A Meta-analysis of Timbre Perception Using Nonlinear Extensions to CLASCAL

John Ashley Burgoyne and Stephen McAdams

Centre for Interdisciplinary Research in Music and Media Technology
Schulich School of Music of McGill University
555 Sherbrooke Street West
Montral, Qubec, Canada H3A 1E3
{ashley,smc}@music.mcgill.ca

Abstract. Seeking to identify the constituent parts of the multidimensional auditory attribute that musicians know as timbre, music psychologists have made extensive use of multidimensional scaling (MDS), a statistical technique for visualising the geometric spaces implied by perceived dissimilarity. MDS is also well known in the machine learning community, where it is used as a basic technique for dimensionality reduction. We adapt a nonlinear variant of MDS that is popular in machine learning, Isomap, for use in analysing psychological data and re-analyse three earlier experiments on human perception of timbre. Isomap is designed to eliminate undesirable nonlinearities in the input data in order to reduce the overall dimensionality; our results show that it succeeds in these goals for timbre spaces, compressing the output onto well-known dimensions of timbre and highlighting the challenges inherent in quantifying differences in spectral shape.

1 Introduction

As any computer musician knows, timbre is one of the most important compositional parameters, and yet it remains one of the most under-theorised. Part of the reason for the relative lack of theory may be due to the fact that, unlike pitch, timbre is a multidimensional auditory attribute, and it it is difficult to draw general conclusions about timbre as a whole without first identifying its constituent parts. Nonetheless, there have been a number of attempts to uncover the underlying dimensionality of timbre space over the past few decades, most based on perceptual experiments with synthesised and recorded tones. Early experiments with synthetic tones identified spectral centroid and attack time as primary components of timbre, in addition, at times, to a third dimension that was more difficult to interpret and dependent on the stimulus set (Grey 1977; Grey and Gordon 1978). Later studies with more sophisticated models came to similar conclusions, and suggested that the third component might be a measure of irregularity in the spectral envelope (Krumhansl 1989) or spectral flux (McAdams et al. 1995). A recent confirmatory study has verified that spectral centroid, attack time, and spectral irregularity are indeed recovered, whereas spectral flux is not (Caclin et al. 2005).

R. Kronland-Martinet, S. Ystad, and K. Jensen (Eds.): CMMR 2007, LNCS 4969, pp. 181–202, 2008.

All of these studies are based on a statistical technique known as multidimensional scaling (MDS) (Torgerson 1958). The basic idea of MDS is to take the set of proximities between all members of some set of data points, e.g., sample timbres, and to model them as distances in a Euclidean space of as few dimensions as possible. In the context of timbre, these proximities are usually taken from psychological experiments in which human subjects have rated their perception of the (dis)similarity between timbre pairs. The trouble with MDS in this context is that its classical form was designed to interpret a single set of dissimilarities among items, not the average over all subjects of an experiment. The first robust solution to this problem was the INDSCAL algorithm (Carroll and Chang 1970), which models a special weight on each dimension for each subject in the experiment in order to better fit model distances to the set of empirical dissimilarities. The more sophisticated CLASCAL algorithm reduces the number of parameters in INDSCAL by modelling weights not for individual subjects but for a smaller number of aggregate subject groups, called latent classes (Winsberg and De Soete 1993). CLASCAL and its variants are the standard techniques for analysing timbre spaces today.

There is another problem with these techniques, however: being linear, they consider all distances estimated by the human subjects to be equally reliable and of equal relative scale. Although some relatively straightforward extensions to MDS can treat the latter problem, e.g., CONSCAL (Winsberg and De Soete 1997), the former requires more aggressive modifications. One such modification, known as Isomap, replaces large distances in the original distance matrices with so-called geodesic distances along a hypothetical manifold (Tennenbaum et al. 2000). Previous work with Isomap has demonstrated that it and its relatives can uncover meaningful musical relationships that traditional linear MDS will always miss (Burgoyne and Saul 2005).

This paper combines the CLASCAL and Isomap models to re-analyse the data from three major studies of timbre: Grey 1977, Grey and Gordon 1978, and McAdams et al. 1995. Section 2 provides a more detailed explanation of these algorithms and best practises for interpreting their results. Section 3 presents the results of our new scaling and compares them to the original studies. Section 4 concludes with suggestions for future applications of nonlinear scaling to the study of musical timbre.

2 CLASCAL and Isomap

2.1 CLASCAL

Traditional MDS was designed to handle a single set of pairwise proximities only. A number of models have been presented to adapt MDS for multiple-subject experiments, of which the most important for studying timbre has been CLASCAL (Winsberg and De Soete 1993). The CLASCAL model seeks to minimise the approximation error in the following equation:

$$d_{ijk} \approx \left[\sum_{r=1}^{R} w_{\mathcal{C}(i),r}(x_{jr} - x_{kr})^2\right]^{1/2} , \tag{1}$$

where d_{ijk} is the dissimilarity rating that subject i assigned to stimulus pair (j, k), R is the number of dimensions in the output set, $w_{\mathcal{C}(i),r}$ is a special weight on dimension r for the so-called latent class $\mathcal{C}(i)$ to which CLASCAL has assigned subject i, and x_{jr} and x_{kr} are the co-ordinates along dimension r for stimuli j and k. Latent classes are meant to represent groups of subjects who pursue similar rating strategies. The number of latent classes used is a compromise between over-parametrisation, e.g., the INDSCAL model, which assigns each subject to its own class, and over-generalisation, e.g., ignoring differences between subjects by taking the average over all dissimilarity matrices. A Monte Carlo likelihood-ratio technique is used to determine the optimal number of classes (Hope 1968; Aitkin et al. 1981). The class weights can be interpreted as the rating strategies used by each class: relatively high weights for a particular dimension suggest that members of the class use that dimension more than others when distinguishing timbres.

