MPEG-7 AUDIO SPECTRUM BASIS AS A SIGNATURE OF VIOLIN SOUND

Aleksander Kaminiarz, Ewa Łukasik

Institute of Computing Science, Poznań University of Technology.
Piotrowo 2, 60-965 Poznań, Poland
e-mail: Ewa.Lukasik@cs.put.poznan.pl

ABSTRACT

The goal of the paper is to examine how robust MPEG-7 Audio Spectrum Basis features are as signatures of instruments from the same group. Instruments analyzed are contemporary concert violins competing in the international violinmaker competition. They have been recorded for research purposes, thus the set of sounds for each instrument and recording conditions are the same – 30 s long musical excerpts and a set of individual sounds. Audio Spectrum Basis captures the statistically most regular features of the sound feature space thus it has been expected to well characterize instruments. The results confirmed the expectations. Since violinmakers follow the same ideal model of instrument construction and use similar material for their creation, differences of their sound are tiny, Audio Spectrum Basis enabled discrimination of several instruments as more dissimilar then the others. However these outliers have been placed by jury musicians during competition on both boundaries of the ranking.

1. INTRODUCTION

The goal of this paper is to examine how good is the MPEG-7 Audio Spectrum Descriptor in distinguishing timbral differences between the contemporary concert violin tones and to examine its predictive power for expert rankings of violin quality. The research related to the machine discrimination of the sound of violins has been inspired by violinmakers competitions, where human experts rank the instruments according to the quality of their sound. During the 10th Henryk Wieniawski International Violinmakers Competition in Poznań in 2001 the sound of competing instruments was recorded and stored in AMATI database [9] along with jury ratings (features rated were e.g. the timbre, the loudness and the playability). The set of sounds comprised individual sounds played in open strings (bowed and plucked), chromatic and dyadic scales and a 30 s. excerpt from J.S. Bach Partita No. 2 in D minor for Solo Violin (BWV 1004) – Sarabande. The collection has already been a benchmark for some research projects. It is planned to transfer it to the digital library [13], along with the MPEG-7 metadata based retrieval mechanism, to be available for a wider audience.

Dealing with the instruments of the same type gives new constraints to the problem of musical instruments recognition. The differences in their timbre may be minute, hardly heard even by very experienced listeners.

The history of the study of acoustic properties of the violin is very long [5] and there is still a significant interest in this area. Therefore the application of new parameterization and machine learning methods in this domain seems to have a considerable research value. Additionally it writes into the domain of Music Information Retrieval, where instruments recognition projects use MPEG-7 descriptors, e.g. [2][3][7][8][11].

In this paper we concentrate only on the MPEG-7 Audio Spectrum Basis (ASB) descriptor - a container for basis functions that are used to project a signal spectrum onto a lower dimensional sub-space suitable for probability model classifiers. Before proceeding with the classification, a closer insight into spectral basis descriptors as a signature for the instruments seemed to be useful for understanding information they provide.

MPEG-7 Audio Framework has already been used as a source of features for examining the violin sound in the previous paper of one of the authors [10]. The research described there concerned individual sounds and features used were from the group of the Harmonic Instrument Timbre Descriptors. The experiments described there showed that the most distinctive descriptor for violin timbre is harmonic spectral centroid. It quite sufficiently divided the set of instruments analyzed into the group of the best, according to the jury assessment, and others. Also the worst instrument in the competition has been always distinguished. The experiments confirmed earlier observations, that violin sounds in the collection of competing contemporary instruments are rather similar even if manufactured in various countries of the world (e.g. Poland, Italy, Russia, South Korea or China).

The MPEG-7 Audio Spectrum Basis calculation is a step towards projection (Audio Spectrum Projection, ASP) of a signal spectrum onto the basis reduced in dimension. The advantage of using this representation lies in its application to the actual pieces of music instead of individual sounds. The description using Audio Spectrum Projection is usually compared with the representation using mel-cepstral coefficients (MFCC) [7]. Our experiments reported in [1] have shown high recognition rate of violins described using MFCC and modelled using GMM.

Audio Spectrum Basis feature extraction mainly consists of a Normalized Audio Spectrum Envelope (NASE) calculation step followed by a decomposition algorithm – such as the Singular Value Decomposition (SVD) or the Primary Component Analysis (PCA) optionally combined with the

Independent Component Analysis (ICA). From the set of basis vectors calculated, only the most significant are kept for further projection. This reduced set of basis vectors will be examined in the paper.

