Towards instrument segmentation for music content description: a critical review of instrument classification techniques

Perfecto Herrera, Xavier Amatriain, Eloi Batlle, Xavier Serra
Audiovisual Institute - Pompeu Fabra University
Rambla 31, 08002 Barcelona, Spain
{perfecto.herrera, xavier.amatriain, eloi.batlle, xavier.serra}@iua.upf.es

A system capable of describing the musical content of any kind of sound file or sound stream, as it is supposed to be done in MPEG7-compliant applications, should provide an account of the different moments where a certain instrument can be listened to. In this paper we concentrate on reviewing the different techniques that have been so far proposed for automatic classification of musical instruments. As most of the techniques to be discussed are usable only in "solo" performances we will evaluate their applicability to the more complex case of describing sound mixes. We conclude this survey discussing the necessity of developing new strategies for classifying sound mixes without a priori separation of sound sources.

Keywords: classification, timbre models, segmentation, music content processing, multimedia content description, MPEG-7

Introduction

The need for automatically classifying sounds[1] arises in contexts as different as bioacoustics or military surveillance. Our focus, anyway, will be that of multimedia content description, where segmentation of musical audio streams can be done in terms of the instruments that can be listened to (for example in order to locate a “solo” in the middle of a song). Two main different objectives can be envisioned in this context:

- segmentation according to the played instrument, where culturally accepted labels for all the classes have to be associated with certain feature vectors (hence it is a clear example of supervised learning problem);
- segmentation according to perceptual features, where there are no universal labels for classifying segments but similarity distance functions derived from psychoacoustical studies on what humans intend as “timbral similarity” [1;2;3;4].

The first point will be the subject of this paper, whereas the second one has been partially pursued in one of our recent contributions to the MPEG-7 process [5].

Although a blind or completely bottom-up approach could be feasible for tackling the problem, we can assume that some additional meta-information (e.g. title of the piece, composer, players…) will be available in the moment of performing the classification, because these and other metadata are expected to be part of the MPEG7 standard that should be approved by the end of 2001 [6]. Descriptions compliant with that standard will include, alongside all those textual metadata, other structural, semantic and temporal data about the instruments or sound sources that are being played in a specific moment, the notes/chords/scales they are playing, or the types of expressive musical resources (e.g. vibrato, sforzando…) used by the players. Extracting all those non-textual data by hand is an overwhelming task

[1] The construction of a classification procedure from a set of data for which the true classes are known has also been variously termed *pattern recognitio*n, *discriminatio*n, or *supervised learning* (in order to distinguish it from *unsupervised* learning or *clustering* in which the classes are inferred from the data) [55]. The aim of supervised learning is to derive, from correctly classified cases, a rule whereby we can classify a new observation into one of the existing classes.

and therefore automatic procedures have to be found to perform what has been called the "signal-to-symbol transformation" [7].

Instrument segmentation of complex mixtures of signals is still far from being solved (but see [8], [9], [10] for different approaches). Therefore, one preliminary way of overriding the obnoxious stage of separating components is reducing the scope of the classification systems to only deal with isolated sounds. There is an obvious tradeoff in endorsing this strategy: we gain simplicity and tractability, but we lose contextual and time-dependent cues that can be exploited as relevant features for classifying the sounds. As this has been the preferred strategy in the current literature on instrument classification, this paper will concentrate on them. A review of those studies would not be complete without discussing the features used for classification, but space constraints have prevented us of including it here.

Classification of monophonic sounds

K-Nearest Neighbors

The *K-Nearest Neighbors* algorithm is one of the most popular algorithms for instance-based learning. It first stores the feature vectors of all the training examples and then, for classifying a new instance, it finds (usually using an Euclidean distance) a set of *k* nearest training examples in the feature space, and assigns the new example to the class that has more examples in the set. Although it is an easy algorithm to implement, the K-NN technique has several drawbacks: as it is a lazy algorithm [11] it does not provide a generalization mechanism (because it is only based on local information), it requires having in memory all the training instances, it is highly sensitive to irrelevant features (as they can dominate the distance metrics), and it may require a significant load of computation each time a new query is done.