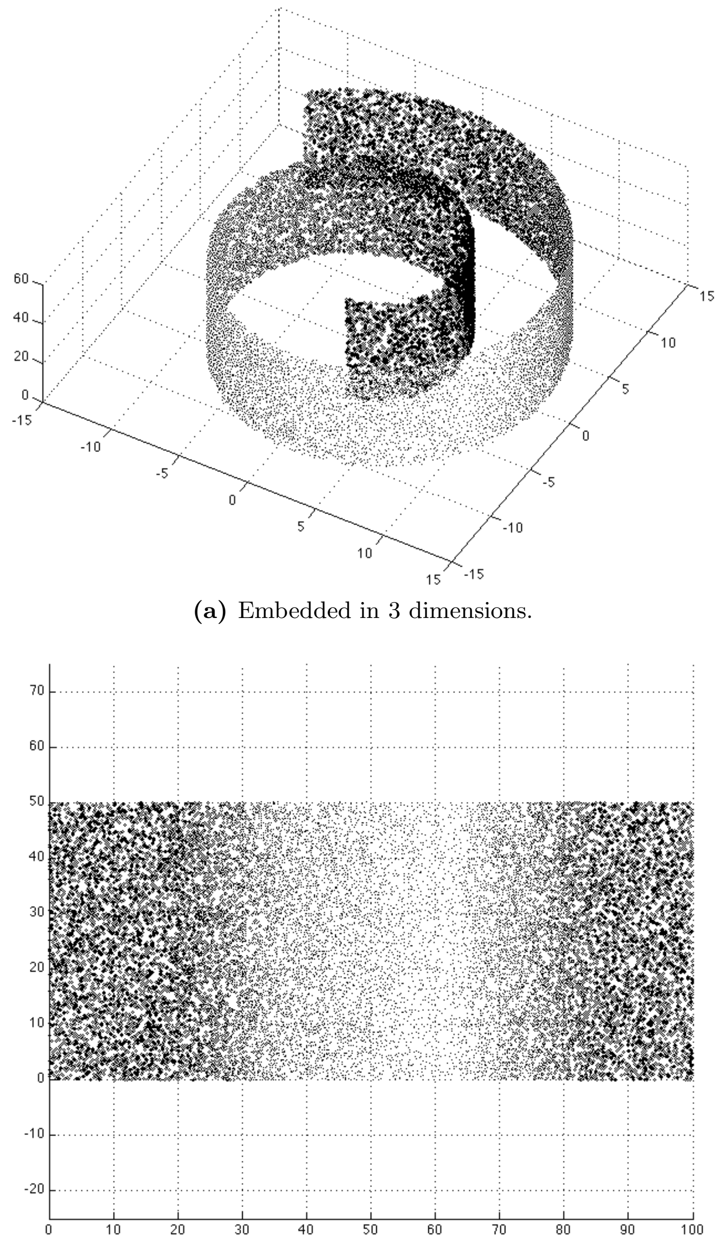
Another potential problem with traditional MDS is that it assumes all of the variance in a data set can be explained by dimensions common to all stimuli. This assumption does not always hold for timbre: many timbres include instrument-specific components such as the sound of the returning hopper in a harpsichord. A more sophisticated version of CLASCAL separates these components, known as *specificities*, using the following model:

$$d_{ijk} \approx \left[\sum_{r=1}^{R} w_{\mathcal{C}(i),r}(x_{jr} - x_{kr})^2 + v_{\mathcal{C}(i)}(s_j + s_k)\right]^{1/2} , \tag{2}$$

where s_j and s_k are the specificities for stimuli j and k and $v_{\mathcal{C}(i)}$ represents the weight subjects in class $\mathcal{C}(i)$ give to specificities when distinguishing timbres (Winsberg and Carroll 1989; McAdams et al. 1995).

2.2 Isomap

Isomap arose as a solution to the dimensionality reduction problem for data sets like the famous ‘Swiss roll’ pictured in Fig. 1 (Tennenbaum et al. 2000). Looking at the plot, it is obvious to a human that the data are arranged on a two-dimensional plane that has been coiled and presented in three dimensions. This fact is not obvious to traditional MDS, which strives to preserve every pairwise distance in the set, including those between the ends of the roll and the inner or outer loops. The ingenious solution in Isomap is to throw away all pairwise distances in the set except those at the local level, i.e., those in a small region immediately surrounding each point in the data set. These regions can be selected as a fixed number k of the nearest neighbours to each point in the data set or as those points that fall within a sphere of fixed radius ϵ around each point. The other distances are then recomputed using an all-pairs shortest-path algorithm, yielding an approximation of the so-called *geodesic distances*, or



(a) Embedded in 3 dimensions.

(b) Unrolled in 2 dimensions.

Fig. 1. The 'Swiss roll' data set. On the left, the data is presented in its original form. On the right, the data is presented as it should be unrolled for human interpretation. Traditional MDS can never arrive at this solution, however, because it seeks to preserve the distances between the ends of the roll and the inner/outer loops.

distances in the lower-dimensional form. After these approximate distances are computed, traditional MDS is applied.

Like many algorithms based on nearest neighbours, Isomap can be improved by incorporating variants of the relative neighbourhood graph, e.g., the Gabriel graph (Gabriel and Sokal 1969; Jaromczyk and Toussaint 1992). The Gabriel graph retains the pairwise distance d_{ab} between points a and b if and only if there is no point c such that $d_{ac}^2 + d_{cb}^2 < d_{ab}^2$. Interpreted geometrically, the Gabriel graph retains the pairwise distance between two points if and only if the minimum-volume hypersphere connecting them, i.e, the diameter sphere, is empty. The Gabriel graph of a set of points is a superset of the minimum spanning tree, and thus, unlike the graphs based on fixed k or ϵ in traditional Isomap, is guaranteed to be connected. It is nonparametric and robust to variations in density throughout the observed space. Because of these desirable properties, all of our experiments with Isomap retained the Gabriel neighbours rather than fixed neighbourhoods specified in the original algorithm.

At first glance, the Swiss roll appears to be a fundamentally different problem than that of estimating timbre spaces. There is little reason to believe that human subjects would willfully twist their ratings of the similarities between timbre pairs into more dimensions than are already present. The larger message of Isomap, however, is that unless a space is perfectly linear, large distances in a scaling model can mask important structures in the data. It seems prudent to check for such structures in psychological data, and because Isomap is based on classical MDS, unlike a number of other nonlinear scaling techniques, it lends itself naturally to combination with CLASCAL. Each subject's dissimilarity matrix is processed according to the Isomap algorithm until the final MDS step. After this preprocessing is complete, the new dissimilarity matrices are fed to CLASCAL.

3 Experiments and Results

As stated earlier, rather than performing a new experiment, we used the data from three earlier experiments to evaluate the effects of Isomap processing. Results based on (1) are presented first, followed by a discussion of the changes to the output spaces after adding specificities to the model as per (2). In interpreting the results, we computed the acoustical features proposed in Peeters et al. 2000 for each timbre stimulus and studied the Pearson correlation coefficients of these features with the dimensions output from CLASCAL.

3.1 Grey 1977

Although the complete experimental results from the Grey 1977 are, the sound stimuli used for the experiments are still available. These stimuli comprise electronically resynthesised imitations of tape recordings of pitch E♭4 (311 Hz) played on two different oboes, an English horn, a bassoon, an E♭ clarinet, a bass clarinet, a flute, two alto saxophone sounds (one played *piano* and one played *mezzo-forte*), a soprano saxophone, a trumpet, a French horn, a muted trombone, and three cello sounds (normal bowing, muted *sul tasto*, and *sul ponticello*). These sounds were then normalised to a consistent perceived loudness, pitch, and duration based on a separate perceptual experiment. We performed a study with 22 new subjects, all professional musicians, using these stimuli, asking each subject to rate the dissimilarity between each of the 120 pairs of stimuli on a continuous scale, which was later converted to range from 0 to 1 for data analysis.

