

GEOMETRIC ICA USING NONLINEAR CORRELATION AND MDS

Samer A. Abdallah and Mark D. Plumbley

Department of Electronic Engineering,
Queen Mary, University of London.
samer.abdallah@elec.qmul.ac.uk
mark.plumbley@elec.qmul.ac.uk

ABSTRACT

We describe a method of visualising geometrically the dependency structure of a distributed representation. The mutual information between each pair of components is estimated using a nonlinear correlation coefficient, in terms of which a distance measure is defined. Multidimensional scaling is then used to generate a spatial configuration that reproduces these distances, the end result being a spatial representation of the dependency between the components, from which an appropriate topology for the representation may be inferred. The method is applied to ICA representations of speech and music.

1. INTRODUCTION

Topographic independent component analysis [1] is concerned with the organisation of a distributed representation, in this case a linear basis decomposition generated by ICA, to reflect some notion of similarity between the elements of the representation. Hyvärinen et al. [1] proposed that these similarity relationships should be defined in terms of statistical dependencies between the elements. Since ICA is specifically intended to minimise these dependencies, they may be termed ‘residual,’ indicating that the ICA algorithm has not fully succeeded, and that further processing is required to achieve the goal of non-redundant factorial representation, as advocated, e.g. by Attneave, [2] as a fundamental objective for perceptual systems. Localising the residual dependencies may facilitate these subsequent processes by allowing certain computations to be implemented more locally.

Topological organisation is also relevant for other representations not generated by ICA: for example, one would intuitively arrange a Fourier representation in one dimension ordered by frequency. Indeed, a salient point about auditory spectrograms is that activity tends to be *localised*, so that coherent features emerge and

are recognisable as such, suggesting that the principle of dependency localisation is at work here too.

In this paper, we investigate the use of residual dependency to generate, not a topology, but a *geometry*, by defining a system of distances in terms of the estimated mutual information between every pair of elements in a given representation. This process is independent of that used to generate the representation, and so may, in principle, be applied to any distributed representation, not just those generated by ICA. The overall procedure may summarised as follows:

1. Define or otherwise generate the initial representation. In the present work, ICA will be used under the tacit assumption that it yields an optimal linear encoding of the data.
2. Estimate the mutual information (MI) between every pair of elements using a method appropriate to the observed statistical structure.
3. Convert the MI estimates into distances in such a way that high dependency implies proximity.
4. Generate a spatial realisation of the induced metric in a Euclidean embedding space using multidimensional scaling (MDS).

2. ICA REPRESENTATIONS OF SPEECH AND MUSIC

We took as our starting point the results of a previous experiment with ICA of speech and music audio signals [4, 3]. The signals were broken up into a sequence of 512-sample vectors \mathbf{x} , and represented as the vectors $\mathbf{s} = \mathbf{W}\mathbf{x}$. The square matrix \mathbf{W} was adapted using the following update rule [5]:

$$\mathbf{W} \mapsto \mathbf{W} + \eta [\mathbf{I} - \langle \gamma(\mathbf{s})\mathbf{s}^T \rangle] \mathbf{W}, \quad (1)$$

where η is a learning rate parameter, and the angle brackets denote a sample average over a batch of training data. The vector-valued function $\gamma : \mathbb{R}^n \mapsto \mathbb{R}^n$

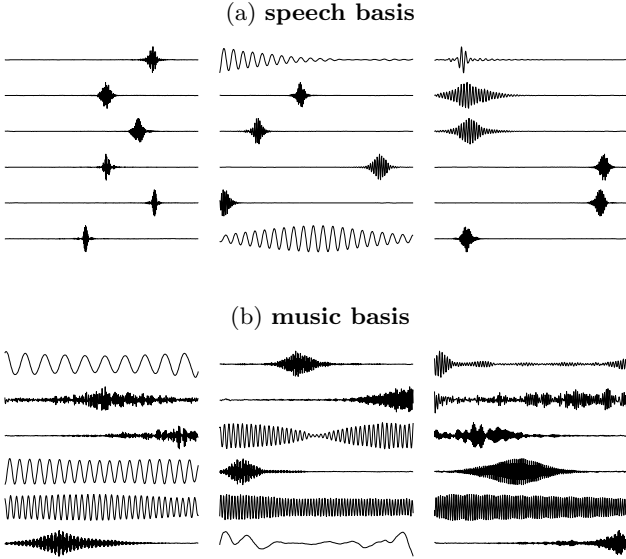


Figure 1: Some of the basis vectors obtained by ICA of (a) speech and (b) music. (Each basis has 512 vectors in total.)

