MEASUREMENTS OF PERCEPTUAL QUALITY OF CONTACT SOUND MODELS

K. van den Doel and D. K. Pai

Department of Computer Science
University of British Columbia
Vancouver, Canada

T. Adam, L. Kortchmar, and K. Pichora-Fuller

School of Audiology and Speech Sciences and
Institute for Hearing Accessibility Research
University of British Columbia
Vancouver, Canada

ABSTRACT

We describe and test methods to construct modal resonance models for solid objects, suitable for the real-time synthesis of sound-effects in simulation and animation. Measurements on typical everyday objects such as a metal vase or a bowl result in several hundred modes, of which only a small fraction is perceptually relevant. We have proposed several heuristics, inspired by psycho acoustical data, to select the modes by perceptual relevance and to order them so that one can increase the quality by adding more modes, at the price of additional computational complexity (progressive synthesis). The resulting synthetic sounds are tested on human subjects in order to determine the quality of the sounds relative to the target sound which they are designed to approximate. The resulting data is used to verify and tune the mode selection methodologies, and to increase our understanding of what determines the subjective quality of a synthetic sound effect.

1. INTRODUCTION

Sound effects, or Foley sounds, are an important component of immersive artificial environments such as simulators or video games. These sound effects are often added by ad hoc methods by very skillful "Foley artists" who use remarkable ingenuity in their creative processes.

A system for the automatic generation of a large and important subset of Foley sounds, namely the sounds made by the contact interactions between solid objects, has been developed [1, 2, 3, 4]. Other work on computing sound effects includes [5, 6, 7, 8, 9]. In our system, a solid object is represented by a modal vibration model which consists of a bank of oscillators driven by the contact forces. A modal resonator bank can be computed very efficiently with an $O(N)$ algorithm [10] for a model of N modes, making real-time synthesis of reasonably complex sonic scenes feasible on desktop computers.

In this article we describe some measurements of the perceived quality of these contact sounds. The contact sound models are "progressive synthesis" models, i.e., one can make a tradeoff between accuracy and computational complexity. Therefore it is important to have a quality metric to quantify the quality of the synthetic sounds.

The modal models were obtained by parameter extraction from recorded impulse responses of real objects. The power spectrum is computed, see Figure 1 for an example, and the peaks are identified as candidates for modes. For each candidate we then perform a phase reconstruction of the complex windowed Fourier transform which is fitted with multiple modes. This "phase unwrapping" has been used before to produce accurate frequency estimates [11]. Our extension also reconstructs the dampings and gains of the modes with high accuracy and is able to identify very closely spaced frequencies. Perceptually, this is important because these modes cause beating, and occur frequently in man-made objects. [1]

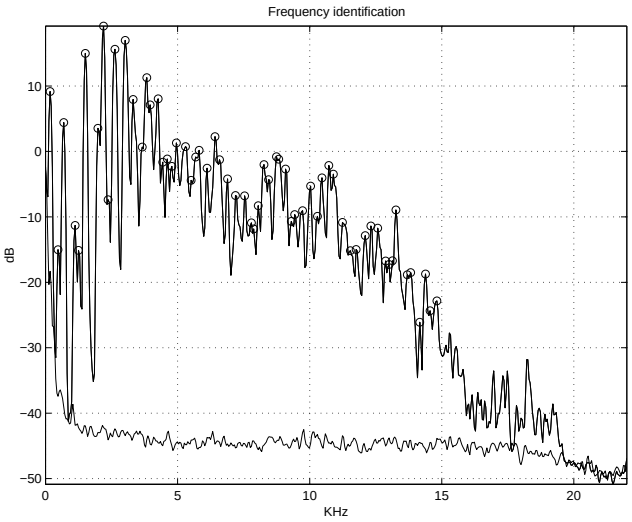

Figure 1: Power spectrum and peaks of the recorded impulse response of a metal vase. The bottom curve is the noise level.