A K-NN algorithm classified 4 instruments almost with complete accuracy in [12]. Unfortunately, they used a small database (with restricted note range to one octave, although including different dynamics), and conclusions should be taken with caution, moreover if we consider the following more thoroughful works.

Martin and Kim [13] (but also see [14]) developed a classification system that used the K-NN with 31 features extracted from cochleagrams. The system also used a hierarchical procedure consisting on first discriminating pizzicati from continuous notes, then discriminating between different "families" (sustained sounds furthermore divided into strings, woodwind and brass), and finally, specifically classifying sounds into instrument categories. With a database of 1023 sounds they achieved 87% of successful classifications at the family level and 61% at the instrument level when no hierarchy was used. Using the hierarchical procedure increased the accuracy at the instrument level to 79% but it degraded the performance at the family level (79%). Without including the hierarchical procedure performance figures were lower than the ones they obtained with a Bayesian classifier (see below).

[15] used a combination of Gaussian classifier[2] and k-NN for classifying 1498 samples into specific instrumental families or specific instrument labels. Using an architecture very similar to Martin and Kim's hierarchy (sounds are first classified in broad categories and then the classification is refined inside that category) they reported a success of 75% in individual instrument classification (and 94% in "family" classification). Additionally they report a small accuracy improvement by only using the best features for each instrument and no hierarchy at all (80%).

[2] The Gaussian classifier was only used for rough discrimination between pizzicati and sustained sounds

A possible enhancement of the K-NN technique consisting on weighting each feature according to its relevance for the task has been used by the Fujinaga team[3] [16;17;18] [19]. In a series of three experiments using over 1200 notes from 39 different timbres taken from the McGill Master Samples CD library the success rate of 50%, observed when only the spectral shape of steady-state notes was used, increased to 68% when tristimulus, attack position and features of dynamically changing spectrum envelope, such as the change rate of the centroid, were added. In the most recent paper, a real-time version of this system was reported.

The fact the best accuracy figures are around 80% and that Martin and Fujinaga have settled into similar figures, can be interpreted as an estimation of the limitations of the K-NN algorithm (provided that the feature selection has been optimized with genetic or other kind of techniques). Therefore, more powerful techniques should be explored.

Naive Bayesian Classifiers

This method[4] involves a learning step in which the probabilities for the classes and the conditional probabilities for a given feature and a given class are estimated, based on their frequencies over the training data. The set of these estimates corresponds to the learned hypothesis, which is formed without searching, simply by counting the frequency of various data combinations within the training examples, and can be used then to classify each new instance.

This technique has been used with 18 Mel-Cepstrum Coefficients in [20]. After clustering the feature vectors with a K-means algorithm, a Gaussian mixture model from their means and variances was built. This model was used to estimate the probabilities for a Bayessian classifier. It then classified 30 short sounds of oboe and sax with an accuracy rate of 85%. Martin [14] enhanced a similar Bayesian classifier with context-dependent feature selection procedures, rule-one-out category decisions, beam search, and Fisher discriminant analysis for estimating the Maximum A Priori probabilities. In [13] performance of this system was better than that of a K-NN algorithm at the instrument level (71% accuracy) and equivalent to it at the family level (85% accuracy).

Discriminant Analysis

Classification using categories or labels that have been previously defined can be done with the help of *discriminant analysis*, a technique that is related to multivariate analysis of variance and multiple regression. Discrimination analysis attempts to minimize the ratio of within-class scatter to the between-class scatter and builds a definite decision region between the classes. It provides linear, quadratic or logistic functions of the variables that "best" separate cases into two or more predefined groups, but it is also useful for determining which the most discriminative features are and the most alike/different groups. One possible drawback of the technique is its reduced generalization power, although Jackknife tests (cross-validating with leave-one-case-out) can protect against overfitting to the observed data.