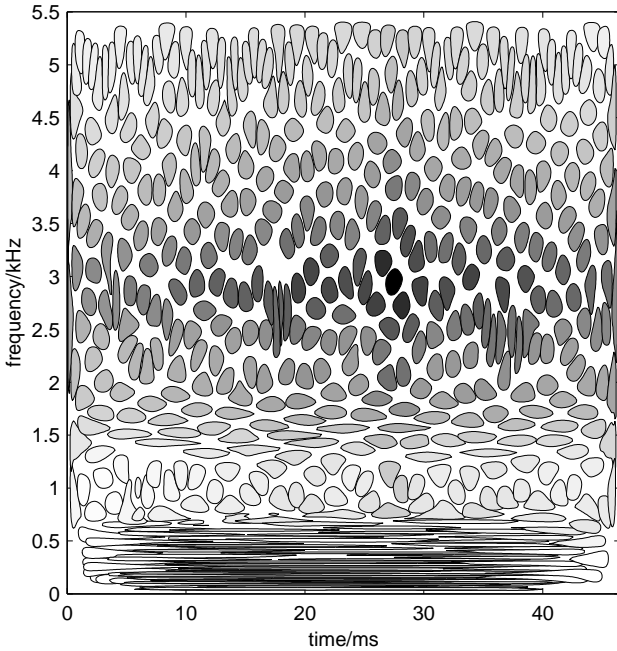


Figure 2: Speech derived ICA basis: each of the 512 basis vectors is plotted as a region in time-frequency, obtained by taking a contour of the Wigner distribution of that basis vector (see [3] for more details.) The shading encodes the nonlinear correlation (see § 3) between each component and the one at approximately (28 ms, 3 kHz), showing that the dependencies are local in time-frequency and that it should be possible to recover the time-frequency organisation through an analysis of the dependency structure.

is defined element-wise as $\boldsymbol{\gamma}(\mathbf{s}) \equiv (\gamma(s_1), \dots, \gamma(s_n))$, where γ is a real-valued function. If the components s_i are assumed to have a non-Gaussian marginal probability density function $p(s)$, then γ is the nonlinear score function $\gamma(s) = -(d/ds) \log p(s)$. In our experiments, a Cauchy prior was used:

$$p(s) = \frac{1}{\pi(1+s^2)}, \quad \gamma(s) = \frac{2s}{1+s^2}. \quad (2)$$

Some of the ICA basis vectors, given by the columns of $\mathbf{A} = \mathbf{W}^{-1}$, are illustrated in fig. 1.

The speech-derived basis vectors are well localised in both time and frequency, and form a neat tiling of the time-frequency plane, as shown in fig. 2. We might expect this two-dimensional organisation to emerge in the subsequent analysis, and indeed it does.

3. ESTIMATING MUTUAL INFORMATION USING NONLINEAR CORRELATION

ICA tends to produce decorrelated outputs, and so pair-wise correlations will not generally be a useful measure of statistical dependency. Instead, Hyvärinen et al. [1] suggest that the dependencies can be captured as energy correlations, that is, as correlations between the *squares* of the components. However, estimates of second-order statistics of the squares of what are already highly super-Gaussian variables [3, ch. 5] are likely to be poorly defined with high asymptotic variance. Hence, we propose an alternative based on nonlinearly rectified activity correlations.