2. PROGRESSIVE MODAL MODELS

For typical objects hundreds or even thousands of modes are identified. Many of these modes do not contribute to the sound model because they are inaudible. These spurious modes may exist as actual modes but be too weak to be heard, or they may be errors caused by noise in the data.

One could simply use all the modes for the real-time synthesis of the contact sounds, but this would waste a lot of CPU cycles. On a typical desktop system such as a 1 Ghz Pentium III system, about 1000 modes can be synthesized in real-time at a CD quality sampling rate of $44100Hz$, using JASS [12, 13].

[1] Geometrical symmetries often result in mode degeneracy, i.e., multiple modes with the same frequency. Small deviations from symmetry then cause those degenerate frequencies to split into close multiplets.

For the metal vase whose power spectrum is depicted in Figure 1 we found 179 modes. By laborious trial and error it seems that only 10-15 of the modes are important perceptually. The question is how to select the appropriate modes automatically.

Let us state the problem formally.

The modal model $\mathcal{M} = \{\boldsymbol{F}, \boldsymbol{D}, \boldsymbol{A}\}$, consists of three vectors of length M: the modal frequencies $\boldsymbol{F}$ in Hertz, the decay rates $\boldsymbol{D}$ in s^{-1}, and the gains $\boldsymbol{A}$ of the modes. Our task is now to select the N perceptually relevant modes (or, equivalently, eliminate the inaudible modes) and sort them in order of increasing importance, such that the sequence of impulse responses for $K = 1, \ldots, N$

$$y_K(t) = \sum_{n=1}^{K} A_n e^{-D_n t} \sin(2\pi F_n t) \qquad (1)$$

is optimal, i.e., converges to the most accurate approximation of the original as fast as possible, so that for any other ordering $\mathcal{M}'$ of the modes, $y_K(t)'$ is of lower "quality" than $y_K(t)$. We also want to know the value of N, i.e., the optimal model.

Modal models can have a large number of modes, from several hundred to several thousand, and for progressive modal synthesis we would like to order the modes by perceptual importance. A subset of M modes can be chosen in $2^M - 1$ possible ways, making an exhaustive evaluation of all possible sounds impossible for most practical values of M.

We have developed several heuristics for ordering the modes, and eliminating the inaudible modes, but their effectiveness has not been measured. These range from a simple technique (which we call the "naive " method) which orders the modes based on the gain of each mode, A_i, [3] to more sophisticated methods based on perceptual criteria which account for masking.

To approximate inter-modal masking effects by the human ear we consider masking of narrowband noises with frequencies F_n and power A_n^2/D_n. This is the power of the mode integrated over time, and also its excitation under a stationary excitation with a flat frequency spectrum. The masking effects of overlapping spectral sources can be assumed to add linearly [14] in a first approximation, though non-linear addition effects have been used to improve perceptual audio coders [15].

We use a masking curve consisting of two straight lines on the Bark scale, see Figure 2. This has been used before in perceptual audio coding [15]. This spreading function is parameterized by a threshold level a_v, which determines the overall threshold level for masking, and two slopes $s_l = 25dB/\mathrm{Bark}$ and $s_u = (22dB - L/5)/\mathrm{Bark}$. The upper slope, s_u, is dependent on the level of the mode L in dB, as louder sounds mask more higher frequencies. The threshold parameter a_v should be chosen a small as possible, to eliminate as many inaudible modes as possible. An experimental determination of a_v is described below. We eliminate the inaudible modes by constructing the upper envelope of all the masking curves for all modes and removing all modes that fall below it.

In Figure 3 we show the masking method being applied to the measured modes of a metal vase. We show the level of each mode normalized to a reference playback level of 80dB (level immediately after impact) and the masking threshold curve. All modes below the threshold curve are considered inaudible by the algorithm.

The surviving modes are then sorted according to "perceptual importance". However, it is not clear what this should be. One method is to sort the surviving modes by their A^2/D value, which we call the "energy" based method. Another reasonable approach is to sort them by how much louder they are than the masking threshold. We call this the "loudness" method.

