NONLINEAR DIMENSIONALITY REDUCTION APPROACHES APPLIED TO MUSIC AND TEXTURAL SOUNDS

Stéphane Dupont, Thierry Ravet, Cécile Picard-Limpens, Christian Frisson

NUMEDIART Institute for New Media Art Technology
University of Mons (Belgium)
31 Boulevard Dolez, B-7000 Mons, Belgium

ABSTRACT

Recently, various dimensionality reduction approaches have been proposed as alternatives to PCA or LDA. These improved approaches do not rely on a linearity assumption, and are hence capable of discovering more complex embeddings within different regions of the data sets. Despite their success on artificial datasets, it is not straightforward to predict which technique is the most appropriate for a given real dataset. In this paper, we empirically evaluate recent techniques on two real audio use cases: musical instrument loops used in music production and sound effects used in sound editing. ISOMAP and t-SNE are being compared to PCA in a visualization problem, where we end up with a two-dimensional view. Various evaluation measures are used: classification performance, as well as trustworthiness/continuity assessing the preservation of neighborhoods. Although PCA and ISOMAP can yield good continuity performance even locally (samples in the original space remain close-by in the low-dimensional one), they fail to preserve the structure of the data well enough to ensure that distinct subgroups remain separate in the visualization. We show that t-SNE presents the best performance, and can even be beneficial as a pre-processing stage for improving classification when the amount of labeled data is low.

Index Terms— Dimensionality reduction, multimedia information retrieval, audio and music analysis.

1. INTRODUCTION

Current trends in information technologies underlie the exponential growth in digital multimedia data collections. At the same time, increasingly diverse approaches for extracting descriptors characterizing the content items are being researched and developed, leading to very high-dimensional representations. Dimensionality reduction techniques fortunately offer various opportunities. These can be used as a pre-processing step prior to the application of further machine learning techniques such as classification, as well as for generating visualizations of the data set while preserving as much of the structures present in the high-dimensional space, such as class membership as well as the main intra-class variation factors. In this paper, we focus on the second target, which offers potential in various forms of entertaining or useful exploration of multimedia content databases.

The potentially non-linear structure of the manifolds close to which the data is lying in the high-dimensional space and the spatially local specificity of principal components naturally triggered research in the area of non-linear embedding methods. Also, to be perceptible by the human, representations need to be of very low dimension, further justifying non-linear techniques. In recent years, a range of such approaches have hence been proposed, some designed to preserve distances on non-linear manifolds, such as Isometric Feature Mapping (ISOMAP [1]), some better able to preserve neighborhoods, such as Stochastic Neighborhood Embedding (SNE [2]). Many techniques have been published. We direct the reader to [3, 4]. They propose recent reviews of dimensionality reduction techniques in unsupervised settings. Existing techniques essentially differ in the properties of the high-dimensional data that they attempt to preserve, and in their use of class labels or not (supervised or unsupervised).

Despite the high level of research activity, there is a saying that goes like "always try PCA first", and some authors [4] even highlight the fact that advanced method do not outperform the traditional PCA on many real-world tasks. Nevertheless, observing the literature, there seem to be many real-world tasks where those techniques are beneficial, although it remains unclear which method is more appropriate to specific use cases. Recent dimensionality reduction techniques have scarcely been applied to audio material. In [5], SNE is used to create a 2D intuitive screen-based interface representing large collections of textural sounds. However, no comparison with other approaches is proposed as the evaluation is rather related to user experience of a complete sound

This research has received funding from Région Wallonne (Belgium) through the Numediart long-term research programme (grant nbr. 716631), and the MediaWorkflows project (grant nbr. 1117549).

Possibly, problems where the structure of the underlying low-dimensional manifolds are well covered by the high dimensional data set will behave well.

"browser" tool. In [6], several techniques are applied as preprocessing step prior to music genre classification relying on k-NN. It is shown that t-SNE significantly outperforms the competitors when the number of retained dimensions is low.