The paper is structured as follows. Section 2 discusses extraction of Audio Spectrum Basis and Audio Spectrum Projection for violin sound. Section 3 illustrates the results of experiments and Section 4 concludes the paper.

2. AUDIO SPECTRUM BASIS EXTRACTION

Spectrum based features are the most frequently used representations of audio signals for classification. Since the dimensionality of the spectrum feature space is large and it does not conform to the psychoacoustic scale of human sound perception, a variety of methods have been applied to diminish the number of spectrum based features. MPEG-7 Audio Framework provides a group of such tools, namely Audio Spectrum Envelope (ASE), Audio Spectrum Basis (ASB) and Audio Spectrum Projection (ASP). The block diagram of the extraction procedure based on the standard [6] is presented in Figure 1.

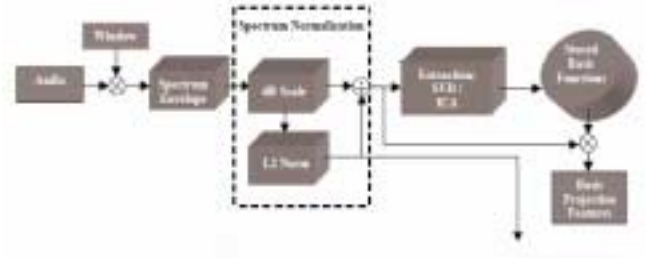

Figure 1 – Extraction method for the AudioSpectrumBasisType and the AudioSpectrumProjectionType [6]

The recordings are sampled with the frequency 44100 Hz. The waveform is divided into blocks of the length close to 30 ms using Hamming window (the standard advises such a window length for psychoacoustic reasons) and with the overlap of 10ms. First Audio Spectrum Envelope (ASE) is calculated using 2048-point FFT giving 1024 equally spaced spectral lines. The power spectral coefficients are grouped in logarithmic sub-bands according to the standard. The use of logarithmic frequency scale is supposed to approximate the response of the human ear. Two frequency edges loEdge and hiEdge limit the frequency range. The spectral resolution r of the frequency bands can be chosen according to the formula:

$$r=2^{j} \text{ octaves } (-4\leq j\leq+3) \qquad (1)$$

By default the loEdge is 62,5Hz and hiEdge is the upper limit of hearing, i.e. 16 kHz. The bands are spanned within 8 octaves, logarithmically centred at the frequency of 1kHz. The resolution may be chosen arbitrarily. In our case j=-2, $r=2^{-2}$ of an octave, so the number of bands is $B_{in}=8/r=2^3/2^2=2^5=32$. Frequency lines below and above loEdge and hiEdge have to be summed up into individual coefficients. Therefore two additional bands are added: from 0Hz to loEdge and from hiEdge to the Nyquist frequency – in our case from 16 kHz to 22.05 kHz (i.e. 6.05 kHz). We get b=34 logarithmic bands. Taking into account the actual spectral resolution, 7 octave span would be sufficient in our calculations, however the default values from the standard have been applied.

It is assumed, that the power spectrum coefficients within the band contribute to both neighbouring bands with a certain weight. The solution presented in Figure 2 [6][7].

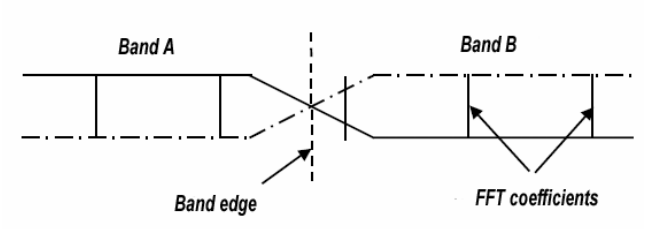

Figure 2 –Weighting the contribution of FFT power coefficients sharing two bands for linear - log conversion [6][7]

Next ASE is converted into the decibel scale. Each decibel-scale spectral vector is normalized with the RMS (root mean square) energy envelope, thus yielding a normalized (L2 norm) log-power version of the ASE called NASE. Figure 3 represents ASE and NASE plot of the sound A played on the open string of violin.