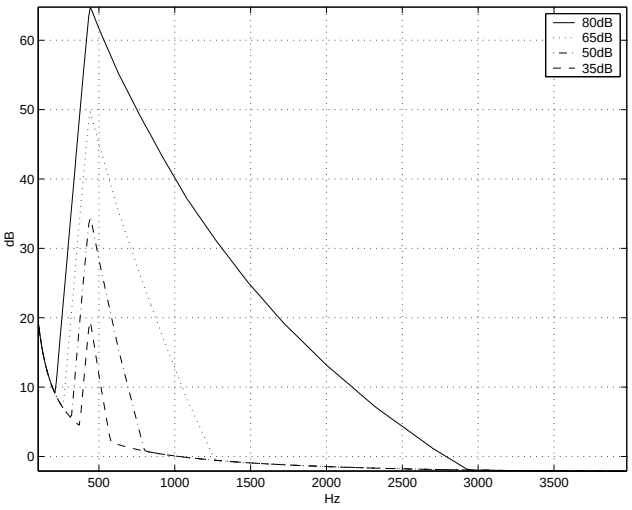

Figure 2: Masking curves for a source at 440 Hz for various levels. Masking threshold $a_v = 15dB$.

In this paper we describe psychophysical experiments aimed at understanding how perceptual quality is affected by mode selection. This will then enable us to select the best mode selection method, and quantify the trade-off between quality and computational complexity. We have compared the "naive" and the "energy" based methods and plan to include measurements of the "loudness" method in future studies.

3. EXPERIMENTAL PROCEDURE

Summary

We have designed a set of three interrelated experiments to answer the following questions: 1) What value(s) of the masking threshold parameter a_v should be used? 2) For a given mode-sorting method, which of the synthetic impulse responses (each with a specific number of modes) can be distinguished by listeners? 3) For a given method, how do the synthetic sounds relate to the target sound and to each other?

We have focused on two target sounds, made by hitting a metal vase and a ceramic bowl. We have reconstructed modal models using the recorded impulse responses of these objects with the techniques described in Section 2. For each object, the threshold parameter a_v was determined by constructing the maximal modal model for a value of $a_v = 20dB$, and decreasing the value until a difference was heard by the subject (Experiment 1). Using these results we chose a value of a_v and generated synthetic impulse responses for every possible number of modes using the "naive" method (no masking analysis, ordering the modes by gain) and the "energy" based method as described in Section 2. The goal of Experiment 2 was to see how many of these sounds are distinguishable by the subjects. Using data from Experiment 2 we chose 4 sets of 11 synthetic impulse responses (two methods, two objects). In Experiment 3 we then measured the quality of the synthetic sounds compared to the target sound as well as the perceived

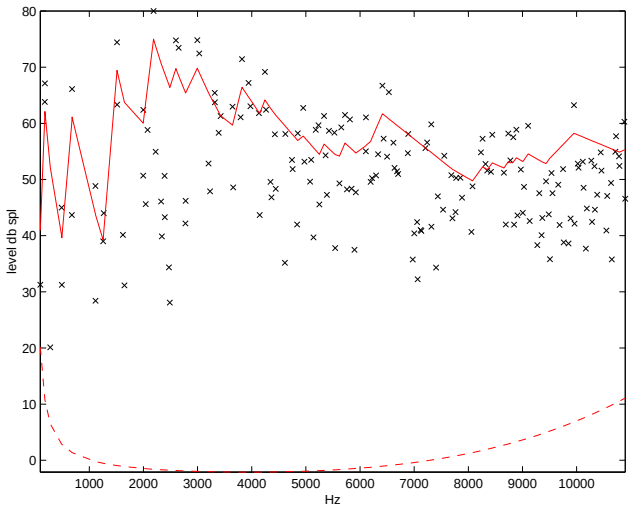

Figure 3: Of the 179 modes of a vase, only 45 modes survive using a masking method with $a_v = 5dB$ and a playback level of 70dB. We also show the absolute threshold of hearing.

differences between the synthetic sounds themselves.