In this paper, we apply some of the recent non-linear methods to two audio content use cases of interest to the creative community: one relying on musical instrument loops used in rhythmic music composition/production, and another one relying on sound effects used in sound editing. Section 2 present the dimensionality reduction approaches used in the study: PCA, ISOMAP and t-SNE, an improved variant of SNE. Section 3 presents the experimental protocol and evaluation metrics, as well as the experimental results, together with a discussion. We conclude in Section 4.

2. DIMENSIONALITY REDUCTION APPROACHES

In this work, we have been comparing the unsupervised methods PCA, ISOMAP and t-SNE, which have been reimplemented in C language. Building on the classical method of multi-dimensional scaling (MDS), ISOMAP attempts to preserve the pairwise distances between data samples as much as possible, with the additional refinement that it takes into account the distribution of neighboring data points. On the other side, t-SNE, a variation of SNE, is part of a class of techniques that rather attempt to preserve the whole structure of neighborhoods. Both methods are summarized below.

2.1. MDS and ISOMAP

The MDS approach suffers from the use of Euclidean distances between data points. If the samples actually lie in the vicinity of a curved manifold, then the approach may consider two points as being close-by while they are actually distant if the distance was measured along the manifold itself. Building on MDS, ISOMAP hence introduced an additional step of estimating geodesic distances between the data points through shortest graph paths, in effect allowing to discover the intrinsic geometry of the multidimensional manifold [1], while attempting to preserve the local distances in the resulting low-dimensional projection. More precisely, the algorithm first builds a neighborhood graph, in which every data point x_i is being connected to its k nearest neighbors. The shortest path between two points in this graph, computed using a shortest path algorithm, is then used as a approximation of the geodesic distance. Applying the MDS approach to the obtained pairwise geodesic distance matrix leads to the ISOMAP low-dimensional representation y_i or each data point x_i. In practice, the MDS/ISOMAP solution is obtained though eigendecomposition of the pairwise distance matrix.

ISOMAP presents some weaknesses, related to the possible erroneous connections created in the neighborhood graph, or else when parts of the manifold are not covered by the data (holes in the manifold). In our experiments, we will however observe that ISOMAP already outperforms PCA (and hence MDS, being equivalent to PCA when Euclidean distance is being used) in the preservation of local neighborhoods.

2.2. SNE and t-SNE

The popularity of approaches derived from MDS have inspired variants, in particular through methods attempting to preserve local properties of the data sets in a "softer" fashion. In particular, SNE (Stochastic Neighborhood Embedding) has been designed to preserve neighborhood identity [2]. It does so using a cost function that favors the probability distributions of points belonging to the neighborhoods of other points to be similar in the high-dimensional space and in its low-dimensional embedding. In the original formulation, a Kullback-Leibler (KL) divergence has been used to measure that similarity. In more details, we first estimate the probability the sample x_i in the high-dimensional space would pick sample x_j as its neighbor using the following expression:

$$p_{j|i} = \frac{\exp(\frac{-\|x_i - x_j\|^2}{2\sigma_i^2})}{\sum_{k \neq i} \exp(\frac{-\|x_i - x_k\|^2}{2\sigma_i^2})} \quad (1)$$

where σ_i is the standard deviation of a Gaussian centered on x_i. Similarly, we model the probability that y_i, the low dimensional counterpart of x_i, would take y_j as its neighbor using the following expression:

$$q_{j|i} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq i} \exp(-\|y_i - y_k\|^2)} \quad (2)$$

where the model of proximity is Gaussian, similarly to the model for $p_{j|i}$, but with a standard deviation set to $1/\sqrt{2}$. SNE then proposes to find a representation for which the probabilities $q_{j|i}$ are faithful to $p_{j|i}$. This is achieved by minimizing the mismatch between $q_{j|i}$ and $p_{j|i}$ measured using a KL-divergence:

$$C = \sum_i D_{KL}(P_i \| Q_i) = \sum_i \sum_j p_{j|i} \ln \left(\frac{p_{j|i}}{q_{j|i}} \right) \quad (3)$$

where P_i represents the probability distribution over all data points given data point x_i, and Q_i the equivalent probability distribution in the low-dimensional mapping. Note that SNE performs a binary search for the value of σ_i that produces a P_i with a fixed perplexity specified by the user (f.i. set to 20 in our experiments), were the perplexity is defined based on the Shannon entropy of P_i measured in bits.