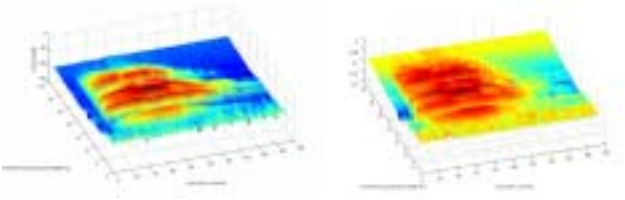

Figure 3 – ASE and NASE plot of the sound A played on the open string of violin

For further reduction of the number of meaningful bands the Audio Spectrum Basis is calculated on which an audio spectrum is usually further projected. Basis functions may be extracted, according to the standard, from SVD or PCA optionally followed by ICA algorithms. The SVD (Singular Value Decomposition) , which is used in this paper, is defined for audio spectrum envelope $\mathbf{X}$ as follows:

$$\mathbf{X}=\mathbf{U}\mathbf{S}\mathbf{V}^{\mathbf{T}} \qquad (2)$$

The basis functions are stored in the columns of a matrix $\mathbf{V}^{\mathbf{T}}$ in which the number of rows corresponds to the length of the spectrum vector and the number of columns corresponds to the number of basis functions. Since the values on the diagonal of matrix $\mathbf{S}$ are diminishing very quickly, the number of columns may be reduced (according to the standard 3 and 10 columns are kept) and this is the source of the reduction of features number. In the Figure 4 the SVD decomposition of the matrix $\mathbf{X}$ representing the spectrum envelope of the sound played on the open A-string is presented:

3. EXPERIMENTS

24 instruments from AMATI collection [9] have been taken for the experiments. The set of instruments contained both

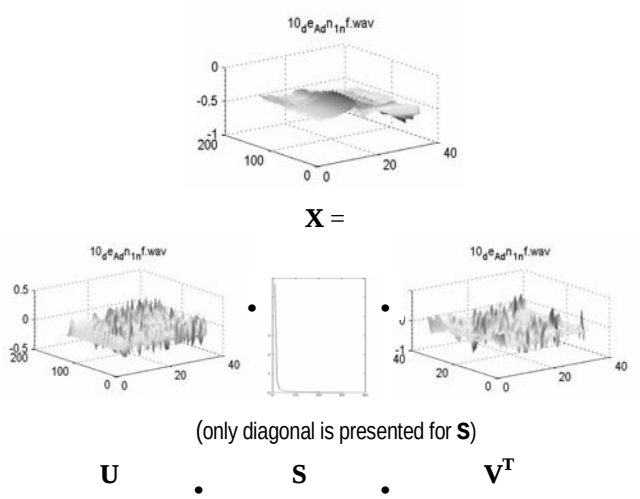

Figure 4 – "Mechanism" of SVD decomposition of the audio spectrum (open A-string violin sound) – for **S** matrix only values on diagonal are presented (exact values are in the text).

groups of instruments: those ranked high and ranked low in the competition. The program in MATLAB has been written for that purpose. Although the full collection of sounds from AMATI has been analysed, including single sounds played on each of four open strings, and diatonic scale, the most informative seemed to be results of the analysis of a 30 s. excerpt from J.S. Bach Partita No. 2 in D minor for Solo Violin (BWV 1004) – Sarabande. For each excerpt (each instrument) an individual ASB has been computed as a signature of a violin.

To get the insight into the values of the audio spectrum bases for the instruments and their ability to characterize them, several visualization methods have been applied, including distance maps and Multidimensional Scaling (MDS) described in the next Sections.

3.1 Spectral Basis vectors visualization

MPEG-7 standard recommends using from 3 to 10 basis vectors for a signal representation with the assumption, that signals projected on this reduced number of basis vectors still contain most of the signal energy and distinctive characteristics of signals are kept. The proportion of the information retained for k basis functions I(k) in the case of music excerpts played on a violin (an exemplary instrument) are following: I(1)=0,70021, I(3)=0,73221, I(7)=0,82079, I(10)=0,86196. The first three basis vectors calculated for the excerpt of J.S. Bach Partita have been presented in Fig. 5. It may be observed, that the difference between k=1 vectors are relatively small - more diversified are next two vectors - however retaining proportionally smaller values. It confirms the fact, that the differences between instruments are tiny and concern small details. It is not surprising – the instruments are of the same type, all contemporary, and following the standard model in violinmaking.

Definitely each basis vector is responsible for a particular feature of the violin timbre. To find out which of them are the most distinctive, a thorough procedure of vectors weighting should be performed for all vectors, until e.g. the distances are more similar to the jurors result. This subject will be developed in future research.