Consider two random variables U and V which are jointly Gaussian. Their correlation coefficient is defined as

$$\text{corr}[U, V] \stackrel{\text{def}}{=} \frac{\text{cov}[U, V]}{\sqrt{\text{var } U \text{ var } V}}, \quad (3)$$

where $\text{cov}[\cdot, \cdot]$ denotes the covariance and $\text{var} \cdot$ the variance. The mutual information (MI) can then be expressed in terms of the correlation coefficient only as

$$I(U, V) = -\frac{1}{2} \log [1 - (\text{corr}[U, V])^2]. \quad (4)$$

If U and V are not jointly Gaussian, then (4) will not hold. However, since the MI is invariant to invertible transformation of the variables [6], if two invertible functions f and g could be found such that $f(U)$ and $f(V)$ were jointly Gaussian, then we would have $I(U, V) = I(f(U), g(V))$, which could be expressed in terms of the nonlinear correlation $\text{corr}[f(U), g(V)]$. Thus, we conjecture that for *any* two invertible functions f and g , the nonlinear correlation provides a lower bound on the mutual information:

$$I(U, V) \geq -\frac{1}{2} \log \{1 - (\text{corr}[f(U), g(V)])^2\}, \quad (5)$$

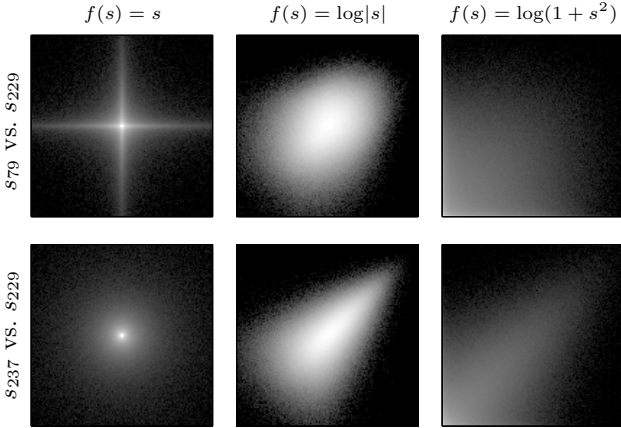


Figure 3: Joint histograms of $f(s_i)$ vs. $f(s_j)$ for three functions f and two pairs of components from the music-derived ICA representation. Both s_{237} and s_{229} (bottom row) are essentially sinusoidal components at 937 Hz, but in quadrature phase: they have a circularly symmetric distribution and are strongly dependent, though uncorrelated. In the top row, s_{79} , a sinusoidal component at 883 Hz, is almost independent of s_{229} .

with equality being reached as $f(U)$ and $g(V)$ approach joint Gaussianity.

Now suppose that the joint density $p(u, v)$ is symmetric in both axes, (as those in the first column of fig. 3 are) with $p(u, v) = p(\pm u, \pm v)$. For such variables, it can easily be shown that the MI is invariant to rectification, with $I(U, V) = I(|U|, |V|)$, in which case the above conjecture should extend to even functions f and g which are invertible on \mathbb{R}^+ , for example, $f(u) = u^2$, which corresponds to measuring energy correlations. In practice, these energies tend to be extremely super-Gaussian, which means both that the bound in (5) may be a poor one, and that estimates of $\text{corr}[U^2, V^2]$ will have high variance. Instead, a *compressive* rectification is required, such as $f(u) = g(u) = \log|u|$, so that $f(U)$ and $g(V)$ are closer to being Gaussian and the nonlinear correlation provides a more accurate picture of their mutual dependence.

Validation The procedure was tested using synthetic data generated by nonlinear transformation of a bivariate Gaussian (z_1, z_2) of known correlation:

$$s_1 = e^{\lambda z_1}, \quad s_2 = e^{\lambda z_2}, \quad \text{corr}[z_1, z_2] = \rho, \quad (6)$$

with $\lambda = 1.5$. For each of several values of ρ , the nonlinear correlation $\text{corr}[f(s_1), f(s_2)]$ was estimated using a number of functions f . Each correlation was measured 160 times using a sample size of 1600, so that the sample mean and standard deviation of the estimates could be computed. The results are illustrated in fig. 4. All