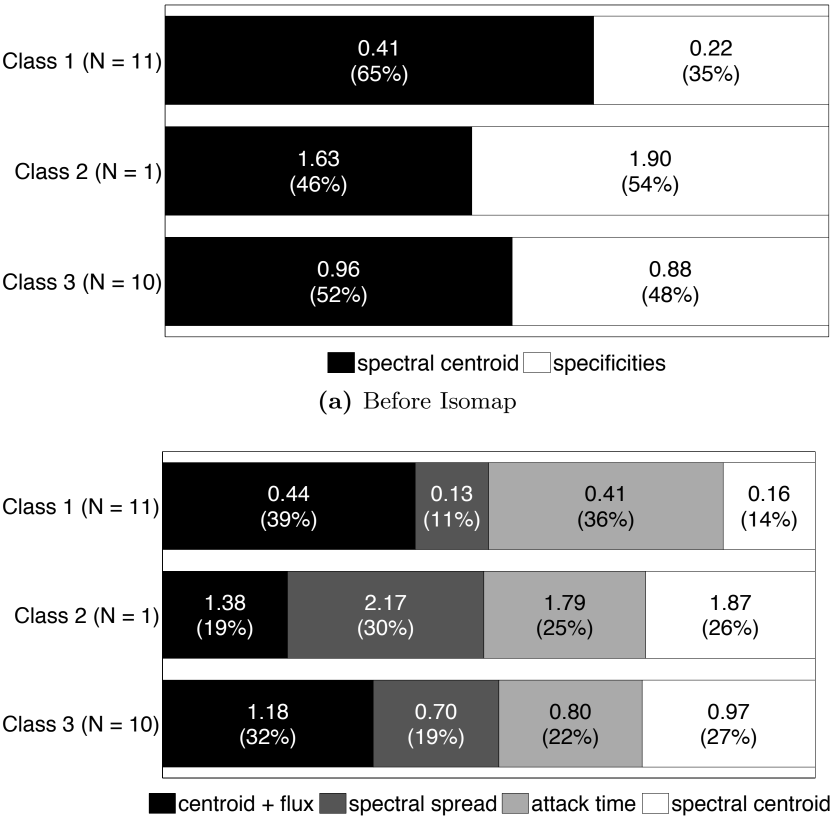
The optimal CLASCAL model prior to Isomap processing contains two latent classes and three dimensions. In descending order of prominence, the dimensions correlate with spectral centroid ($r = 0.90$), spectral slope ($r = 0.79$), and log attack time ($r = 0.71$) for the stimuli. Both latent classes weight these dimensions fairly evenly (see Fig. 2a); the primary difference between the classes appears to be that subjects in Class 1 (9 subjects) made use of a wider range of the rating scale than those in Class 2 (13 subjects). The co-ordinates of each stimulus are listed for reference in Table 1, and Fig. 3a presents a plot of the stimuli positioned in the space. In this plot and all future plots, points are connected according to their minimum spanning tree, which includes all pairs of nearest neighbours; points in the plot that appear to be close together but are not connected by a dark line are in fact farther from each other than they appear.

After Isomap processing, the optimal CLASCAL model contains three latent classes and four dimensions. Again in descending order of prominence, these dimensions correlate with a combination of spectral centroid ($r = 0.91$) and spectral flux ($r = 0.80$), spectral spread ($r = -0.74$), log attack time ($r = 0.86$), and – curiously – spectral centroid again ($r = 0.76$). Note that the first two dimensions correlate highly with the first two dimensions prior to Isomap processing ($r = 0.84$ and $r = 0.71$). Subjects in Class 1 (11 subjects) appear to have used a rating strategy emphasising spectral centroid and log attack time, whereas subjects in Class 3 (10 subjects) weighted the four dimensions more evenly. Class 2 contains a single subject who weighted the first dimension relatively less and used more of the scale overall than those in Class 3. The raw and relative weights are available in Fig. 2b. A full set of co-ordinates appears in Table 1 and a plot of the leading three dimensions of the space is presented in Fig. 3b. Despite the extra dimension and some mild changes to the minimum spanning tree with respect to the saxophones and oboes, the overall structure is similar to the space prior to Isomap processing.

The CLASCAL technique is more sophisticated than the MDS techniques that were available to Grey originally, and so some differences in the output dimensions are to be expected. The overall structure of our timbre space, however, is similar to the original published space both before and after Isomap processing.

3.2 Grey and Gordon 1978

Like Grey 1977, the complete experimental results of Grey and Gordon 1978 are no longer available, but we were able to locate the sound stimuli used. These stimuli were mostly the same as the Grey 1977 stimuli, but for four pairs of instruments – (a) oboe 1 and bass clarinet, (b) bassoon and French horn, (c) cello *sul tasto* and normal cello, and (d) muted trombone and trumpet) – the



(a) Before Isomap

(b) After Isomap

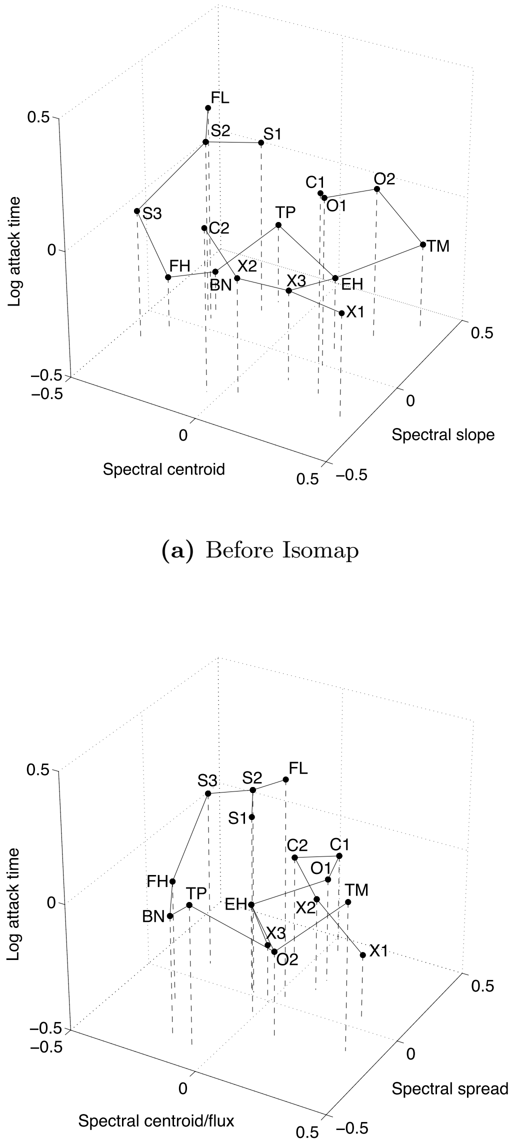
Fig. 2. Raw and relative weights on dimensions for latent classes of subjects using Grey 1977 stimuli. Although the overall listening strategies (relative weights) are similar before Isomap, the raw weights reveal that subjects in Class 1 used a wider range of the rating scale than subjects in Class 2.

spectral envelopes were exchanged during synthesis. The purpose of these exchanges was to test the effect of changes in spectral envelope on timbre perception by comparing the MDS space resulting from these timbres to that of the original space. Again, we conducted a new experiment using the same 22 subjects on these stimuli, asking them to rate the dissimilarity between all 120 pairs on a continuous scale.