More recently, t-SNE proposed a cost function inspired by SNE, but using a Student-t distribution rather than a Gaussian distribution to compute the similarity between two points in the low-dimensional space [7]. This heavy-tailed distribution in the low-dimensional space significantly alleviate the so-called "crowding" problem observed with SNE where far away data samples, for instance low density areas in between natural clusters, come close together in the low-dimensional

space. Additionally, t-SNE actually uses a symmetrized version of SNE, as opposed to the original formulation where $p_{j|i}$ was not necessarily equal to $p_{i|j}$.

The minimization of the cost function in Equation 3 is performed using a gradient descent method. In this paper, we will show that t-SNE outperforms ISOMAP in terms of preservation of local neighborhoods.

2.3. Supervised dimensionality reduction

In addition, researchers have proposed supervised methods, as well as ways to introduce class label information in the unsupervised methods, which was also of interest in our work. Commonly used techniques include Linear Discriminant Analysis (LDA)) as well as Heteroscedastic Discriminant Analysis (HDA [8]). In [9], the authors propose to modify the distance matrix used in MDS/ISOMAP, i.e. increasing distances of data points belonging to different classes, yielding a supervised version of the method (called S-ISOMAP), evaluated on both toy and real data sets. More principled approaches have then been proposed, in which local distance learning is applied in order to increase class discrimination [10, 11]. In this paper, we investigated S-ISOMAP, and further implemented a supervised version of t-SNE following the same lines as S-ISOMAP. Experimental results did not show the benefit of a supervised mode. It did not degrade the results, but did not bring a significant gain in performance neither. Hence, results presented hereafter only include the unsupervised settings.

3. EXPERIMENTAL RESULTS

Two data sets have been used in our evaluations. We first used a production music library (ZeroG ProPack). This library contains more than ten thousand loops and samples of various instruments and music styles. We manually annotated the files within 7 classes of instruments: Brass, Drums, Vocals, Percussion, Electric Bass, Acoustic Guitar and Electric Guitar. After discarding more complex sounds or effects, we ended up with 4380 samples to be used in our evaluations. We also applied the compared approaches to a library of sound effects and ambiances (BBC Sound Effects Library - Original Series), some sounds being hence closer to sound textures. We annotated 873 files from this library in 5 classes of sound: Animals, Mechanic, Impact, Vocals and Crowd.

3.1. Application to sound textures and music

In recent research, a large body of work in the music information retrieval literature has been devoted to the design of feature extraction algorithms for the purpose of characterizing, analyzing, searching or classifying audio content. In this paper, we consider the application of dimensionality reduction preserving timbral properties of sounds. From previous research [12], we ended-up using two groups of features, covering the spectral envelope and the noisiness of the sounds (together with their dynamic and statistical properties), both being important for characterizing the perceived timbre. The state-of-the-art feature set that we used contains:

- Mel-Frequency Cepstral Coefficients (MFCC) as used in [13], computed using 30 ms frames every 10 ms, using a filterbank of 20 filters covering the audible frequency range, and keeping the first 12 coefficients. To be able to capture the temporal characteristics and statistics of the MFCCs, we actually used as features the MFCCs means along the sound file duration, as well as their standard deviation, skewness and kurtosis; the means of the first order temporal derivatives of the MFCCs, as well as their standard deviation, and the means of the second order temporal derivatives of the MFCCs, as well as their standard deviation.
- Spectral Flatness (SF), which is a correlate of the noisiness (opposite to sinusoidality) of the spectrum computed on the same audio frames as MFCCs. It is computed as the ratio between the geometric and arithmetic means of the spectrum energy values. As proposed in [13], the spectrum was divided into 4 sub-bands for computing the flatness: 250-500Hz, 500-1000Hz, 1000-2000Hz and 2000-4000Hz. Here too, we used the mean of the SF over the sound extract duration, as well as its standard deviation, skewness and kurtosis.