Surprisingly the only study using this technique, and not thoroughly, has been the one by Martin and Kim. They only used LDA for estimation of the mean and variance for the gaussians of each class to be fed to an enhanced naive Bayesian classifier. Perhaps it is commonly assumed that the classification problem is much more complex than that of a quadratic estimation, but it means taking from granted something that has not been experimentally verified, and maybe it should be done.

Following this line, in a pilot study carried in our laboratory with 120 sounds from 8 classes and 3 families we have got 85% (Jackknifed: 75%) accuracy using quadratic linear discriminant functions in

[3] The feature relevance was determined with a genetic algorithm
[4] Here *naive* means that it assumes feature independence

two steps (sounds are first assigned to family, and then they are specifically classified). Given that the features we used were not optimized for segmentation but for searching by similarity, we expect to be able to get still better results when we include other valuable features.

Binary trees

Binary trees, in different formulations, are pervasively used for different machine learning and classification tasks. They are constructed top-down, beginning with the feature that seems to be the most informative one, that is, the one that maximally reduces entropy. Branches are then created from each one of the different values of this descriptor (in the case of non binary valued descriptors a procedure for dichotomic partition of the value range must be defined). The training examples are sorted to the appropriate descendant node, and the entire process is then repeated recursively using the examples of one of the descendant nodes, then with the other. Once the tree has been built, it can be pruned to avoid overfitting and to remove secondary features. Although building a binary tree is a recursive procedure, it is anyway faster than the training of a neural network.

Binary trees are best suited for approximating discrete-valued target functions but they can be adapted to real-valued features as Jensen's binary decision tree [21], which exemplifies their application to instrument classification. In his system the trees are constructed by asking a large number of questions (e.g. "attack time longer than 60 ms?"), then, for each question, data are split into two groups, goodness of split (average entropy) is calculated and finally the question that renders the best goodness is chosen. Once the tree has been built using the learning set, it can be used for classifying new sounds (each leaf corresponds to one specific class) but also for making explicit rules about which features better discriminate an instrument from another. Unfortunately results regarding the classification of new sounds have not yet been published (but see Jensen's thesis [22] for an attempt on log-likelihood classification functions).

An application of the C4.5 algorithm [23] can be found in [24], where a database of 18 classes and 62 features was classified with accuracy rates between 64% and 68% depending on the test procedure.
A final example of a binary tree for audio classification, although not specifically tested with musical sounds, is that of Foote [25]. His tree-based supervised vector quantization with maximization of mutual information uses frame-by-frame 12 Mel-cepstral coefficients plus energy for partitioning the feature space into a number of discrete regions. Each split decision in the tree involves comparing one element of the vector with a fixed threshold, that is chosen to maximize the mutual information between the data and the associated labels that indicate the class of each datum. Once the tree is built, it can be used as a classifier by computing histograms of frequencies of classes in each leaf of the tree and using distance measures between histogram templates derived from the training data and the resulting histogram for the test sound.

Support Vector Machines

SVMs are a very recently developed technique that is based on statistical learning theory [26]. The basic training principle behind SVMs is finding the optimal linear hyperplane such that the expected classification error for unseen test samples is minimized (i.e. they look for good generalization performance). According to the structural risk minimization inductive principle, a function that classifies the training data accurately and which belongs to a set of functions with the lowest complexity will generalize best regardless of the dimensionality of the input space. Based on this principle, a linear SVM uses a systematic approach to find a linear function with the lowest complexity. For linearly non-separable data, SVMs can (nonlinearly) map the input to a high dimensional feature space where a linear hyperplane can be found. Although there is no guarantee that a linear solution will always exist in the high dimensional space, in practice it is quite feasible to construct a working solution. In sum, training a SVM is equivalent to solving a quadratic programming with linear constraints and as many variables as

data points. A SVM was used in [27] for the classification of eight solo instruments playing musical scores from well-known composers. The best accuracy rate was a 70% using 16 MCCs and sound segments that were 0.2 seconds long. When she attempted classification on longer segments an improvement was observed (83%) although there were two instruments very difficult to classify (trombone and harpsichord). Another worth-mentioning feature of this study is the use of truly independent sets for the learning and for the test sets (and they were mainly "solo" phrases from commercial recordings).