Stimuli

The stimuli consisted of recordings of the impulse responses and two sets of synthesized impulse responses for each of the two original contact sounds, a metal vase and a ceramic bowl (target sounds). The impulse response was recorded by striking the object with a metal hammer and recording the resulting sound at a distance of 50cm. Each of the sets of synthesized sounds for each object was constructed using a specific mode ordering and elimination method as described in Section 2. The first method, which we call the "naive" method uses the raw measured modes sorted by gain. The second method, which we call the "energy" method uses a masking analysis to eliminate inaudible modes and orders the remaining modes by energy.

All stimuli were sampled at 16 bit resolution at a sampling rate of 22050 Hz. The level was equalized over the initial 5 ms of the signals. Stimuli were presented at the same level to all participants. All signal presentations were delivered binaurally over Sennheiser HD 265 linear headphones to subjects seated in an IAC double walled sound-attenuating booth. A Tucker-Davis Technology System II was used for the digital-to-analog conversion and control of stimulus parameters.

Experiment 1: Selecting the threshold parameter

For each object, eight subjects were presented with sequences of synthetic impulse response pairs (X, Y). Y was the impulse response using all modes, X the impulse response with mode elimination as determined by the threshold a_v. A starting value of $a_v = 20dB$ was used and this value was reduced by $1dB$ for every pair. The subjects were asked to select the first differentiable signal of the sequence. This procedure yields the optimal value of a_v, which is the smallest value such that no change is detected, within 1dB.

The participants were young adult volunteers: four women and four men (aged 24 to 32; mean 27; SD = 2.7) who had clinically normal hearing (pure-tone thresholds from 250 to 4000 Hz less than or equal to 25 dBHL in both ears).

Experiment 2: Finding the differentiable sounds

In Experiment 2, subjects were presented with four distinct pair sequences consisting of impulse responses with different numbers of modes. There were 179 pairs for the metal vase (no masking), 128 for the bowl (no masking), 100 pairs for the vase using the "energy" method, and 41 for the bowl with this method. The tokens were ordered according to the number of modes used to construct the sound. Beginning with the stimulus with the maximum number of modes as the reference, participants were asked to indicate whether or not the reference token was differentiable from the next token in the sequence (i.e., stimulus with one less mode).

The comparison token of each non-differentiable pair was discarded, and the reference token compared to the next consecutive stimulus (i.e., stimulus with two less modes). Using an adaptive procedure, the comparison token was incremented until it was discriminated from the reference token. The first comparison token that could be discriminated from the reference token became the new reference token. The comparisons continued in a similar fashion until the one mode stimulus was tested. This process of comparisons yielded a much smaller subset of differentiable tokens, dependent on the subject. The results for all subjects were used to select the set of 11 synthetic tokens for the next experiment.

Experiment 3: Perceived differences in sounds

In Experiment 3, sequences of 78 sound signals were generated, consisting of all possible pairings of the 11 signals selected in Experiment 2, as well as the real sound. For each object and method, such a sequence (4 in total) was presented to each participant. A magnitude estimation procedure was employed to obtain measures of dissimilarity between tokens. The sequences of token-pairs were presented in random order. Participants estimated the magnitude of dissimilarity between the tokens of each pair using a 10-point scale of dissimilarity: A value of 1 represented identical tokens and 10 represented the most dissimilar pair. To orient subjects to the scale, a standard pair comprised of the recorded sound and the single-mode token was initially presented as the most dissimilar pair. A magnitude estimation value of 10 was anchored to this token-pair. For each test trial following orientation, participants selected a number on the estimation scale whose ratio to the anchor of 10 represented the dissimilarity of the test pair, relative to that of the standard pair [16, 17]. For example, when a stimulus pair seemed to be half as dissimilar as the standard pair, a value of 5 would be assigned to that pair.