3.2. Evaluation metrics

Three metrics have been used in this empirical study:

- The classification error rate. More precisely, we perform k-NN classification in the low-dimensional embedding, as often applied in such evaluation.
- Trustworthiness with respect to the original representation, measured according to the proposal in [14]. In trustworthy representations, visualized proximities hold in the original data as well. Hence, data points originally farther away entering the neighborhood of a sample in the low-dimensional projection decrease the trustworthiness. They indeed create neighborhood relationships that are not present in the data. The trustworthiness measure is defined as:

$$T(k) = 1 - \frac{2}{Nk(2N - 3k - 1)} \sum_{i=1}^{N} \sum_{j \in U_k(i)} (r(i,j) - k)$$

where N is the number of data samples, $r(i,j)$ is the rank of the data sample j in the ordering according to the distance from i in the original data space, and $U_k(i)$ the set of those data samples that are in the neighborhood of size k of the sample i in the visualization display but not in the original data space.

Data sets definitions and labels available as supplementary material at http://www.numediart.org/tools/mediacycle/

http://www.zero-g.co.uk/

- Continuity with respect to the original representation, again, according to the proposal in [14]. Here, samples that are originally in the neighborhood and that are pushed away in the low-dimensional projection decrease the continuity. It is computed similarly to trustworthiness, but with the roles of the original data space and visualization display reversed.

Classification performance using nearest neighbors in 2D is a good indicator of the interestingness of the visualization as it informs on the grouping of samples belonging to natural classes into significant clusters. Note that this metric is strongly influenced by the informativeness of the feature set for the selected classification task. We will show that non-linear techniques, and t-SNE in particular, are able to preserve very well the class structure down to 2D with proper features for the considered classification problem, while PCA performs poorly when applied on the same features.

Trustworthiness and continuity create an evaluation setting similar to precision-recall used in information retrieval. Actually, continuity can intuitively be related to recall, while trustworthiness is to be related to precision as "false positives" in the proximity of samples in the projection decrease the value of this metric. Note that these measures are important to evaluate the preservation of the original neighborhoods in the low-dimensional projection. However, they do not in general inform on the preservation of the class structure, which also depends on the proper choice of features. However, besides the separation of classes, an important and desirable property of dimensionality reduction approaches is the preservation of the natural structure of samples within classes, which can be measured using trustworthiness and continuity. We will show that the non-linear techniques used here provide higher local trustworthiness and continuity than PCA.

Ideally, we could assess the generalization capabilities regarding the above metrics, using held out data not seen by the dimensionality reduction approach. However, starting from SNE, the original formulations of the proposed methods do not provide parametric transforms that allow extending the reduction to data points not seen in the "training" set. Although there are very recent opportunities for such out-of-sample extensions, in particular based on artificial neural networks [15], they have not been used here. This is left to future work.

We applied these three metrics to compare three algorithms of dimension reduction: PCA, ISOMAP and t-SNE. This paragraph presents additional details on the meta-parameters used within these methods. For ISOMAP, the neighborhood graph was constructed by employing the 10 nearest neighbors for each sample. For t-SNE, the perplexity of the conditional probability distribution was set to 20; and we performed 2000 iterations of gradient descent. In our experiments, we used the refinements proposed in [7], including a momentum term in the gradient descent as well as tricks referred to as "early compression" and "early exageration" in [7]. In order to evaluate the performance, we used the feature set described earlier and summarized here: delta MFCC (mean and standard deviation), delta delta MFCC (mean and standard deviation), MFCC (mean, standard deviation, skewness and kurtosis), SF (mean, standard deviation, skewness and kurtosis). We normalized each feature to zero-mean and unity-variance.