Before Isomap processing, the optimal CLASCAL model contains three classes and two dimensions. Like the leading two dimensions of the Grey 1977 space, the dimensions correlate with spectral centroid ($r = 0.93$) and spectral spread ($r = -0.67$). All three latent classes suggest fairly even rating strategies, although Class 1 (5 subjects) shows a slight preference for spectral spread and Class 2 (9 subjects) shows a slight preference for spectral centroid relative to Class 3 (8 subjects), which distributes the weights most evenly (see Fig. 4a). All

Table 1. Grey 1977 instruments and fitted model co-ordinates along common CLASCAL dimensions and specificities

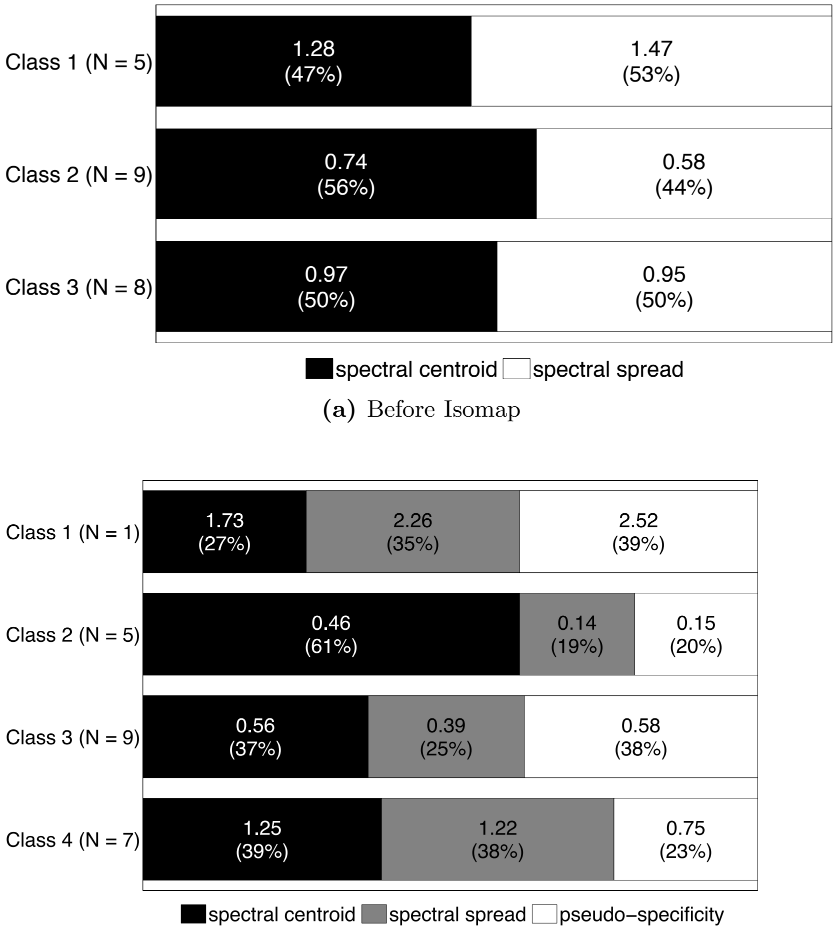
	Instrument	Before Isomap							After Isomap					
		no specificities			with specificities				no specificities				with spec.	
		1	2	3	1	2	3	$\sqrt{s}$	1	2	3	4	1	$\sqrt{s}$
BN	Bassoon	−0.288	0.111	−0.332	−0.272	0.244	−0.236	0.145	−0.181	0.351	−0.026	0.224	−0.203	0.391
C1	E♭ clarinet	0.284	−0.164	0.280	0.230	−0.270	0.179	0.179	0.051	−0.399	−0.099	−0.074	0.177	0.324
C2	Bass clarinet	−0.050	−0.344	0.144	−0.145	−0.260	0.212	0.063	−0.070	−0.295	−0.088	0.230	0.005	0.369
EH	English horn	0.203	0.097	−0.203	0.130	−0.035	−0.261	0.155	−0.078	−0.006	−0.135	−0.288	0.132	0.310
FH	French horn	−0.380	−0.055	−0.296	−0.396	0.165	−0.145	0.192	−0.290	0.160	−0.021	0.177	−0.233	0.332
FL	Flute	−0.280	0.065	0.343	−0.066	0.184	0.366	0.207	0.032	−0.050	0.359	0.205	−0.237	0.351
O1	Oboe 1	0.212	−0.005	0.165	0.195	−0.100	0.055	0.230	0.107	−0.216	−0.087	−0.218	0.280	0.000
O2	Oboe 2	0.281	0.241	0.100	0.307	0.012	−0.059	0.279	0.126	0.188	−0.158	−0.262	0.184	0.341
S1	Cello *normale*	−0.151	0.201	0.180	0.053	0.242	0.177	0.071	−0.018	0.089	0.269	−0.013	−0.194	0.219
S2	Cello *sul tasto*	−0.312	0.098	0.188	−0.120	0.249	0.259	0.000	−0.078	−0.020	0.304	0.017	−0.240	0.192
S3	Cello *sul ponticello*	−0.421	−0.177	0.019	−0.400	0.088	0.178	0.063	−0.309	−0.109	0.188	0.089	−0.320	0.243
TM	Muted trombone	0.402	0.363	−0.147	0.434	0.082	−0.269	0.241	0.377	0.142	0.085	−0.278	0.225	0.507
TP	Trumpet	−0.127	0.280	−0.180	−0.027	0.269	−0.182	0.173	−0.092	0.371	0.052	−0.071	−0.103	0.366
X1	Alto saxophone *mf*	0.422	−0.263	−0.082	0.182	−0.398	−0.127	0.251	0.326	−0.068	−0.247	0.133	0.285	0.382
X2	Alto saxophone *p*	0.051	−0.305	−0.039	−0.106	−0.270	0.004	0.126	0.103	−0.144	−0.130	0.195	0.108	0.290
X3	Soprano saxophone	0.155	−0.143	−0.141	0.000	−0.200	−0.148	0.148	−0.005	0.008	−0.268	−0.067	0.134	0.283



(a) Before Isomap

(b) After Isomap

Fig. 3. Grey 1977 timbre spaces. Axes are labelled with their acoustic correlates, and to aid visualisation, points are connected according to the minimum spanning tree in the common CLASCAL spaces. The two structures are similar, but the space groups more tightly after Isomap.



(a) Before Isomap

(b) After Isomap

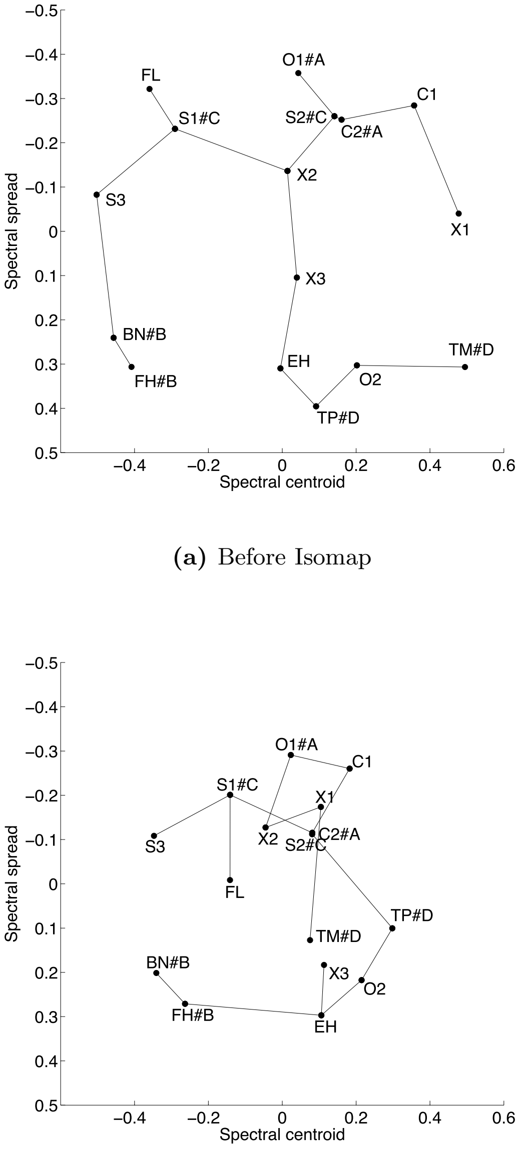
Fig. 4. Raw and relative weights on dimensions for latent classes of subjects using Grey and Gordon 1978 stimuli

co-ordinates are listed in Table 2 for reference and a plot of the output space appears in Fig. 5a.