The participants in the last two experiments were young adult volunteers possessing clinically normal hearing (pure-tone thresholds from 250 to 4000 Hz less than or equal to 25 dBHL in both ears). The same eight listeners participated in both experiments, 6 women and 2 men (aged 21 to 32; mean 27; SD = 3.4).

4. RESULTS

In Table 1 we summarize the results of Experiment 1. The optimal value of a_v turned out to vary more than expected from subject

to subject, possibly reflecting individual differences in ability to distinguish different sounds, or it could be due to judgment biases. Informal experiments have also shown that the optimal value of a_v is sensitive to background noise. In noisy environments a lower value of a_v can be tolerated. Given the distribution of optimal values, for the subsequent experiments we took the median a_v (16) to generate the synthetic tokens for the "energy" method.

Subject	Vase	Bowl
1	5	13
2	16	19
3	15	17
4	17	20
5	14	20
6	18	13
7	15	16
8	3	19

Table 1: Optimal values for the threshold parameter a_v in dB for each subject, for each object.

The number of discriminated tokens which was measured in the second experiment varied widely from user to user, see Table 2. Based on this data, we selected 11 tokens for each of the

Subject	Vn	Ve	Bn	Be
1	64	79	3	3
2	27	13	7	2
3	25	9	3	3
4	12	11	10	4
5	79	25	29	15
6	59	27	26	8
7	78	59	59	15
8	12	13	10	4

Table 2: Number of tokens distinguished by each subject for each object (v = vase, b = bowl) using the two mode-selection methods (n = naive, e = energy).

two methods and each of the two objects by selecting the tokens that were most often selected as being different from the previous one (when ordered by number of modes). Ties were resolved by choosing tokens to be as diverse in mode number as possible. We believe this procedure to result in the most diverse sounding set of tokens for all subjects.

The perceived differences, as measured in the third experiment, contain an important subset, namely the differences between the target sound (the real sound) and various approximations to it with different numbers of modes. In Figure 4 we plot the perceived difference between the recorded impulse response and the resynthesized sound as a function of the number of modes of the synthetic sound. The difference scale ranges from 0 to 9, 0 being indistinguishable. It can be seen that the perceived difference with the real sound decreases faster for the method using masking and energy sorting, as expected. The apparent lack of convergence and monotonicity is probably due to errors in subject responses.

The data for all perceived differences was analyzed using multidimensional scaling methods. In Figures 5, 6, 7, and 8, we plot perceptual maps of the perceived differences between pairs. The data points are fitted to a plane in such a way to make the best correspondence between Euclidean distance on the plane and measured difference. The sum of the squares of the differences between geometric and measured distances was minimized and plotted using PERMAP 9.2. We have also generated maps using the city-block metric and by minimizing the sum of the absolute values of the differences with only very small differences in the maps. We have no a priori reason to believe the data should fit in two dimensions. The difference between sounds has most likely a large number of perceptual dimensions (pitch, roughness, timbre, duration, complexity, decay rate, etc.). The two dimensional plots nevertheless allow for some interesting observations. In all cases we can see a cluster of sounds around the real sound, which are all good approximations to the target. The synthetic sounds with very few modes also cluster together. These are very poor approximations to the real sound. Having 1, 2 or 3 modes does not seem to lead to any appreciable improvement. These sounds were perceived similar to each other, even though they have different amount of modes. Finally we observe a cluster of sounds (2 clusters for the vase with the naive method) with an intermediate amount of modes. The sounds can be heard on the website accompanying these proceedings, or on [18].

5. CONCLUSIONS

Modal models have been shown to very useful for the real-time generation of high quality sound-effects for animation and simulation. The efficiency of the synthesis can be improved by several orders of magnitude by a careful selection of the modal model. Several heuristics were proposed, inspired by knowledge of masking effects occurring in human sound processing. The resulting synthetic sounds were then tested on subjects in order to assess the various mode selection methods.