Artificial Neural Networks

A very simple feedforward network with a backpropagation training algorithm was used, along with K-NN, in [12]. The network (a 3/5/4 architecture) learnt to classify sounds from 4 very different instruments (piano, marimba, accordion and guitar) with a high accuracy (best figure 97%), although slightly better figures were obtained using the simplest K-NN algorithm (see above).

A comparison between a multilayer perceptron, a time-delay network, and a hybrid self-organizing network/radial basis function can be found in [28]. Although very high success rates were found (97% for the perceptron, 100% for the time-delay network, and 94% for the self-organizing network) it should be noted that the experiments used only 40 sounds from 10 different classes and ranging one octave only.
Examples of self-organizing map [29] usage can be found in [30], [31],[32], [33]. All these studies use some kind of auditory pre-processing for getting the features that are fed to the network, then build the map, and finally compare the clustering of sounds made by the network with human subjects similarity judgments ([1], [34]). From these maps and comparisons the authors advance timbral spaces to be explored, or confirm/disconfirm theoretical models that explain the data. It can be seen then, that the classification we get from these kind of systems are not directly usable for instrument recognition, as they are not provided with any label to be learnt. Nevertheless, a mechanism for associating their output clusters to specific labels seems feasible to be implemented (e.g. the radial basis function used by Cemgil, see above). The ARTMAP architecture [35] eventually implements this strategy by a complex topology: an associative memory is connected with an input network that self-organizes binary input patterns, with an output network that does the same with binary and real-valued patterns, and with an orienting subsystem that may alter the input processing depending on output and associative memory states. Fragoulis et al [36] successfully used an ARTMAP for the classification of 5 instruments with the help of only ten features (slopes of the first five partials, time delays of the first 4 partials respective to the fundamental, and high frequency energy). The errors (2%) were attributed to not having taken into account different playing dynamics in the training phase.

The most thorough study on instrument classification using neural networks is, perhaps, that of Kostek's [37], although it has been a bit neglected in the relevant literature. Her team has carried out several studies [38] [39] on network architecture, training procedures, and number and type of features, although the number of classes to be classified has been always too small. They have used a feedforward NN with one hidden layer, and their classes were trombone, bass trombone, English horn and contrabassoon, instruments with somehow similar sound. Accuracy rates use to be higher than 90%, although they vary depending on the type of training and number of descriptors.

Although some ANN architectures are capable of approximate any function, and therefore neural networks are a good choice when the function to be learned is not known in advance, they have some drawbacks: first of all, the computation time for the learning phase is very long, tweaking of their parameters can also be tedious and prohibitive, and over-fitting (excessive number of bad selected examples) can degrade their generalization capabilities. On the positive side, figures coming from available studies do not quite outperform other simpler algorithms but anyway neural networks may

exhibit one advantage in front of some of them: once the net has learnt, the classification decision is very fast (compared to K-NN or to binary trees).

Higher Order Statistics

When signals have Gaussian density distributions, we can describe them thoroughly with second order measures like the autocorrelation function or the spectrum. There are some authors who claim that musical signals, as they have been generated through non-linear processes, do not fit a Gaussian distribution. In that case, using *higher order statistics* or polyspectra, as for example skewness of bispectrum and kurtosis of trispectrum, it is possible to capture all information that could be lost if using a simpler Gaussian model. With these techniques, and using a Maximum Likelihood classifier, Dubnov and his collaborators [40] have showed that discrimination between 18 instruments from string, woodwind and brass families is possible although they only provide figures for a classification experiment that used generic classes of sounds (not musical notes).