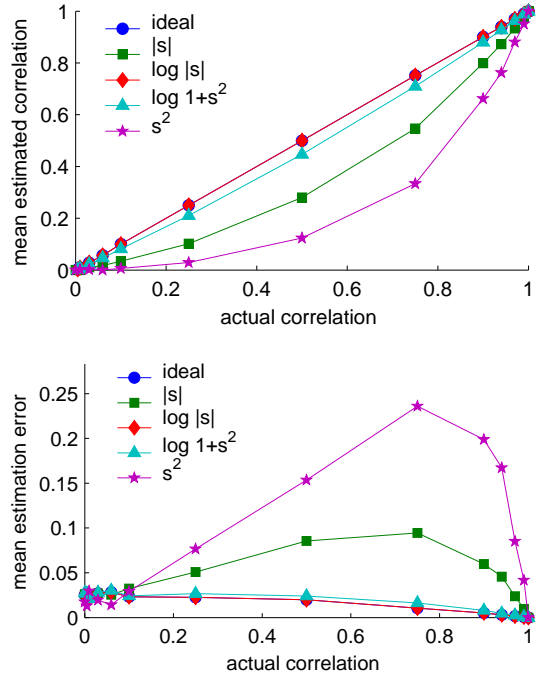


Figure 4: Nonlinear correlation estimates from synthetic data, generated by nonlinear transformation of Gaussian variables with a known correlation coefficient. The points marked ‘ideal’ show the mean and standard deviation of the correlation estimated directly from the Gaussian source data. The other points indicate estimates obtained using several nonlinear correlation measures.

the correlation estimates are smaller than ρ , the true correlation of the Gaussian source, so the inequality (5) holds for this data. As expected, the correlation estimates obtained with $f(s) = s^2$ and $f(s) = |s|$ have the highest variance and are therefore the least reliable.

The motivation behind the use of $f(s) = \log 1 + s^2$ as a nonlinearity is this: if the observed components s_i have added uncorrelated noise, then small activations due to the noise will be greatly amplified in $\log s_i$. The nonlinear correlation will tend to underestimate the mutual information between the components, as illustrated in fig. 5. Using $f(s) = \log 1 + s^2$ tends to counteract this effect by ‘squashing out’ the noise. We suggest that $f(s) = -\log p(s)$ be used generally, where $p(s)$ is the estimated marginal density of the components. When this density is sharply peaked, we assume that there is little noise, but when the peak is rounded, we assume that this is due to noise, and $f(s)$ will attenuate values near the peak. In this way, the nonlinearity is tailored to the observed distribution of the data.

Results As a further validation with real data, the mutual information between two pairs of components obtained from the music-derived ICA representation

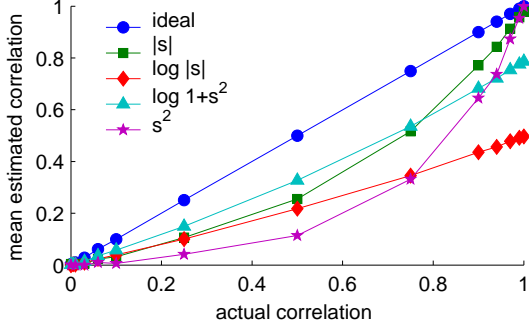


Figure 5: Nonlinear correlations estimated from synthetic data with additive noise; that is, $s_i = e^{\lambda z_i} + \epsilon_i \sqrt{2}$ where the ϵ_i are Gaussian, uncorrelated, and of unit variance.

$f(s)$	s	$\log s$	$\log 1+s^2$	s^2
<i>Almost independent components</i>				
correlation	0.000	0.0652	0.0419	0.006
direct	0.0457	0.0746	0.0685	0.185
<i>Dependent components</i>				
correlation	0.000	0.3228	0.4573	0.2663
direct	0.3282	0.4089	0.4022	0.1945

Table 1: Estimates of MI (in bits) computed by two methods from the histograms in fig. 3. Correlation coefficients were computed from the histograms, not from the original samples; (4) was then used to estimate the MI. The ‘direct’ method is the standard summation for discrete random variables [6], treating the histogram as a discrete probability distribution.

was estimated using linear and nonlinear correlations and by directly summation of the joint histograms shown in fig. 3; the results are shown in table 1. The linear correlations are minimal, as expected, but the nonlinear correlations agree fairly well with the direct estimates. In the case of the two dependent components, the nonlinear correlation measure using $f(s) = \log 1 + s^2$ gives a higher mutual information than the direct estimate, apparently refuting our earlier conjecture (5). It must be noted, however, that the direct method is only an approximation computed from a noisy discrete approximation to the underlying continuous density.