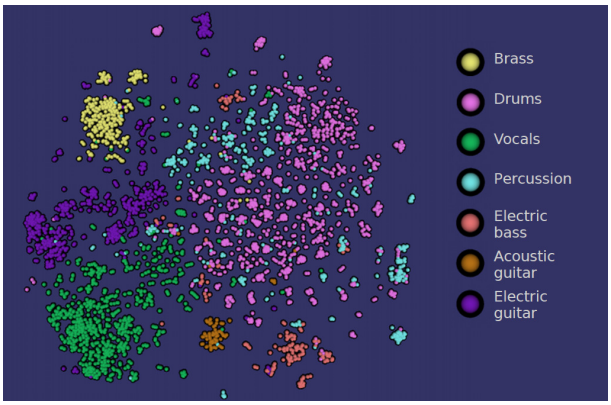
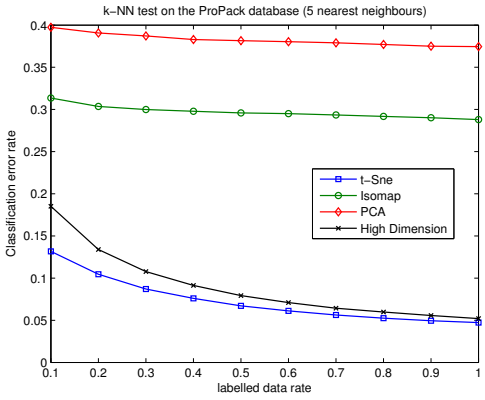


Fig. 1. Example of 2D visualization obtained using t-SNE on the musical loops database

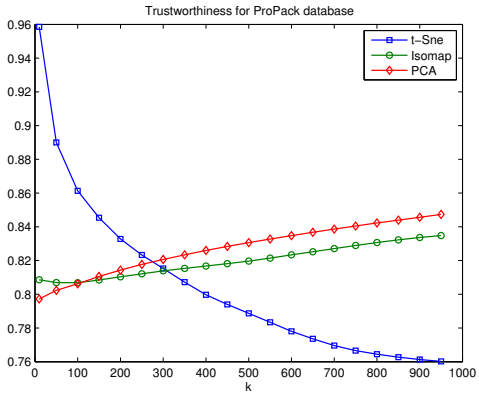
3.3. Results and discussion

An example of visualization obtained using t-SNE on the musical loops database is presented in Figure 1. We can observe that instrument categories are well separated, and also that sub-groups emerge for some instruments. For instance, the different subgroups within the Drums samples correspond to timbrally or stylistically different rhythms. Results according to the evaluation measures are then presented in Figures 2 and 3 when the target dimensionality is set to two, and hence is appropriate for visualization. On the classification results using k-NN (with k=5), we observe that the linear approach PCA is outperformed by the non-linear ISOMAP, which is itself outperformed by t-SNE. Interestingly, classification performance on the reduced dimensional space obtained through t-SNE is generally better than performance obtained when the classification is set in the original feature space, which is surprising given that we reduced down to two dimensions. Note that results are presented for different percentages of randomly selected labeled data (from 10% to 100%, f.i. 10% means that only 438 samples have been labeled using their instrument class in the production music database), hence also simulating a semi-supervised setting were only part of the whole corpora can be annotated with the desired class labels. The results suggests that on this kind of data, an efficient dimensionality reduction scheme (in particular t-SNE) is a useful pre-processing step for semi-supervised classification.

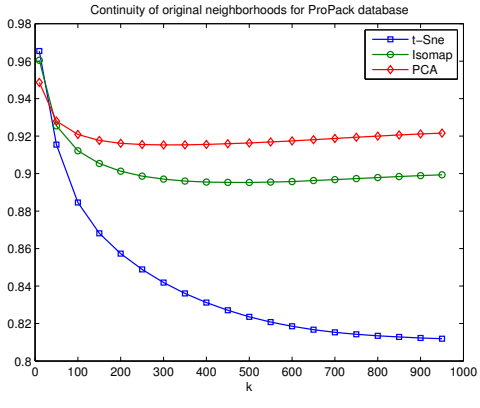
Trustworthiness and continuity are presented for different values of the size k of the neighborhood used in the computation of those measures. Results show the superiority of t-SNE over ISOMAP and PCA, but when k is rather small only. For