Rough Sets

Rough sets [41] are a novel technique for evaluating the relevance of the features used for description and classification. It has been developed in the realm of knowledge-based discovery systems and data mining (although similar, not to be mistaken with *fuzzy sets*). In rough set theory any set of similar or indiscernible objects is called an elementary set and forms a basic granule of knowledge about the universe; on the other hand, the set of discernible objects are considered rough (imprecise or vague). Vague concepts cannot be characterized in terms of information about their elements; however they may be replaced by two precise concepts, respectively called the *lower approximation* and the *upper approximation* of the vague concept. The lower approximation consists of all objects that surely belong to the concept whereas the upper approximation contains all objects that possibly belong to the concept. The difference between both approximations is called the *boundary region* of the concept. The assignment of an object to a set is made through a membership function that as a probabilistic flavor. Once information is conveniently organized into information tables this technique is used to assess the degree of vagueness of the concepts, the interdependency of attributes and therefore the alternatives for reducing complexity in the table without reducing the information it provides. Information tables regarding cases and features can be interpreted as conditional decision rules of the form IF {feature x} is observed THEN {isanYobject}, and consequently they can be used as classifiers. An elementary but formal introduction to rough sets can be found in [42]. Applications of this technique to different problems, including those of signal processing [43], alongside with discussion of software tools implementing these kinds of formalisms, are presented in [44]. When applied to instrument classification [45] reports accuracy rates higher than 80% for classification of the same 4 instruments mentioned in the ANN's section. The main cost of using rough sets is however the need for quantization of features' values, a non-trivial issue indeed, because in the previous study different results were obtained depending on the quantization method (see also [46] and [47]). On the other hand, when compared to neural networks or fuzzy sets rules, rough sets have several benefits: they are cheaper in terms of computational cost and the results are similar to those obtained with the other two techniques.

Towards classification of sounds in more complex contexts

Although we have found that there are several techniques and features which provide a high percent of success when classifying isolated sounds, it is not clear that they can be applied directly and successfully to the more complex task of segmenting monophonic phrases or complex mixtures. Additionally, many of them would not accomplish the requirements discussed in [14] for real-world sound-source recognition systems. Instead of assuming a preliminary source separation stage that facilitates the direct applicability of those algorithms, we are committed with an approach of signal *understanding without separation* [48].

This means that with relatively simple signal-processing and pattern-classification techniques we elaborate judgments about the musical qualities of a signal (hence, describing content). Provided that desideratum, we can enumerate some apparently useful strategies to complement the previously discussed methods:

- *Content awareness* (i.e. using metadata when available): the MPEG-7 standard provides descriptors that can help to partially delimitate the search space for instrument classification. For example, if we know in advance that the recording is a string quartet, or a heavy-metal song, several hypotheses regarding the sounds to be found can be used for guiding the process.
- *Context awareness*: contextual information can be conveyed not only from metadata, nor from models in a top-down way. It also can spread from local computations at the signal level by using descriptors derived from analysis of groups of frames. Note transition analysis, for example, may provide a suitable context [49].
- *Use of synchronicities and asynchronicities:* co-modulations or temporal coherence of partials may be used for inferring different sources, as some CASA systems do [50;8].
- *Use of spatial cues:* in stereophonic recordings we can find systematic instrument positioning that can be tracked for reducing the candidate classes.
- *Use of partial or incomplete cues*: contrasting with the problems of source separation or analysis for synthesis/transformation, our problem does not demand any complete characterization or separation of signals and, consequently, incomplete cues might be enough exploited.
- *Use of neglected features*: as for example articulations between notes, expressive features (e.g. vibrato, portamento) or what has been called “specificities” of instrument sounds [3].
- *Combining different subsystems*: different procedures can make different estimations and errors. Therefore a wise combination may yield better results than figuring out what is the best or what is good in each one [51] [52]. Combinations can be done at different processing stages: at the feature computation (concatenating features), at the output of the classification procedures (combining hypothesis), or also in a serial layout where the output of one classification procedure is the input to another procedure (as Martin’s *MAP+Fisher projection* exemplifies).
- *Use of more powerful algorithms for representing sequences of states:* Hidden Markov Models [53] are good candidates for representing long sequences of feature vectors that define an instrument sound, as [54] have demonstrated for generic sounds.