Before Isomap processing, CLASCAL suggests a space with 2 classes and 3 dimensions. Consistent with the dimensions recovered for the full set of 88 subjects, the first two dimensions correlate with log attack time ($r = -0.82$), spectral centroid ($r = -0.89$); the third dimension, however, correlated best with spectral spread ($r = 0.73$) rather than spectral flux ($r = 0.33$) as in the original paper. The primary difference between Class 1 (14 subjects) and Class 2 (10 subjects) is scale, Class 2 making fuller use of the scale than Class 1, although Class 1 shows a slight preference for log attack time and Class 2 shows a slight preference for

Table 2. Grey and Gordon 1978 instruments and fitted model co-ordinates along common CLASCAL dimensions and specificities

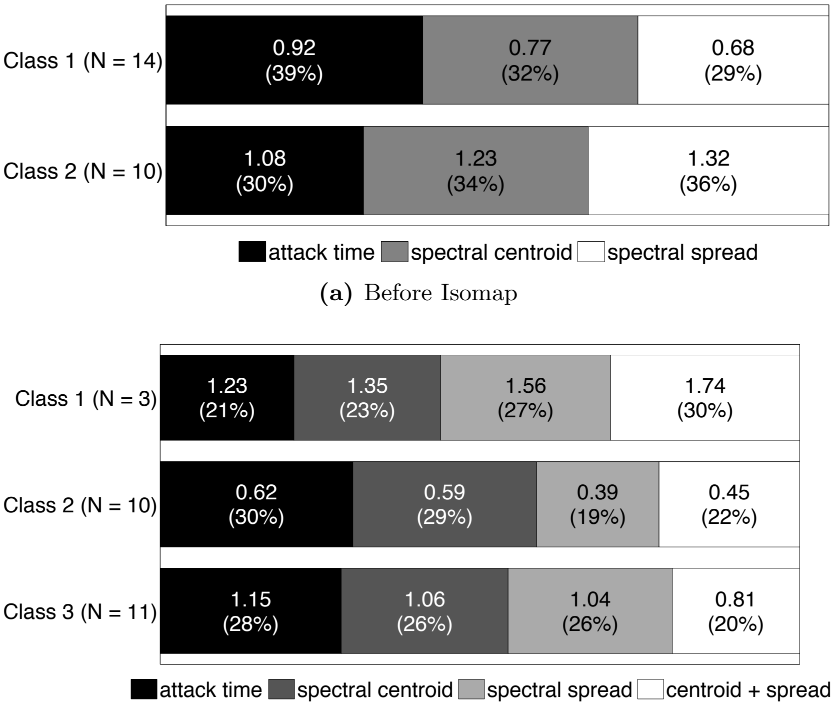
	Instrument	Before Isomap		After Isomap				
		no specificities		no specificities			with spec.	
		1	2	1	2	3	1	$\sqrt{s}$
BN	Bassoon with French horn envelope	−0.456	0.241	−0.341	0.201	−0.041	−0.270	0.307
C1	E♭ clarinet	0.357	−0.284	0.182	−0.260	−0.013	0.168	0.241
C2	Bass clarinet with oboe 1 envelope	0.161	−0.252	0.081	−0.112	0.051	0.092	0.000
EH	English horn	−0.005	0.310	0.106	0.297	−0.023	0.060	0.346
FH	French horn with bassoon envelope	−0.408	0.306	−0.263	0.271	−0.050	−0.224	0.327
FL	Flute	−0.359	−0.322	−0.141	−0.009	0.470	−0.184	0.431
O1	Oboe 1 with bass clarinet envelope	0.044	−0.357	0.023	−0.291	−0.008	0.079	0.274
O2	Oboe 2	0.202	0.303	0.215	0.218	0.125	0.112	0.339
S1	Cello *normale* with *sul tasto* envelope	−0.291	−0.232	−0.141	−0.201	0.308	−0.136	0.362
S2	Cello *sul tasto* with *normale* envelope	0.141	−0.260	0.081	−0.115	0.057	0.091	0.045
S3	Cello *sul ponticello*	−0.502	−0.083	−0.347	−0.108	0.176	−0.364	0.032
TM	Muted trombone with trumpet envelope	0.495	0.307	0.075	0.127	−0.551	0.151	0.540
TP	Trumpet with muted trombone envelope	0.091	0.395	0.298	0.100	0.095	0.105	0.344
X1	Alto saxophone *mf*	0.478	−0.040	0.105	−0.174	−0.300	0.191	0.298
X2	Alto saxophone *p*	0.014	−0.136	−0.045	−0.127	−0.159	0.058	0.239
X3	Soprano saxophone	0.039	0.105	0.113	0.183	−0.137	0.070	0.326



(a) Before Isomap

(b) After Isomap

Fig. 5. Grey and Gordon 1978 timbre spaces. Axes are labelled with their acoustic correlates, and to aid visualisation, points are connected according to the minimum spanning trees in their common CLASCAL spaces. In the post-Isomap space, the minimum spanning tree incorporates distance information from the unplotted third dimension, which is why it crosses itself.



(a) Before Isomap

(b) After Isomap

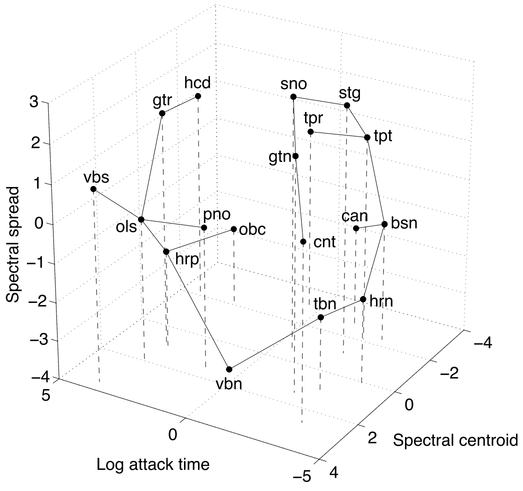
Fig. 6. Raw and relative weights on dimensions for latent classes of subjects using McAdams et al. 1995 stimuli

spectral spread in their respective rating strategies (see Fig. 6a). A full set of co-ordinates appears in Table 3 and the space is plotted in three dimensions in Fig. 7a.