(a) k-NN classification error rate

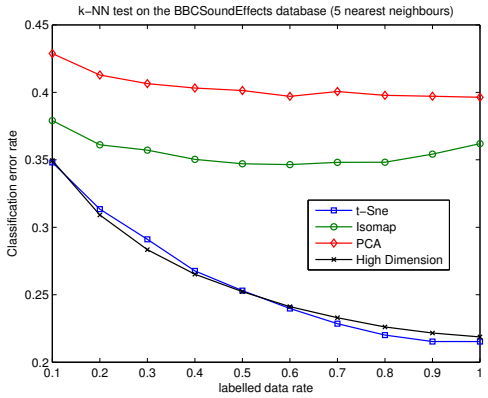


(b) Trustworthiness measure

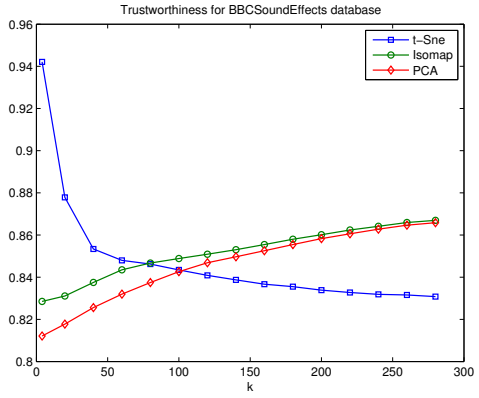


(c) Continuity measure

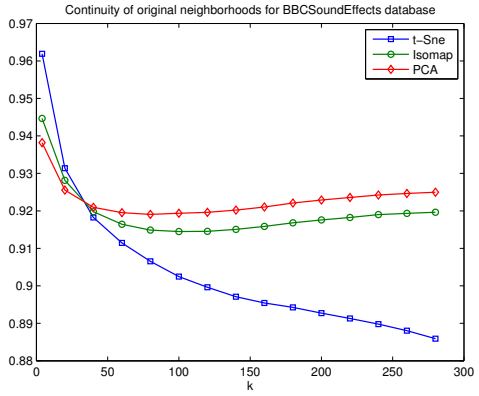
Fig. 2. Experimental results on production music database. 4380 musical loops, seven instruments categories: Brass, Drums, Vocals, Percussion, Electric Bass, Acoustic Guitar and Electric Guitar. For k-NN evaluation, results are presented for different proportions of labeled data. For the trustworthiness/continuity evaluations, the whole data set is used, but the size of the neighborhood is gradually increased from 10 to 1000, in order to assess the quality of the dimensionality reduction both locally and also more globally.



(a) k-NN classification error rate



(b) Trustworthiness measure



(c) Continuity measure

Fig. 3. Experimental results on sound effects database containing recordings from five ambiance sound effects categories: Animals, Mechanic, Impact, Vocals and Crowd.

large values of k, the situation is reversed, with PCA quickly reaching the best performance. This is in line with the specificities of these methods. Indeed, PCA does not distort the space and is not affected that much by more local correlations in the original feature dimensions. On the other side, methods such as t-SNE are designed to better preserve local structure, which is an important factor for preserving classification performance. We made use of the visualization methods within a real-time sound "browser" tool allowing to listen to audio samples by clicking or hovering the mouse on the created 2D maps. In informal discussions around this tool, end-users acknowledged that t-SNE-based visualizations were more useful and interesting that PCA-based ones. Despite subgroups of some sound categories (like specific timbres of brass instruments, of some rhythmic guitar playing phrases) got sometimes farther away from other instances of their category, the visualization created less surprising and annoying phenomena, such as sounds from different categories appearing intermingled in the visualization.

We also observe that continuity is quite good with PCA. With this method, neighbors in the original space are projected on the same area in the low-dimensional one. However, far away samples may also be projected onto the same area. Continuity is hence preserved, being conceptually similar to recall, while trustworthiness is not preserved as well, being conceptually similar to a measure of precision. ISOMAP and t-SNE are able to non-linearly warp the space to project more distant points to different areas of the low-dimensional space, hence better preserving trustworthiness.