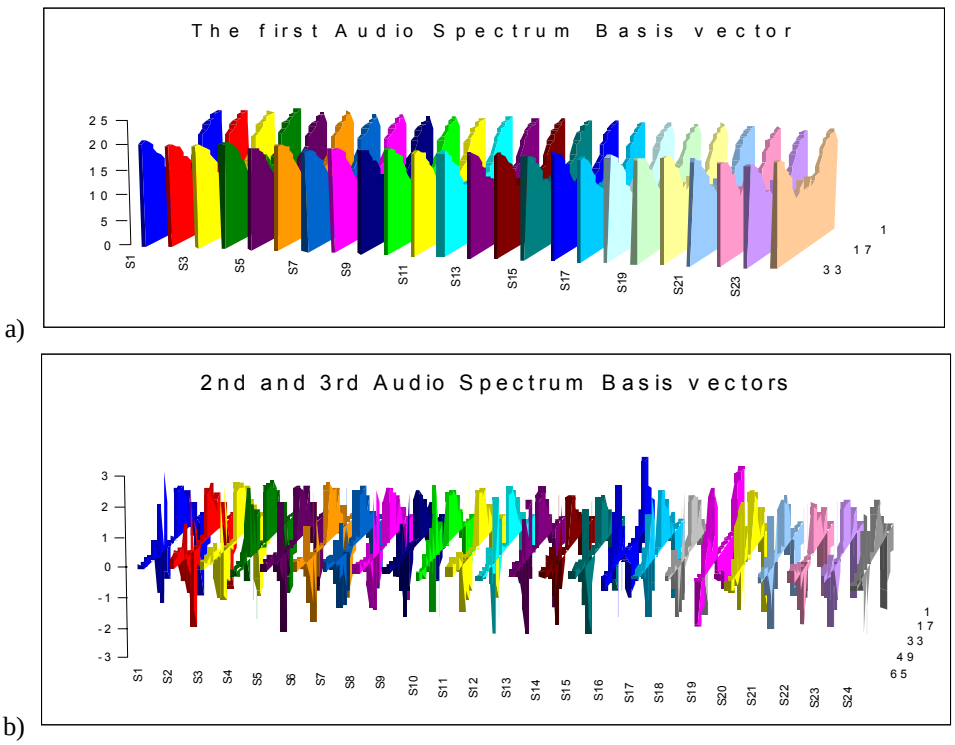

Figure 5 – Values of first three basis vectors for a set of violins:
a) values of the first basis vector, b) 2nd and 3rd vectors concatenated

3.2 Distance measure

To calculate the dissimilarity of basis functions, a Manhattan distance has been used:

$$d(x,y)=\sum_{l=1}^{N}\left|x_l-y_l\right| \qquad (3)$$

where x_l, y_l are elements of basis vectors of two different instruments, N – number of vector elements (N=34).

The choice of the distance measure has been rather arbitral to show only tendency of the results. Manhattan distance is simple to calculate and reliable. Since the procedure of comparing the violin sound has some common features with the comparison of faces, we have taken into the consideration the results from [12], where Manhattan distance was the second (after Mahalanobis and before Euclidean and Angle) to give the most distinct results.

To visualize the distances between all instruments the dissimilarity map has been drawn. It is presented in Fig. 6 for the excerpt from Partita of J.S. Bach. We can read from the map that seven instruments are different from the others: good ones, no 30 (3rd in the ranking), 118 (4th), 46 (10th), 93 (12th), and weaker ones, no 108 (38th), 11 (47th), and 49 (51st). It is interesting to note, that the similar map created for other sounds, e.g. diatonic scale or individual sound played on open string showed more diversification. One possible explanation of this fact is that the violinist, while playing the actual piece of music, controls more the instrument, than while playing the scale or individual sounds (it was clearly visible from other examples, for which the map has been drawn). Another reason may be related to the length and diversity of notes played in actual music passage. Therefore the musical excerpt seemed to give more consistent results.

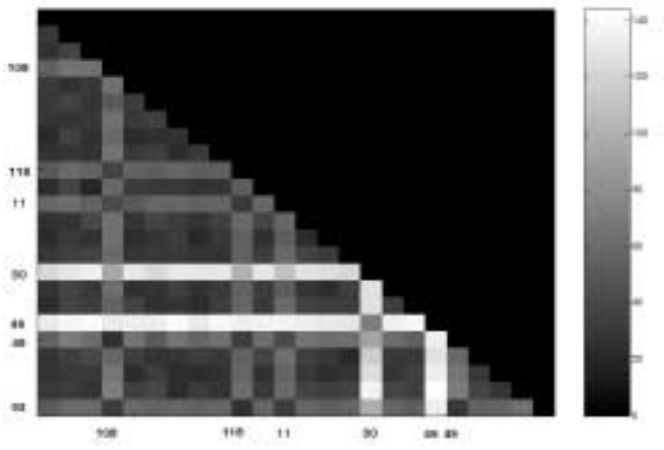

Figure 6 – Dissimilarity map of the first three basis vectors of 24 instruments calculated for the excerpt from J.S. Bach Partita in d-minor (brighter are more distant sounds)