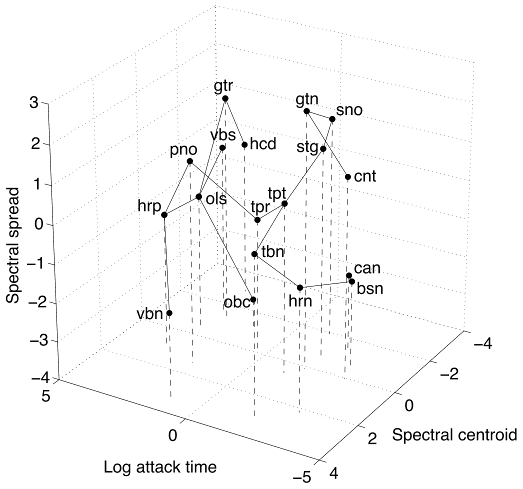
After Isomap processing, the optimal CLASCAL model comprises three classes and four dimensions. The first three dimensions correlate with the same acoustic features as those of the pre-Isomap space, log attack time ($r = -0.81$), spectral centroid ($r = -0.68$), and spectral spread ($r = 0.82$); the fourth dimension appears to be some combination of spectral centroid ($r = -0.69$) and spectral spread ($r = 0.78$). The first and third dimensions of the spaces correlate highly with each other ($r = 0.97$ and $r = 0.77$), but it is the fourth dimension after Isomap that correlates with the second dimension before Isomap ($r = 0.81$). Class 1 (3 subjects) shows a slight preference for Dimensions 3 and 4, Class 2 (10 subjects) shows a slight preference for Dimensions 1 and 2, and Class 3 (11 subjects) weights all four dimensions evenly (see Fig. 6b). Table 3 contains a list of the exact co-ordinates for each stimulus and Fig. 7b provides a projection of this space into three dimensions.

Table 3. McAdams et al. 1995 instruments and fitted model co-ordinates along common CLASCAL dimensions and specificities

Instrument		Before Isomap			After Isomap					
		no specificities			no specificities				with spec.	
		1	2	3	1	2	3	4	1	$\sqrt{s}$
hrn	French horn	−3.87	0.18	−1.73	−2.96	2.21	−0.79	0.54	−2.82	2.54
tpt	Trumpet	−2.35	−2.11	1.30	−1.18	0.59	0.40	−2.50	−1.52	2.10
trn	Trombone	−2.82	0.94	−2.10	−1.86	3.00	0.14	−0.78	−2.25	2.74
hrp	Harp	3.14	0.96	−1.49	3.26	0.87	−0.56	0.53	2.84	1.85
tpr	Trumpar (trumpet/guitar hybrid)	0.31	−2.81	0.78	0.35	−0.07	−0.52	−2.63	−0.23	2.53
ols	Oboleste (oboe/celesta hybrid)	3.22	2.02	−0.27	2.79	−0.26	−0.42	2.17	3.15	1.58
vbs	Vibraphone	3.73	3.58	0.97	2.84	−1.53	0.41	3.71	3.60	3.00
sno	Striano (string/piano hybrid)	−0.70	−0.51	2.58	−1.14	−1.94	1.65	−0.86	−0.90	2.70
hcd	Harpsichord	4.22	−2.07	1.33	3.01	−2.91	0.00	−2.01	2.19	4.02
can	English horn	−1.64	−2.62	−1.35	−2.26	−1.37	−2.00	0.02	−1.91	2.76
bsn	Bassoon	−2.91	−2.34	−0.92	−2.99	−0.47	−1.67	−0.92	−2.63	2.29
cnt	Clarinet	−3.40	2.66	0.58	−3.14	0.04	1.20	2.43	−2.25	3.41
vbn	Vibrone (vibraphone/trombone hybrid)	0.37	1.37	−3.91	1.00	3.46	−1.69	0.97	0.51	4.06
obc	Obochord (oboe/harpsichord hybrid)	2.91	−2.19	−2.02	1.40	−1.24	−3.27	0.24	1.46	3.48
gtr	Guitar	3.12	1.05	2.09	2.47	−1.23	1.84	0.83	2.54	2.21
stg	Strings	−2.25	−1.13	2.41	−1.73	−0.63	1.44	−2.12	−1.65	2.33
pno	Piano	1.13	1.59	−0.27	1.80	1.38	1.23	−0.06	1.29	2.32
gtn	Guitarnet (guitar/clarinet hybrid)	−2.19	1.41	2.01	−1.64	0.10	2.60	0.42	−1.41	2.68



(a) Before Isomap



(b) After Isomap

Fig. 7. McAdams et al. 1995 timbre space before Isomap. Axes are labelled with their acoustic correlates, and to aid visualisation, points are connected according to the minimum spanning trees in their common CLASCAL spaces.

3.3 Specificities

As stated earlier, all of the above models were derived without allowing for specificities, i.e, using (1) rather than (2). Before Isomap processing, the specificities make little difference. An analogous Monte Carlo likelihood-ratio test to the one used for determining the optimal number of classes prefers models without specificities for Grey and Gordon 1978 and McAdams et al. 1995. Although a model with specificities is preferred for Grey 1977, the model structure changes very little on account of them. The optimal model including specificities still contains two classes and three dimensions. All three respective pairs of dimensions (with and without specificities) correlate highly with each other ($r = 0.89$, $r = 0.74$, and $r = 0.89$), although after including specificities, the second dimension correlates better with a psychoacoustical model for perceived roughness ($r = -0.77$; see von Bismarck 1974) than it does with spectral spread. Like the model without specificities, both latent classes weight the dimensions fairly evenly; the difference between them is that Class 1 (7 subjects) incorporates specificities into its rating strategy whereas Class 2 (15 subjects) does not (see Fig. 8). The output space, which is structurally almost identical to the space without specificities, is presented in Fig. 9, and a complete co-ordinate listing appears in Table 1.

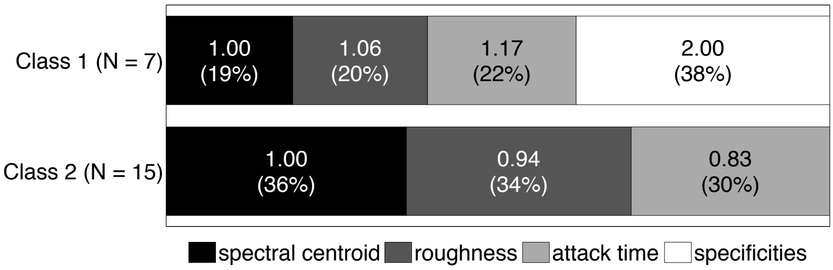


Fig. 8. Raw and relative weights for each latent class in Grey 1977 before Isomap processing and including specificities

After Isomap processing, however, the Monte Carlo test prefers models with specificities in all cases. Moreover, although the optimal number of latent classes remains unchanged, the inclusion of specificities reduces the optimal number of common dimensions to just one: spectral centroid for Grey 1977 ($r = 0.83$) and Grey and Gordon 1978 ($r = 0.91$) and log attack time for McAdams et al. 1995 ($r = -0.78$). These single dimensions correlate highly with the leading dimensions in their respective spaces before and after Isomap, with and without specificities; all other information has been pushed out into the specificity dimensions. It is impossible to visualise these models, unfortunately, although their co-ordinate values are listed in their respective tables. The latent classes within each of these models differ primarily in the relative weight they place on the specificities (see Fig. 10).