Larger dimensional t-SNE embeddings were also evaluated in terms of k-NN classification. Keeping 2, 3 or 10 dimensions and using a condition in which 10% of the data has a class label, we obtained respectively 13.2%, 12.2% and 11.6% classification error on the production music database (k-NN classification error in the original space was 18.5%). Hence, keeping a larger number of dimensions seems beneficial for classification. However, t-SNE only performs marginally worse when the target space contains only 3 or even 2 dimensions. It is hence also very appropriate for visualization of the organization of audio datasets in cases were the preservation of local neighborhoods is wished, while low-dimensional PCA and ISOMAP present visualizations with a strongly degraded value.

4. CONCLUSIONS

In this paper, we presented an experimental study of several dimensionality reduction approaches for creating two-dimensional maps of sound collections, including music and sound effects and ambiances. The evaluations were initially conducted using unsupervised methods. We showed that while simple linear PCA is to be preferred when the global structure is to be visualized, the non-linear technique t-SNE considerably outperforms ISOMAP and PCA when the goal is to preserve local neighborhoods. Interestingly, we also observed that classification performance is boosted when applying this dimensionality reduction scheme as a pre-processing step, even when going down to as low as two dimensions. Of course, these conclusions are valid for moderate size corpora, as the sub-corpora used here and of interest to creative in the music and sound editing fields. Initial experiments using supervised versions of ISOMAP and t-SNE were also conducted, but results were not conclusive. This can be the subject of future work, together with the application of non-linear techniques to larger scale data sets.

5. REFERENCES

[1] J. B. Tenenbaum, V. Silva, and J. C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[2] Geoffrey Hinton and Sam Roweis, "Stochastic neighbor embedding," *Advances in neural information processing systems*, vol. 15, pp. 833–840, 2003.

[3] John A. Lee and Michel Verleysen, *Nonlinear dimensionality reduction*, Springer, New York; London, 2007.

[4] L.J.P. van der Maaten, E. O. Postma, and H. J. van den Herik, "Dimensionality reduction: A comparative review," Tech. Rep. TiCC-TR 2009-005, Tilburg University, 2009.

[5] T. Grill and A. Flexer, "Visualization of perceptual qualities in textural sounds," in *International Computer Music Conference (ICMC 2012)*, Ljubljana, Slovenia, Sept. 2012.

[6] Charles Lo, "Nonlinear dimensionality reduction for music feature extraction," Tech. Rep. CSC2515, Toronto University, 2012.

[7] Laurens van der Maaten and G.E. Hinton, "Visualizing high-dimensional data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[8] Nagendra Kumar and Andreas G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, vol. 26, no. 4, pp. 283 – 297, 1998.

[9] Xin Geng, De-Chuan Zhan, and Zhi-Hua Zhou, "Supervised nonlinear dimensionality reduction for visualization and classification," *Transactions on Systems, Man and Cybernetics, Part B*, vol. 35, no. 6, pp. 1098–1107, Dec. 2005.

[10] Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski, "Information retrieval perspective to nonlinear dimensionality reduction for data visualization," *Journal of Machine Learning Research*, vol. 11, pp. 451–490, Mar. 2010.

[11] Kerstin Bunte, Barbara Hammer, Axel Wismüller, and Michael Biehl, "Adaptive local dissimilarity measures for discriminative dimension reduction of labeled data," *Neurocomputing*, vol. 73, no. 7-9, pp. 1074–1092, Mar. 2010.

[12] Stéphane Dupont, Christian Frisson, Xavier Siebert, and Damien Tardieu, "Browsing sound and music libraries by similarity," in *128th Audio Engineering Society (AES) Convention*, London, UK, May 22-25 2010.

[13] Geoffroy Peeters, "A large set of audio features for sound description (similarity and classification)," in the CUIDADO project. Paris, IRCAM, 2004.

[14] Jarkko Venna and Samuel Kaski, "Local multidimensional scaling," *Neural Networks*, vol. 19, no. 6, pp. 889–899, July 2006.

[15] Martin Renqiang Min, Laurens van der Maaten, Zineng Yuan, Anthony J. Bonner, and Zhaolei Zhang, "Deep supervised t-distributed embedding," in *International Conference on Machine Learning (ICML 2010)*, Haifa, Israel, June 2010, pp. 791–798.