Conclusions

We have discussed the most commonly used techniques for instrument classification. Although they provide a decent starting point for the more realistic problem of detection and segmentation of musical instruments in real-world audio, conclusive statements after performance figures can be misleading because of inherent biases in each one of the algorithms. Enhancing or tuning them for the specificities of dealing with realistic musical signals seems a more important task than selecting the best existing algorithm. Consequently other complementary strategies should be addressed in order to achieve the kind of signal understanding we aim at.

References

[1] Grey, J. M., "Multidimensional perceptual scaling of musical timbres," *Journal of the Acoustical Society of America*, 61, pp. 1270-1277, 1977.

[2] Krumhansl, C. L., "Why is musical timbre so hard to understand?," in Nielzenand, S. and Olsson, O. (eds.) *Structure and perception of electroacoustic sound and music* Amsterdam: Elsevier, 1989, pp. 43-53.

[3] McAdams, S., Winsberg, S., de Soete, G., and Krimphoff, J., "Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes," *Psychological Research*, 58, pp. 177-192, 1995.
[4] Lakatos, S., "A common perceptual space for harmonic and percussive timbres," *Perception and Psychophysics*, in press, 2000.
[5] Peeters, G., McAdams, S., and Herrera, P. Instrument Sound Description in the context of MPEG-7. Proc. of the ICMC 2000.
[6] ISO/MPEG-7. Overview of the MPEG-7 Standard. 15-6-2000. Electronic document: http://drogo.cselt.stet.it/mpeg/standards/mpeg-7/mpeg-7.htm
[7] Green, P. D., Brown, G. J., Cooke, M. P., Crawford, M. D., and Simons, A. J. H., "Bridging the Gap between Signals and Symbols in Speech Recognition," in Ainsworth, W. A. (ed.) *Advances in Speech, Hearing and Language Processing* JAI Press, 1990, pp. 149-191.
[8] Ellis, D. P. W., "Prediction-driven computational auditory scene analysis." Ph.D. thesis MIT. Cambridge, MA, 1996.
[9] Bell, A. J. and Sejnowski, T. J., "An information maximisation approach to blind separation and blind deconvolution," *Neural Computation*, 7 (6), pp. 1129-1159, 1995.
[10] Varga, A. P. and Moore, R. K. Hidden Markov Model decomposition of speech and noise. Proc. of the ICASSP. pp. 845-848, 1990.
[11] Mitchell, T. M., *Machine Learning* Boston, MA: McGraw-Hill, 1997.
[12] Kaminskyj, I. and Materka, A. Automatic source identification of monophonic musical instrument sounds. Proc. of the IEEE International Conference On Neural Networks. 1, 189-194, 1995.
[13] Martin, K. D. and Kim, Y. E. Musical instrument identification: A pattern-recognition approach. Proc. of the 136th meeting of the Acoustical Society of America. 1998.
[14] Martin, K. D., "Sound-Source Recognition: A Theory and Computational Model." Ph.D. thesis, MIT. Cambridge, MA, 1999.
[15] Eronen, A. and Klapuri, A. Musical instrument recognition using cepstral coefficients and temporal features. Proc. of the ICASSP. 2000.
[16] Fujinaga, I., Moore, S., and Sullivan, D. S. Implementation of exemplar-based learning model for music cognition. Proc. of the International Conference on Music Perception and Cognition. 171-179, 1998.
[17] Fujinaga, I. Machine recognition of timbre using steady-state tone of acoustical musical instruments. Proc. of the 1998 ICMC. 207-210, 1998.
[18] Fraser, A. and Fujinaga, I. Toward real-time recognition of acoustic musical instruments... Proc. of the ICMC. 175-177, 1999.
[19] Fujinaga, I. and MacMillan, K. Realtime recognition of orchestral instruments. Proc. of the ICMC. 2000.
[20] Brown, J. C., "Musical instrument identification using pattern recognition with cepstral coefficients as features," *Journal of the Acoustical Society of America*, 105 (3), pp. 1933-1941, 1999.
[21] Jensen, K. and Arnspang, J. Binary decission tree classification of musical sounds. Proc. of the 1999 ICMC. 1999.
[22] Jensen, K, "Timbre models of musical sounds." Ph.D. thesis University of Copenhaguen, 1999.
[23] Quinlan, J. R., *C4.5: Programs for Machine Learning* San Mateo, CA: Morgan Kaufmann, 1993.
[24] Wieczorkowska, A. Classification of musical instrument sounds using decision trees. Proc. of the 8th International Symposium on Sound Engineering and Mastering, ISSEM'99, pp. 225-230, 1999.
[25] Foote, J. T. A Similarity Measure for Automatic Audio Classification. Proc. of the AAAI 1997 Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora. Stanford, 1997.
[26] Vapnik, V. *Statistical Learning Theory*. New York: Wiley. 1998.
[27] Marques, J., "An automatic annotation system for audio data containing music." BS and ME thesis. MIT. Cambridge, MA, 1999.
[28] Cemgil, A. T. and Gürgen, F. Classification of Musical Instrument Sounds using Neural Networks. Proc. of SIU97. 1997.
[29] Kohonen, T., *Self-Organizing Maps* Berlin: Springer-Verlag, 1995.
[30] Feiten, B. and Günzel, S., "Automatic indexing of a sound database using self-organizing neural nets," *Computer Music Journal*, 18 (3), pp. 53-65, 1994.
[31] Cosi, P., De Poli, G., and Lauzzana, G., "Auditory Modelling and Self-Organizing Neural Networks for Timbre Classification," *Journal of New Music Research*, 23, pp. 71-98, 1994.
[32] Cosi, P., De Poli, G., and Parnadoni, P. Timbre characterization with Mel-Cepstrum and Neural Nets. Proc. of the 1994 ICMC, pp. 42-45, 1994.