It was found that the optimal parameters for synthesizing sounds varied substantially amongst participants. The number of synthetic sounds of different complexity (number of modes) that could be distinguished also varied widely from subject to subject. By comparing two methods for selecting the modes for the progressive sound model, it was found that higher quality sounds can be achieved with less modes using a mode selection heuristics based on masking characteristics of the human ear.

A perceptual map of a small set of synthetic sounds utilizing different number of modes and the real sound they intend to approximate showed the sounds cluster in 3-4 groups.

We noticed that subjects often claim to hear a difference between identical sounds, which results in an overestimation of a_v and in an overestimation of the perceived differences between sounds. This problem could be overcome by using forced choice procedures instead, which are however more time-consuming.

Acknowledgements

This work was supported in part by grants from the UBC Peter Wall Institute for Advanced Studies, NSERC, CIHR, and MSFHR. One of us (K. van den Doel) wishes to thank Donald Greenwood for useful discussions regarding tonal masking.

6. REFERENCES

[1] Kees van den Doel, Paul G. Kry, and Dinesh K. Pai, "FoleyAutomatic: Physically-based Sound Effects for Interactive Simulation and Animation," in *Computer Graphics (ACM SIGGRAPH 01 Conference Proceedings)*, 2001.

[2] K. van den Doel and D. K. Pai, "The sounds of physical shapes," *Presence*, vol. 7, no. 4, pp. 382–395, 1998.

[3] K. van den Doel, *Sound Synthesis for Virtual Reality and Computer Games*, Ph.D. thesis, University of British Columbia, 1998.

[4] K. van den Doel and D. K. Pai, "Synthesis of Shape Dependent Sounds with Physical Modeling," in *Proceedings of the International Conference on Auditory Displays 1996, Palo Alto*, 1996.

[5] J. F. O'Brien, P. R. Cook, and G. Essl, "Synthesizing Sounds from Physically Based Motion," in *SIGGRAPH 01*, 2001.

[6] P. R. Cook, "Integration of physical modeling for synthesis and animation," in *Proceedings of the International Computer Music Conference*, Banff, 1995, pp. 525–528.

[7] J. K. Hahn, H. Fouad, L. Gritz, and J. W. Lee, "Integrating sounds and motions in virtual environments," in *Sound for Animation and Virtual Reality, SIGGRAPH 95 Course 10 Notes*, 1995.

[8] T. Takala and J. Hahn, "Sound rendering," *Proc. SIGGRAPH 92, ACM Computer Graphics*, vol. 26, no. 2, pp. 211–220, 1992.

[9] W. W. Gaver, "What in the world do we hear?: An ecological approach to auditory event perception," *Ecological Psychology*, vol. 5, no. 1, pp. 1–29, 1993.

[10] W. W. Gaver, "Synthesizing auditory icons," in *Proceedings of the ACM INTERCHI 1993*, 1993, pp. 228–235.

[11] Judith C. Brown and Miller S. Puckette, "A high resolution fundamental frequency determination based on phase changes of the fourier transform," *J. Acoust. Soc. Am.*, vol. 94, no. 2, pp. 662–667, 1993.

[12] K. van den Doel and D. K. Pai, "JASS: A Java Audio Synthesis System for Programmers," in *Proceedings of the International Conference on Auditory Displays 2001, Helsinki*, 2001.

[13] "http://www.cs.ubc.ca/~kvdoel/jass," .

[14] D. M. Green, "Additivity of Masking," *J. Acoust. Soc. Am.*, vol. 41, no. 6, pp. ?, 1967.

[15] Frank Baumgarte, Charalampos Ferekidis, and Hendrik Fuchs, "A Nonlinear Psychoacoustic Model Applied to the ISO MPEG Layer 3 Coder," in *Preprint 4097, 99th AES Convention, New York, October 1995*, New York, 1995.

[16] T. Carvellas and B. Schneider, "Direct estimation of multidimensional tonal dissimilarity," *J. Acoust. Soc. Am.*, vol. 51, no. 6, pp. 1839–1848, 1971.

[17] S. S. Stevens, "The direct estimation of sensory magnitudes-loudness," *Amer. J. Psychol.*, vol. 69, pp. 1–25, 1956.