The full matrix of nonlinear correlations was measured for both the speech and the music ICA representation, using $f(s) = \log 1 + s^2$. The music-derived correlation matrix is shown in fig. 6, though in the figure, the matrix has been rearranged to clarify its structure.

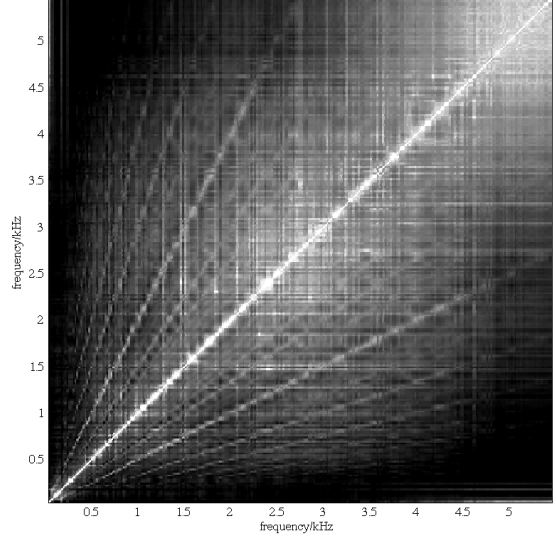


Figure 6: Matrix of nonlinear correlation coefficients obtained with the music-derived basis, with rows and columns reordered and scaled by centre frequency to show strong correlations between frequencies at small whole number ratios. There is also a slight 12-cycle per octave ‘ripple’ effect visible, which is probably due to the semitone quantisation of the Western musical scale.

4. DEFINITION OF DISTANCE MEASURE

The mapping from the mutual information $I(U, V)$ between two random variables to a distance $d(U, V)$ is to some extent arbitrary as long as the distance is a monotonically decreasing function of the MI with $d \rightarrow 0$ as $I \rightarrow \infty$ and $d \rightarrow \infty$ as $I \rightarrow 0$. It may also be desirable for the derived distances to satisfy the triangle inequality (see, e.g., [7]) so that the result qualifies as a metric or a pseudometric, but we do not consider this here. We used the following mapping:

$$d^2(U, V) = -\log \left\{ 1 - e^{-2I(U, V)} \right\}, \quad (7)$$

which was arrived at by considering an hypothetical system with a local Gaussian noise process, and is described elsewhere [3].

Using (4) as an approximation to the mutual information, the full set of pair-wise distances can be written in terms of the matrix of nonlinear correlation coefficients $R_{ij} = \text{corr}[f(s_i), f(s_j)]$, as $D_{ij} = \sqrt{\log R_{ij}^{-2}}$.

5. GEOMETRY BY MDS

We now have a number of objects and a distance between every pair of objects, and we would like to find a configuration of points in a metric space which have the

same pair-wise distance relationships. This is precisely the problem that multidimensional scaling or MDS [8] addresses. More formally, given a distance matrix D_{ij} , a metric space E and a metric $d_E : E \times E \mapsto \mathbb{R}$, we seek n points $x_i \in E$ such that $d_E(x_i, x_j) = D_{ij}$ for all pairs i, j .

Generally, MDS is formulated as an optimisation problem where the goal is to minimise a certain *stress* function, which quantifies the degree of mismatch between the spatial configuration and the target distances. There are some subtleties involved in the choice of stress function, relating to the expected errors in distance estimation; these are discussed further in [3], but in this paper we present results obtained using the following stress function only:

$$J = \sum_{i < j} \left[\frac{d_E(x_i, x_j) - D_{ij}}{D_{ij}} \right]^2. \quad (8)$$

The MDS analysis was performed on the two distance matrices obtained for speech and music data, using Euclidean embedding spaces of dimension 2 to 8; figures 7 and 8 show some of the 2 and 3 dimensional results.