3.3 Multidimensional scaling (MDS)

The relationship between the objects (violin sounds) in the multidimensional space is not easy to present to humans in such a way, that most of the relationship between them are visualized. It is hard to say if the objects form any clusters. The method that assists humans to better perceive the relative distances between sounds described in high-dimensional data sets is multidimensional scaling (MDS) [4]. Since seeing the relative positions of objects in the multidimensional space of attributes is directly impossible the method suggests specific positions (the x–y co-ordinates) of the considered objects in two-dimensional space (the space may be reduced also to three dimensions, but then the visualization is not so straightforward). The two-dimensional positions are chosen in such a way that the distances between the objects in this two-dimensional space match as well as possible the distances between the objects in the original, multi-dimensional space. And although the structure represented by the positions of objects in the new, two-dimensional plane is not the same as the structure of their positions in the original space, such a 'map' of objects may be treated as a good approximation of this multidimensional structure, especially if the Kruskal Stress is small. In our experiments the stress converged quickly to zero, therefore the representation seems to be reliable. Figures 7-10 display mapping of the basis vectors representing violins playing J.S. Bach Partita in 2-D space. First only one basis vector (34 dimensions) is taken into consideration (Fig. 7), then the distances for three basis vectors are presented (Fig. 8), then seven (Fig. 9) and finally the MDS graph for ten vectors is displayed (Fig.10). The numbers on the graphs refer to the instruments - these are competition numbers, that precisely identify violins.

3.4 Analysis of results

The analysis of Figures 7-10 confirms the initial conclusions drawn from the discussion in Sections 3.1 and 3.2. The first basis vector, whose elements have relatively large values, discovers some differences between violins, but no clusters of more similar objects have been detected. The aspect the fist basis vector represents is probably related to the main resonances that characterize individual violins. Perhaps instruments on the border of the cloud may be regarded as the most different from the others. Indeed instruments no 30, 46, 118, 108, 11 and 15 are placed there, but the proximity to their neighbors is comparable with distances within the cloud of objects. After adding the second and the third basis vectors to MDS analysis some objects more distant from the main cloud of instruments appear, meaning that some important features have been discovered.. The Figure 8 is concordant with the Figure 6 showing the most distinct instruments. It is worth noting, that instruments holding the competition numbers 46 and 30 had a high position in the ranking – it may be supposed that the difference of their sound attracted jurors. However other features must have been also important for jurors, as some lower rated instruments have been found in the same group (108, 11, 49, 15).

4. CONCLUSIONS

The goal of the paper was to examine how powerful MPEG-7 Audio Spectrum Basis (ASB) features are as the signatures of instruments from the same group – violins competing during the international violinmaker competition. ASB captures the statistically most regular features of the spectral feature space. From a wide range of experiments we have reported results of those concerning the comparison of sig-

natures of 24 test instruments on which actual piece of music has been played. As expected, the proportion of information retained for the initial basis vectors was substantial: over 73% for three basis vectors, and over 86% for ten basis vectors. The similarity map and multidimensional scaling used for visualization have indicated the violins, that were placed at a distance from the main group of instruments, suggesting their distinctive sound qualities. However not clear explanation may be given why in the group of outliers - the instruments ranked as the best are mixed with instruments ranked low.

Audio Spectrum Basis descriptors have only vaguely indicated the possible factors influencing jurors' decision. However the exact similarity of sounds has not been observed, so possibly the ASB descriptors could play the role of the compact signatures of violin sounds especially if basis vectors were appropriately weighted (this might be performed in the future). In further experiments the ASB will be used to calculate Audio Spectrum Projection needed for instruments recognition. The recognition rate will be compared to the one obtained in [1] for Mel Cepstral coefficients. Conclusions will be verified using a larger set of instruments.

REFERENCES

[1] P. Anioła, E. Łukasik, "Java Library for Automatic Musical Instruments Recognition", *AES Convention Paper 7157*, Vienna, 2007.

[2] M. Casey, "MPEG-7 Sound Recognition Tools", *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6), 2001, pp. 737-747.

[3] M. Casey, "General sound classification and similarity in MPEG-7", *MERL Cambridge Research Laboratory*, 2001.