[18] "http://www.cs.ubc.ca/ kvdoel/icad2002," .

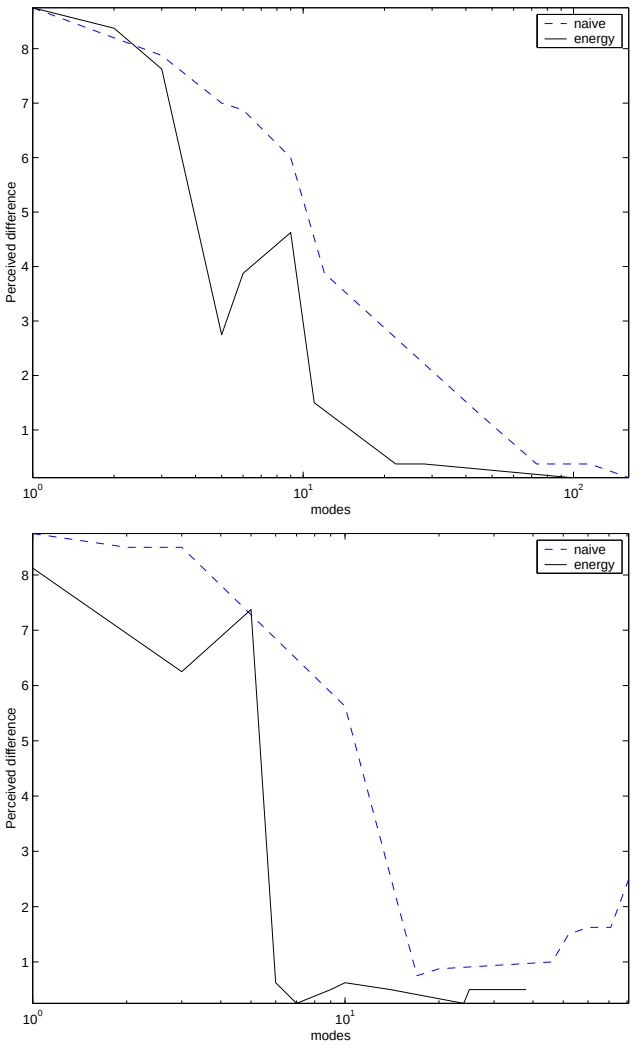

Figure 4: Perceived difference between real and synthetic sound for two mode selection methods as a function of number of modes, averaged over 8 subjects. The upper figure is for the vase, the lower for the bowl.

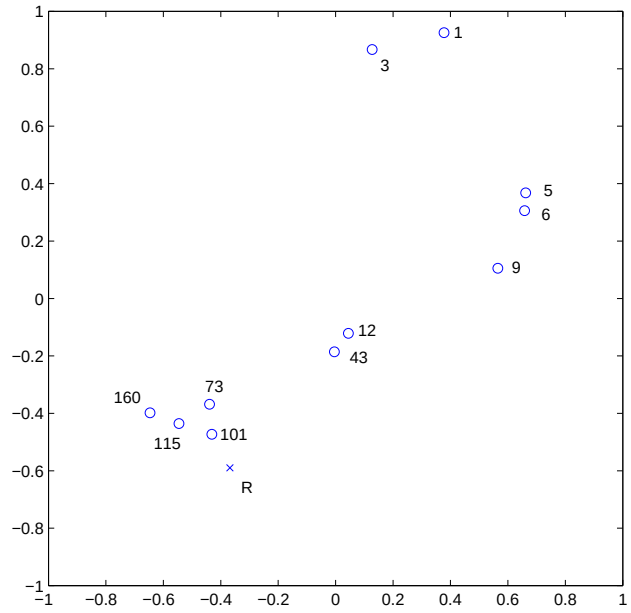

Figure 5: Perceptual map of the vase sounds with the naive method.