In 2-D, the speech basis forms what is essentially a time-frequency manifold, with the components arranged according to their time-frequency localisation as they are in fig. 2. In 3-D, this 2-D manifold persists but takes on a folded shape, seen edge-on in fig. 8(a).

The music-derived representation produces a 3-D configuration similar to a frustrum of cone, or ‘cork-shape,’ with low frequency components at the broad end and high frequency components at the other. Experiments with probe tones suggest that components tuned to the 12-semitone Western scale are approximately localised at the surface of the ‘cork,’ with intermediate tones towards the interior, and that there is an approximate ‘circle-of-fifths’ arrangement of tones around the circumference. These and other harmonic relationships are visible as the straight lines of various slopes in the correlation matrix (fig. 6).

6. RELATIONS WITH OTHER METHODS

The method presented here is most closely related to topographic ICA [1] which also uses residual dependency to define similarity relationship in a distributed representation. One important difference is that topographic ICA requires a predefined topology, into which the ICA basis is fitted, whereas the present geometric approach allows the topology to emerge from the data. The speech basis, for example, maintains its 2-D topology even in a 3-D embedding space. Another difference is that the present method can also be applied post-hoc

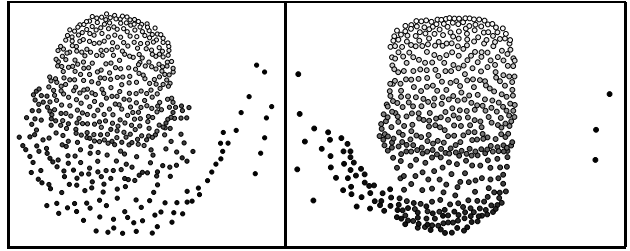


Figure 7: MDS solutions in 2-D obtained for music (left) and speech (right) derived representations. The gray scale indicates the nominal centre frequency of each basis vector, with low frequencies in black at the bottom of the figures. Further analysis of the speech based configuration makes it clear that it is essentially a *time-frequency* representation, with the time-axis horizontal in this figure. The ‘tail’ extending to the left contains low-frequency units that are not localised in time.

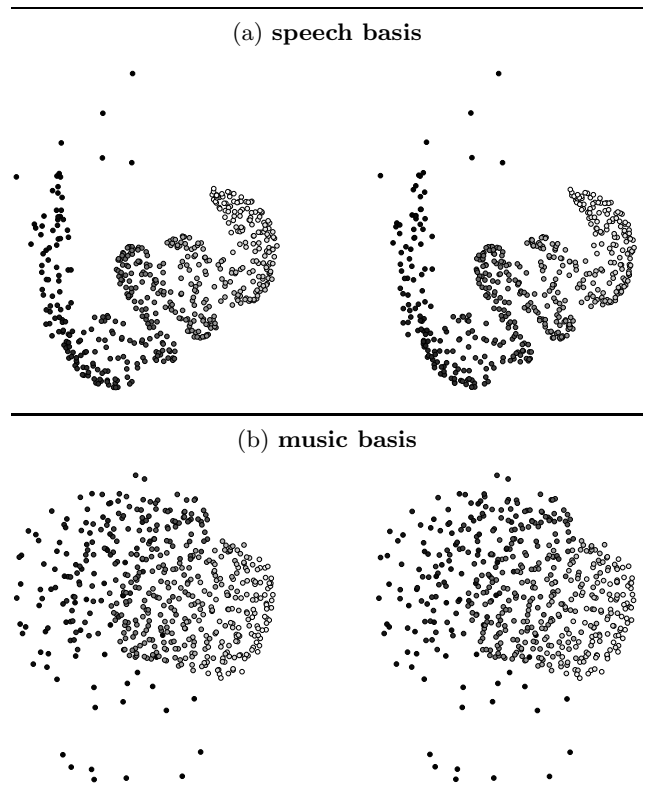


Figure 8: Stereo pairs of 3-D MDS results for speech basis (top) and music basis (bottom). The left-eye images are on the left. The grey scale encodes the estimated centre frequency of each unit. In three dimensions, the time-frequency plane visible in fig. 7 becomes a curved manifold, viewed edge-on in this figure, so that time axis is perpendicular to the page. The music representation is approximately a frustrum of cone, with its axis pointing right, slightly down, and slightly out of the page.

to any distributed representation as long as a suitable method for estimating the residual dependencies can be found.

