriptor of their scale, but this description is only a rough approximation between families of different shapes. The resonances of each object are a consequence of the three dimensional geometric properties. It is impossible to capture these in one number and therefore we can only represent one aspect of the size parameter at a time. We chose the maximum geometric dimension for simplicity. Any other descriptor, like volume, would also have been possible and would have yielded the same qualitative results.

Sounds were recorded in the same manner as the spheres. Prior to analysis by tAIM, the signals were band-pass filtered with cut-offs at 100 Hz and 16 kHz, and normalised to RMS =1. Averages of up to three recordings for each size were calculated. The FIMO sphere sounds were

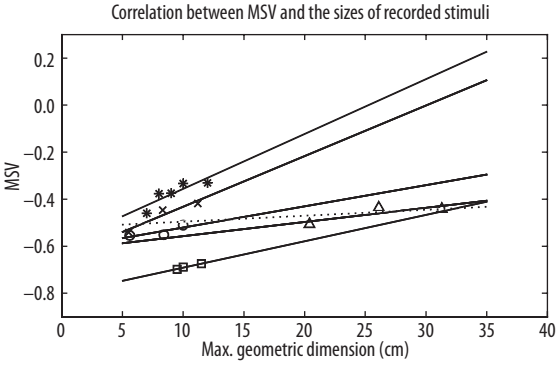

Figure 8. Correlation between maximum geometric dimension of the recorded objects and the Model Size Variable. The correlation between all shapes and sizes is indicated with a dashed line. The symbols represent: -Δ- Cone, r=0.84; -O- Egg, r=0.81; -X- Heart, r=0.95; -□- Fimo sphere, r=0.98; -✶- Polystyrene sphere, r=0.86; ---- All shapes, r=0.155

recorded differently, because the impact with a small metal ball would not have been loud enough. FIMO spheres were recorded 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.

Figure 8 shows the MSVs obtained from tAIM for all objects, including the polystyrene spheres. All object families show the same behaviour: bigger objects produce higher MSVs. Correlations within each family are all positive and significant. The only signals that provided misleading results were the medium and large cones; this can probably be explained by the quality of the recordings of these objects. The correlation between all families and sizes is weak: r=0.155 and not significant, indicating that the model is

Table 6. Ratios between the first sphere resonances after F0

Ratio	X.Large	Large	Medium	Small	X.Small
F2/F1	1.49	1.47	1.44	1.35	1.43
F3/F2	1.26	1.29	1.29	1.32	1.19

less useful for relative size discrimination of objects of different shapes. This corresponds to psychophysical results which will be analysed in a subsequent paper.

Discussion

The psychophysical experiments presented in this study were carried out to identify the importance of the spectral cue in size discrimination of transient signals. We established that the frequencies of the main resonances carry the most important size information. In experiment 1 it was demonstrated that listeners can discriminate the size of transient signals from polystyrene spheres to a high standard without training. These results confirmed the validity of the used stimuli. In experiments 2 and 3 signals were spectrally scaled in order to intentionally deceive participants to test the importance of F1 Results show that the signal with the lowest F1 was consistently chosen as the larger signal, regardless of its size before scaling. This indicates that spectral information plays a major role as a cue for size discrimination of transient signals. This demonstrates, to our knowledge for the first time, that there is useful information available for size discrimination above 8 kHz. We conclude that listeners base their size judgement mainly on F1 if available and on higher resonances when it is not available.

F1 is the most important psychophysical cue for size discrimination and it is also very prominent for a topological reason: geometrically, F1 is related to the diameter of the sphere and the shear waves across its centre. This can be shown by calculating the first mode of resonance as the relationship of speed of sound in polystyrene (~813 m/s) and twice the sphere diameter. The values are within 6% of each other. Other resonances are a result of higher modes and surface waves. F0 is probably the result of the impact between the small metal ball and the polystyrene spheres. It is very broad in frequency compared to the other resonances and its maximum is flat. Therefore F0 peak frequency is not a good predictor of size. Furthermore, the results of experiment 3 show that F0 alone is not as good a cue for size discrimination as F1. Further studies could include a physical simulation in order to investigate the physical sources of the resonances. This would shed further light on which physical aspects of objects of different shape and size affect the spectrum and the perception.