[4] T.F. Cox, M.A.A. Cox, *Multidimensional Scaling*, Chapman and Hall, London, 1994.

[5] C.M. Hutchins, *A History of Violin Research*, JASA, 73, pp.1421-1432, 1983.

[6] ISO/IEC 15938-4, *Information Technology – Multimedia Content Description Interface* – Part 4: Audio, 2001.

[7] H.G. Kim, N. Moreau, T. Sikora, *MPEG-7 Audio and Beyond. Audio Content Indexing and Retrieval*, John Wiley & Sons Ltd. 2005.

[8] B. Kostek, *Perception-Based Data Processing in Acoustics*, Springer-Verlag, Berlin Heildelberg, 2005.

[9] E. Łukasik, "AMATI-Multimedia Database of Musical Sounds", *Proc. Stockholm Music Acoustics Conference*, KTH Stockholm 2003, pp. 79-82.

[10] E. Łukasik, "MPEG-7 Musical Instrument Timbre Descriptors Performance in Discriminating Violin Voices", *Proc. IEEE Workshop "Signal Processing 2004"*, pp. 87-90.

[11] P. Szczuko, P. Dalka, M. Dąbrowski, B. Kostek, "MPEG-7 based low level descriptor effectiveness in the automatic musical sound classification", AES Convention Paper 6105, Berlin 2004.

[12] W.S. Yambor, B.A. Draper, J.R. Beveridge, "Analyzing PCA-based Face Recognition Algorithms: Eigen-vector Selection and Distance Measures", [in] *Empirical Evaluation Methods in Computer Vision*, H. Christensen and J. Phillips (eds.), World Scientific Press, Singapore, 2002.,pp.39-60.

[13] URL: http://www.wbc.poznan.pl/dlibra

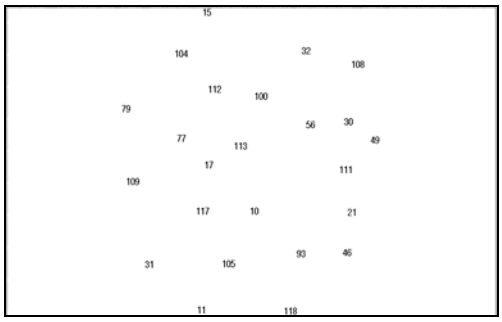

Figure 7 – Distances of violin sound (J.S. Bach) represented by the Multidimensional Scaling for the first basis vector

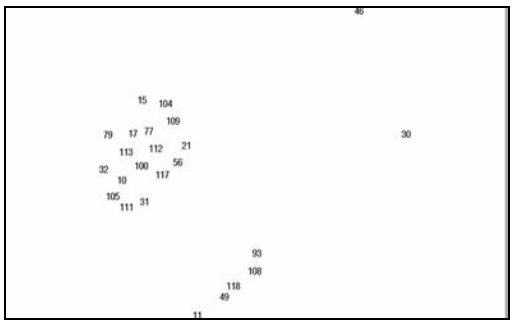

Figure 8 – Distances of violin sound (J.S. Bach) represented by the Multidimensional Scaling for three first basis vectors

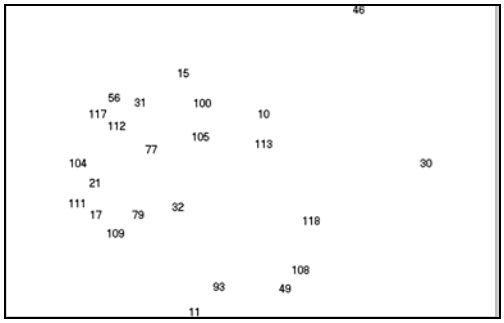

Figure 9 – Distances of violin sound (J.S. Bach) represented by the Multidimensional Scaling for seven first basis vectors

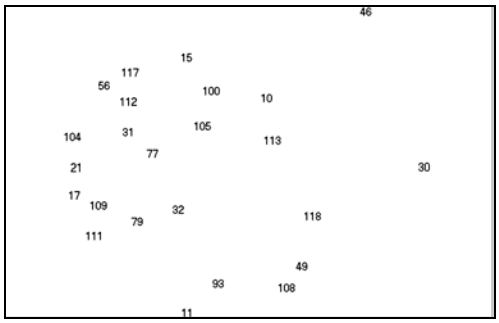

Figure 10 Distances of violin sound (J.S. Bach) represented by the Multidimensional Scaling for ten first basis vectors