It may be though that Kohonen's self-organising map [9] achieves a comparable result, but the similarity is only superficial. The SOM algorithm finds low-dimensional manifolds in high-dimensional data, but the data *already* has a fully specified geometric structure, in terms of coordinates in a Euclidean space. By contrast, the present method *generates* the coordinates (using MDS) from a set of distances. The objects being arranged (in this case, the components of a distributed representation) need have no pre-existing representation in a metric space.

There is also a more fundamental difference: each unit in a Kohonen map represents an observation in its entirety, and distances in the Kohonen map represent similarity between patterns, or points in the input space. Both topographic ICA and the present system work with distributed-activity representations; there is no one-to-one mapping between points in the map and points in the data space, and distances in the map do not represent similarity between patterns. Instead, these distances represent a different kind of similarity, which is discussed in more detail in [3, ch. 8].

Hyvärinen et al. [1] have noted a relationship between topographic ICA and independent subspace analysis [10], which carries over into the present system: if the components produced by ICA fall into several independent subspaces, these should appear as widely separated clusters in the MDS solution. This could be used to determine the appropriate dimensionality for the subspaces. It should also be possible to find these clusters by a direct analysis of the nonlinear correlation matrix without actually performing the MDS.

7. CONCLUSIONS AND FURTHER WORK

The method of nonlinear correlations is clearly capable of capturing some of the residual dependencies in the ICA representations examined here, producing some nontrivial geometric structures. The speech-based representation form what is essentially a time-frequency manifold, in which both the basis and the geometric configuration were developed in an entirely data-driven way. The the music based representation is quite different and reflects the harmonic and scalar structure of Western music.

Two areas that requires further investigation are the mapping from mutual information to distance, and the choice of stress function to use in the MDS algorithm. In both cases, it may be possible to find more effective alternatives to the ones used here.

The purpose of the geometric mapping procedure is not merely visualisation, but to allow the implementation of meaningful 'local' computations in the MDS embedding space; we are currently investigating several possibilities, such as spatial filtering and local divisive normalisation [11].

8. REFERENCES

- [1] Aapo Hyvärinen, Patrik Hoyer, and Mika Inki, "Topographic independent component analysis," *Neural Computation*, vol. 13, no. 7, pp. 1527–1558, 2001.
- [2] Fred Attneave, "Some informational aspects of visual perception," *Psychological Review*, vol. 61, no. 3, pp. 183–193, 1954.
- [3] Samer A. Abdallah, *Towards Music Perception by Redundancy Reduction and Unsupervised Learning in Probabilistic Models*, Ph.D. thesis, Department of Electronic Engineering, King's College London, 2002.
- [4] Samer A. Abdallah and Mark D. Plumbley, "If edges are the independent components of natural scenes, what are the independent components of natural sounds?," in *3rd Intl. Conf. on Independent Component Analysis and Signal Separation, ICA2001*, San Diego, 2001, pp. 534–539.
- [5] Jean-François Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. on Signal Processing*, vol. 44, no. 12, pp. 3017–30, Dec. 1996.
- [6] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, John Wiley and Sons, New York, 1991.
- [7] G. J. O. Jameson, *Topology and Normed Spaces*, Chapman and Hall, London, 1974.
- [8] Trevor Cox and Michael A. A. Cox, *Multidimensional Scaling*, Chapman Hall/CRC, London, 2001.
- [9] Teuvo Kohonen, *Self-Organizing Maps*, Springer, Berlin, 1995.
- [10] Aapo Hyvärinen and Patrik Hoyer, "Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces," *Neural Computation*, vol. 12, no. 7, pp. 1705–1720, 2000.
- [11] Odelia Schwartz and Eero P. Simoncelli, "Natural sound statistics and divisive normalisation in the auditory system," in *Advances in Neural Information Processing Systems*, Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, Eds., 2001, vol. 13, pp. 166–172.