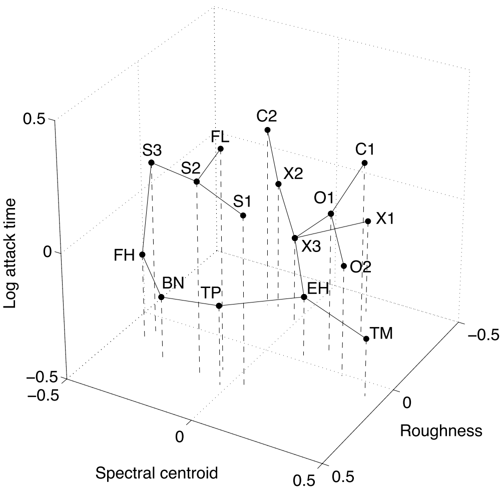


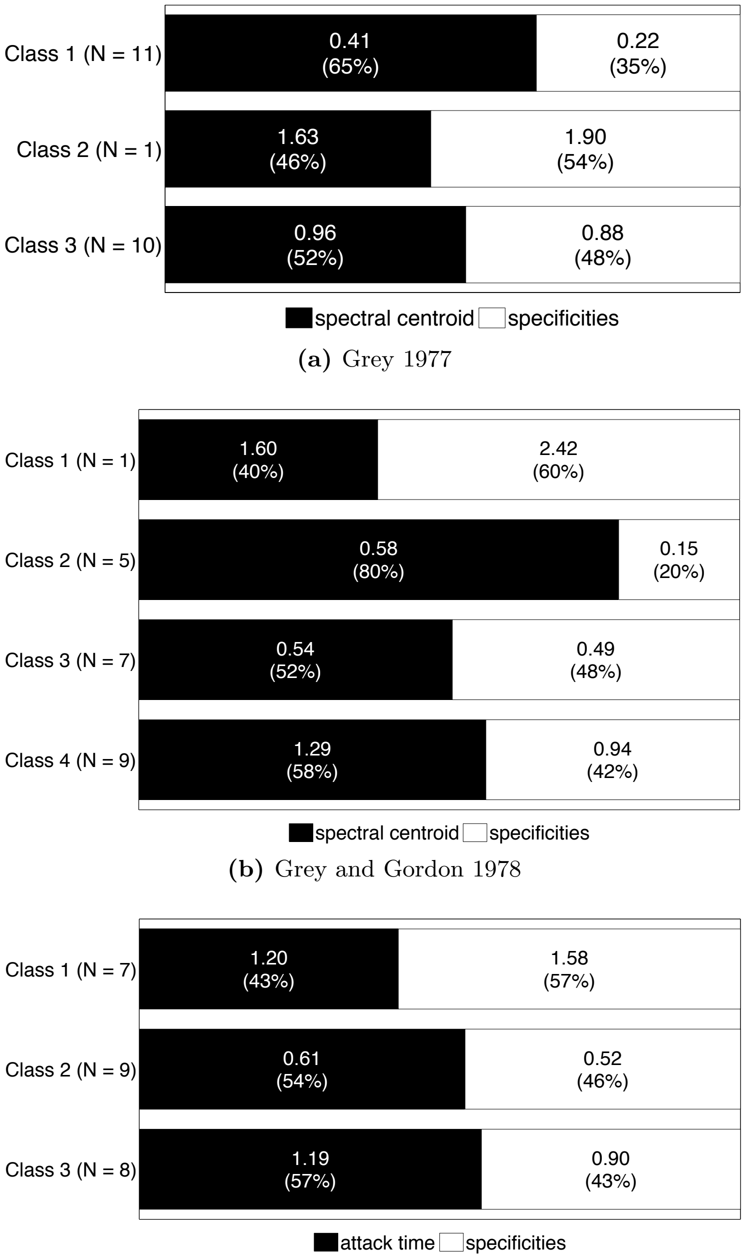
Fig. 9. Grey 1977 timbre space before Isomap and including specificities. Structurally, it is nearly identical to the analogous space without specificities.

One can see from inspecting the specificity values in the co-ordinate listings for all three spaces that it is difficult to interpret the precise meaning of specificities. Higher values denote timbres that are heard as more unique than others, but other than an informal analysis like that in McAdams, Winsberg, Donnadieu, Soete, and Krimphoof 1995, there is no scientific means to determine from our data what makes each timbre sound unique. We do find, however, that certain specificity dimensions (vectors valued at zero for all instruments but one) correlate at $p < 0.01$ with certain acoustic features, e.g., the string sound in McAdams et al. 1995 with spectral flux ($r = 0.75$). These correlations are undesirable but strictly dependent on the choice of stimulus set; confirmatory studies with artificial timbres, e.g., Caclin et al. 2005, can and should check for such correlations before conducting rating experiments.

4 Discussion and Future Work

4.1 Perceived Dimensions of Timbre

Consistent with previous studies, it is clear from all of the spaces presented above that humans use log attack time and spectral centroid when distinguishing between timbres. It is also clear that we use at least one other component, which manifests itself in the above spaces as spectral shape (spectral slope, spectral spread), perceived roughness, or specificities. Whenever specificities are included in the models above, however, the spectral shape dimensions disappear. This behaviour suggests that all of these spectral shape dimensions are poor approximations of the elusive third component of timbre, so poor that in most



(a) Grey 1977

(b) Grey and Gordon 1978

(c) McAdams et al. 1995

Fig. 10. Raw and relative weights for each latent class after processing with Isomap and including specificities in the model

cases, it is statistically advantageous to avoid the approximation by reverting to specificities. Specificities are a sort of null model in that they take up as much unexplained variance as possible without making any assumptions about the relationships among timbres.

Ultimately, these results should not be surprising. Practitioners of sound synthesis are well aware of the trade-off between modelling a given spectral shape precisely and keeping the number of parameters manageable, e.g., basic FM synthesis produces a coarser approximation of a desired spectrum but requires many fewer parameters than an additive synthesis model; subtractive synthesis models would fall somewhere in between. In no case will a single parameter provide a meaningful approximation. Moreover, interpolating between synthesis parametrisation when replicating time-varying spectra can be quite challenging because the spaces involved are generally not Euclidean. It would be well worth investigating whether the higher dimensions of timbre perception or the specificities are ultimately correcting for problems with the assumption that it is possible to represent timbre space as a Euclidean one, which is the default assumption in CLASCAL and its relatives.

4.2 Isomap Processing

Although Isomap was ultimately successful in reducing the output dimensionality, it was not able to do so without resorting to specificities. In the models without specificities, Isomap increases the models' degrees of freedom (the total number of free parameters accounting for all dimensions and classes) in the Grey 1977 space from 50 to 71, in the Grey and Gordon 1978 space from 37 to 58, and in the McAdams et al. 1995 space from 56 to 79. Another disadvantage of Isomap, before specificities are considered, is that it can yield dimensions that are difficult to interpret. The Grey 1977 and McAdams et al. 1995 spaces include duplicate spectral-centroid dimensions and the Grey and Gordon 1978 space includes the curious pseudo-specificity in its third dimension. Furthermore, the correlation between dimensions pre- and post-Isomap can be surprising, as discussed above for the McAdams et al. 1995 space. It is impossible to know exactly what about the geometry of the timbres in this space causes these behaviours, but the geometric structures it uncovers are consistent with those of the spaces before Isomap processing.