The results of our experiments are in line with those of Houben, Kohlrausch & Hermes [9] and Grassi [10] who also showed that participants were able to discriminate size of spheres (Grassi – 4 spheres, 16 participants; Houben – 7 spheres, 8 participants). However, their results might have been affected by cues other than the size of the spheres because the sounds came from single presentations of rolling or bouncing balls, and also from the objects onto which the spheres fell. In our experiments the sounds were controlled in order to ensure that the only available cues were from the spheres, thus increasing the validity of the results. We also only selected clear recordings of impact sounds and averaged over 300 impacts. As shown in experiment 1, participants were significantly better at discriminating the averaged signals. We thus believe that our results are more repeatable because of the more reliable representations of the sphere sounds. However, this came at the cost of ecologic validity.

Despite this the average score across all pairs of averaged unfiltered and unscaled signals (in experiment 1) was not perfect at 90.1%. Participants thus demonstrated a very good, but not perfect ability to discriminate between the sphere sizes. However, participants were untrained; they had no previous exposure to any of the stimuli and there was no feedback given at any stage. This demonstrates that the task of extracting size from a single impulse sound is a natural human ability for most people. However, we excluded 11 out of 31 participants because they were not consistent in their answers (note that they could have been consistently wrong, but they had to be consistent). It would be interesting in future studies to investigate the reasons why these people are unable to do the task, (if they 'size-deaf') and to what degree this ability is correlated with pitch discrimination ability.

The tAIM model presented in this study is an alteration of the well-known Auditory Image Model AIM [12]. AIM was originally developed as a model to predict the pitch of periodic sounds, and was later extended to normalise for speaker size using the Mellin transform. The transform works on the basis of the approximately constant relationship between formants for vowel identification regardless of speaker size. For example the relationships between the first three formants of the vowel "A" are very similar for a man, a woman or a child even though the absolute formant frequencies are very different [6]. We show here that the same is true for spheres of different size. The similarities in the shape of the spectral envelopes can be seen in Figure 2, and similar to the formants of vowels, Table 6 shows that the resonances of the spheres have a near constant relationship: all ratios are within 10%. However, the vowel sounds consist of many repetitions, and AIM makes use of the repeated nature of the signals by averaging. Here we show that the single transient impulse signal of a struck object is similar in its spectral pattern to a vowel sound but much shorter. It has been suggested that single impulse communication sounds evolved very early on in fish and were useful for simple communication [1]. Later animals presumably extended this by repeating the same message several times in short succession thus increasing the SNR and giving the listeners multiple looks of the individual waveforms. Voiced speech sounds can be described by a repetition of identical single impulse sounds that are only 5–10 ms long. Nevertheless, the cues used for identifying

both seem to be the same. The results of our experiments confirm that single impulse sounds carry useful information about size and potentially about the shape of the object. The role of different shapes on perception of impulse sounds will be investigated in a future paper.

In contrast to AIM which cannot process single non-periodic sounds, tAIM was developed to process transient sounds such as by single pulse resonances, but also to retain as much spectral information about the signal as possible, especially in the higher frequencies that are usually ignored in AIM. A simplified alignment method was developed for tAIM that replaced the strobed temporal integration and image stabilisation in AIM. Both tAIM and AIM produce as output an auditory image that is suitably stable to perform the Mellin transform. The second major difference between the models is the use of the Mellin transform: the transform separates the size information ('Mellin phase') from the shape information ('Mellin frequency'). AIM uses the frequency information to create size normalised images ('Mellin images'), whereas tAIM carries forward the phase information for size analysis. It is worth noting that despite the fact that our signals are not periodic tAIM can also be used for successful size normalisation and to create Mellin images.

We conclude that the output of our model calculation, the Mellin size variable (MSV: calculated by the group delay of the Mellin phase), is a good descriptor of the perception of sizes of objects within the same shape family. To our knowledge, this is the first model that has been developed to calculate the sizes of objects from a single impulse sound.