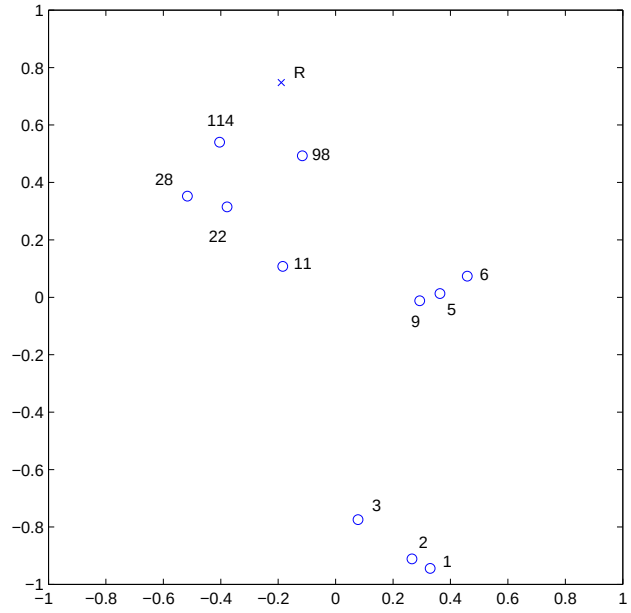

Figure 6: Perceptual map of the vase sounds with the energy method.

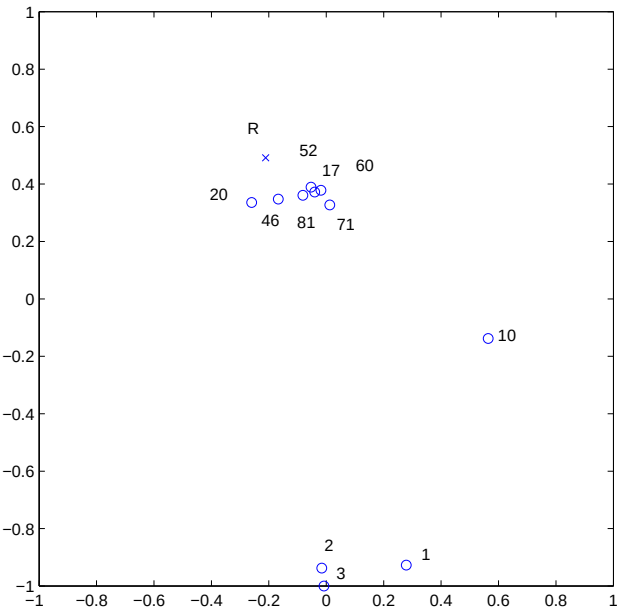

Figure 7: Perceptual map of the bowl sounds with the naive method.

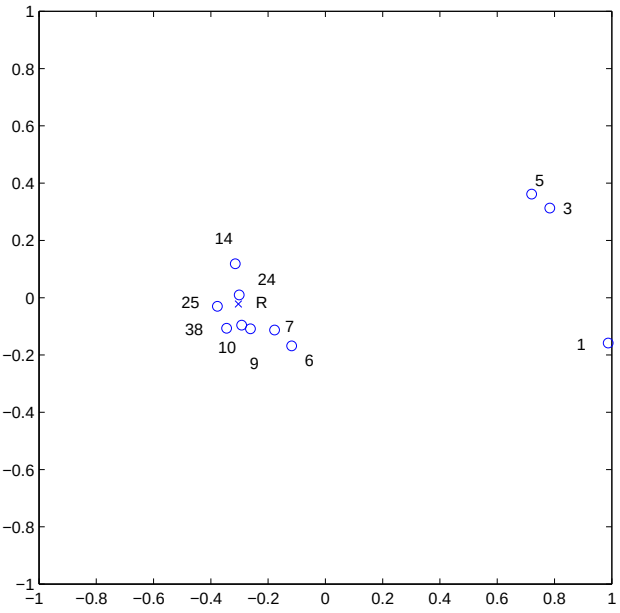

Figure 8: Perceptual map of the bowl sounds with the energy method.