These observations would be a blacker mark against Isomap processing for MDS studies if Monte Carlo testing did not always prefer models with specificities. Fortunately, after the introduction of specificities, the Isomap-processed spaces can reduce dimensionality greatly and become very easy to interpret. The necessary degrees of freedom for an optimal fit reduce from 67 to 38 for Grey 1977 and from 56 to 42 for McAdams et al. 1995; for Grey and Gordon 1978, the necessary degrees of freedom increase just slightly from 37 to 41. For each of the timbre sets considered, the ideal output space consists of a single dimension correlating with the leading dimensions of the pre-Isomap spaces. The trade-off for this heavy degree of compression is the dispersion of remaining variance among the specificities, which are difficult to interpret directly but, as discussed above, may be a statistical proxy for variations in spectral shape. A problem with this severe reduction, however, is that it conceptualises timbre as unidimensional with specificities, whereas many musicians feel that two to four dimensions already lose much of the subtle richness of timbral experience available in the instrument sounds tested.

In generating its geodesic distances, Isomap prefers small dissimilarities to larger ones; by construction, dissimilarities in the transformed matrix are always greater than or equal to dissimilarities in the original matrix. We chose this particular transformation because of its natural combination with the CLASCAL method and relative ease of computation. It is unclear exactly what such a transformation is assuming about subjects' rating strategies, however, other than a larger trust in estimations of dissimilarity between relatively similar timbres over estimations between relatively dissimilar timbres. One improvement would be to incorporate methods such as maximum variance unfolding (MVU) (Weinberger 2004; Weinberger and Saul 2006) that can emphasise particular sets of neighbouring stimulus pairs without requiring the recomputation of other dissimilarities. Such methods could be extended to allow researchers to focus not only on small dissimilarities in the matrix, but possibly on exclusively mid-range dissimilarities or exclusively large dissimilarities as well, which would allow studies to identify differing rating strategies for differing grades of dissimilarity.

Geodesic distances are a means not an end: like MVU, Isomap seeks to identify low-dimensional manifolds in high-dimensional structures. Curiously, an informal analysis suggests that Isomap has no effect on dimensionality or the interpretation of the dimensions when applied *after* CLASCAL analysis, which suggests that for our study, the primary value of Isomap was in eliminating complicated nonlinearities in individual subjects' rating strategies and reducing them to more broadly used acoustic features such as log attack time or spectral centroid. Because the dimensions that disappear on account of Isomap processing all have to do with spectral shape, it is reasonable to assume that many of these nonlinearities are connected to spectral shape in some way. Further confirmatory studies are warranted to explore exactly how Isomap (or its relatives) warp timbre dissimilarity matrices.

5 Conclusion

Designed for uncovering the true dimensionality of Euclidean manifolds, Isomap is also able to simplify the timbre spaces resulting from MDS on empirical timbre dissimilarity matrices. These simplifications are in one sense disappointing: they collapse the spaces to a single shared dimension plus a set of instrument-specific dimensions that are relatively difficult to interpret. These simplified spaces, however, confirm two known dimensional components of timbre, spectral centroid and log attack time, and highlight an important direction for future work on the perception of spectral shape. More generally, the success of Isomap in this domain should encourage all researchers using MDS models to explore how preprocessing dissimilarity matrices before MDS could be valuable and, in particular, how Isomap preprocessing could be tailored to their needs.

Acknowledgements

This research was supported in part by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada to Stephen McAdams. We would also like to thank Bruno Giordano for his insights.

References

[Aitkin, Anderson, and Hinde 1981]Aitkin, M., Anderson, D., Hinde, J.: Statistical modelling of data on teaching styles. Journal of the Royal Statistical Society, Series A (General) 144(4), 419–461 (1981)

[von Bismarck 1974]Bismarck, G., von, G.: Sharpness as an attribute of the timbre of steady sounds. Acustica 30, 159–172 (1974)

[Burgoyne and Saul 2005]Burgoyne, J.A., Saul, L.K.: Visualization of low-dimensional structure in tonal pitch space. In: Proceedings of the International Computer Music Conference, pp. 243–246 (2005)

[Caclin, McAdams, Smith, and Winsberg 2005]Caclin, A., McAdams, S., Smith, B.K., Winsberg, S.: Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. Journal of the Acoustical Society of America 118(1), 471–482 (2005)

[Carroll and Chang 1970]Carroll, J.D., Chang, J.-J.: Analysis of individual differences in multidimensional scaling via an n-way generalization of 'Eckart-Young' decomposition. Psychometrika 35(3), 283–319 (1970)

[Gabriel and Sokal 1969]Gabriel, K.R., Sokal, R.R.: A new statistical appraoch to geographic variation analysis. Systematic Zoology 18, 259–270 (1969)

[Grey 1977]Grey, J.M.: Multidimensional perceptual scaling of musical timbre. Journal of the Acoustical Society of America 61, 1270–1277 (1977)

[Grey and Gordon 1978]Grey, J.M., Gordon, J.W.: Perceptual effects of spectral modifications on musical timbres. Journal of the Acoustical Society of America 63(5), 1493–1500 (1978)

[Hope 1968]Hope, A.C.A.: A simplified Monte Carlo significance test procedure. Journal of the Royal Statistical Society, Series B (Methodological) 30(3), 582–598 (1968)

[Jaromczyk and Toussaint 1992]Jaromczyk, J.W., Toussaint, G.T.: Relative neighborhood graphs and their relatives. Proceedings of the IEEE 80(9), 1502–1517 (1992)

[Krumhansl 1989]Krumhansl, C.L.: Why is musical timbre so hard to understand? In: Nielzen, S., Olsson, O. (eds.) Structure and Perception of Electroacoustic Sound and Music. Excerpta Medica, vol. 846. Elsevier, Amsterdam (1989)

[McAdams, Winsberg, Donnadieu, Soete, and Krimphoof 1995]McAdams, S., Winsberg, S., Donnadieu, S., Soete, G.D., Krimphoof, J.: Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. Psychological Research 58, 177–192 (1995)

[Peeters, McAdams, and Herrera 2000]Peeters, G., McAdams, S., Herrera, P.: Instrument sound description in the context of MPEG-7. In: Proceedings of the International Computer Music Conference (2000)

[Tennenbaum, de Silva, and Langford 2000]Tennenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323 (2000)

[Torgerson 1958]Torgerson, W.S.: Theory and Methods of Scaling. Wiley, Chichester (1958)

[Weinberger 2004]Weinberger, K.Q.: Unsupervised learning of image manifolds by semidefinite programming. In: Proceeding of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2004)

[Weinberger and Saul 2006]Weinberger, K.Q., Saul, L.K.: An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In: Proceedings of the National Conference on Artificial Intelligence (AAAI) (2006)

[Wessel, Bristow, and Settel 1987]Wessel, D.L., Bristow, D., Settel, Z.: Control of phrasing and articulation in synthesis. In: Proceedings of the International Computer Music Conference, pp. 108–116 (1987)

[Winsberg and Carroll 1989]Winsberg, S., Carroll, J.D.: A quasi-nonmetric method for multidimensional scaling via an extended Euclidean model. Psychometrika 54(2), 217–229 (1989)

[Winsberg and De Soete 1993]Winsberg, S., De Soete, G.: A latent class approach to fitting the weighted Euclidean model, CLASCAL. Psychometrika 58(2), 315–330 (1993)

[Winsberg and De Soete 1997]Winsberg, S., De Soete, G.: Multidimensional scaling with constrained dimensions: CONSCAL. British Journal of Mathematical and Statistical Psychology 50, 55–72 (1997)