We also tested a different method of extracting size information based on the spectral centroid frequency (SCF) of the auditory image. This was motivated by Houben et al. [9], who suggested that SCF of the spectrum provided a cue for size discrimination. In the second test of experiment 2, participants were not confused when the F1 values of different sizes were matched. We conclude that despite F1s being scaled to be the same (and F2 and F3 also being well aligned) other information in the scaled signals was used for size discrimination. Since F1-F3 values are all in the lower frequency regions (below 8 kHz in most cases) this implies that this additional information must be in higher frequencies regions above 8 kHz. Simple calculation of SCF for the sounds used in this experiment shows that SCF is indeed correlated with size perception; however, a closer inspection raises doubts. First, discrimination was better for the HPnoF1 signals, despite their SCFs being higher than the HPwithF1 signals (8.5–11.5 kHz, and 6.1–9.8 kHz respectively). If SCFs are the most important cue then we would expect that discrimination is best at low SCFs. Since this is evidently not true in the case of the HPF signals, this suggests that the differences between resonances, rather than the absolute SCF values, are the important cue for discrimination. Secondly, there is not always a correlation between SCF and size. For example in some cases of the BS and LPwithF1 signals, SCF increases with increasing size. We conclude that SCF is not a good descriptor of size perception in all cases and the presented model therefore uses the relative position of the resonances.

tAIM is not a perfect model that works in all circumstances. As it is, the model is limited to comparing two objects with the same shape. Humans have to a degree the ability to estimate the absolute size of objects of different shapes [8]. The model does this to a small degree (see Figure 8), but obviously there is other information that humans also use for object identification including its shape and material. Within object families tAIM did a good job for all tested objects, but it differs from the psychophysical results in one important point: the relevance of F0. While psychophysically F0 contributed constructively to correct size discrimination; in the model it dominates over F1 energetically and thus reduces the accuracy of results (see Figure 7). Nonetheless, Figure 8 shows the MSVs from signals with F0 remaining and the correlation between MSV and size is strong for all objects within shape family.

Conclusions

The presented experiments demonstrate the importance of the spectral cue in the auditory size discrimination of transient signals. The cues for size discrimination are similar in transient sounds and in speech sounds; single impulse responses can therefore be thought of as a simple model of voiced speech. A mathematical model for the analysis of transient signals based on the Mellin transform (tAIM) can predict which of two signals from objects of the same shape is bigger.

References:

1. Patterson RD, Smith DRR, Dinther R, Walters TC et al. Size Information in the Production and Perception of Communication Sounds. In: Yost WA, Popper NA, Fay PR (eds.). Auditory Perception of Sound Sources. New York, Springer Science+Business Media, 2008l; 43–75.
2. Irino T, Patterson RD, Kawahara H. Speech segregation using an auditory vocoder with event-synchronous enhancements. In: IEEE Transactions on Audio, Speech and Language Processing. [Online]. November 2002 Wakayama Univ, Fac Syst Engn, Wakayama 6408510, Japan Univ Cambridge, Dept Physiol Dev & Neuroci, Ctr Neural Basis Hearing, Cambridge CB2 3EG, England. 2002; 2212–21.
3. Ives DT, Smith DRR, Patterson RD. Discrimination of speaker size from syllable phrases. Journal of the Acoustical Society of America. [Online], 2005; 118(6): 3816–22
4. Patterson RD, Robinson KEN, Holdsworth J, McKeown D, et al. Complex sounds and auditory images. In: K. Horner Y Cazals, L. Demany (eds.). Auditory physiology and perception, Proc. 9th International Symposium on Hearing. 1992; Oxford, Pergamon.
5. Bleeck S, Ives T, Patterson RD, Ives DT. Aim-mat: The Auditory Image Model in MATLAB. Acta Acustica, 2004; 90(4): 781–87.

6. Irino T, Patterson RD. Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform. Speech Communication [Online], 2002; 36(3–4): 181–203.
7. Irino T. Noise suppression using a time-varying, analysis/synthesis gamma chirp filterbank. 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258). [Online], 1999; 1: 97–100.
8. Carello C, Anderson KL, Kunkler-Peck AJ. Perception of Object Length by Sound. Psychological Science [Online], 1998; 9(3): 211–14.
9. Houben MMJ, Kohlrausch A, Hermes DJ. Perception of the size and speed of rolling balls by sound. Speech Communication [Online], 2004; 43(4): 331–45.
10. Grassi M. Recognising the size of objects from sounds with manipulated acoustical parameters. Fechner Day 2002: Proceedings of the International … [Online], 2002
11. Irino T, Patterson RD. A Dynamic Compressive Gammachirp Auditory Filterbank. IEEE transactions on audio, speech, and language processing. [Online], 2006; 14(6): 2222–32.
12. Patterson R, Holdsworth J. A functional model of neural activity patterns and auditory images. Advances in speech, hearing and … [Online], 1996; 3547–58.