[33] Toiviainen, P., Tervaniemi, M., Louhivuori, J., Saher, M., Huotilainen, M., and Näätänen, R., "Timbre Similarity: Convergence of Neural, Behavioral, and Computational Approaches," *Music Perception*, 16 (2), pp. 223-241, 1998.
[34] Wessel, D., "Timbre space as a musical control structure," *Computer Music Journal*, 3 (2), pp. 45-52, 1979.
[35] Carpenter, G. A., Grossberg, S., and Reynolds, J. H., "ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-orgnising neural network," *Neural Networks*, 4, pp. 565-588, 1991.
[36] Fragoulis, D. K., Avaritsiotis, J. N., and Papaodysseus, C. N. Timbre recognition of single notes using an ARTMAP neural network. Proc. of the 6th IEEE International Conference on Electronics, Circuits and Systems. Paphos, Cyprus. 1999.
[37] Kostek, B., *Soft computing in acoustics: applications of neural networks, fuzzy logic and rough sets to musical acoustics* Heidelberg: Physica Verlag, 1999.
[38] Kostek, B. and Krolikowski, R., "Application of artificial neural networks to the recognition of musical sounds," *Archives of Acoustics*, 22 (1), pp. 27-50, 1997.
[39] Kostek, B. and Czyzewski, A. An approach to the automatic classification of musical sounds. Proc. of the AES 108th convention. Paris. 2000.
[40] Dubnov, S., Tishby, N., and Cohen, D., "Polyspectra as Measures of Sound Texture and Timbre," *Journal of New Music Research*, vol. 26, no. 4, 1997.
[41] Pawlak, Z., "Rough sets," *Journal of Computer and Information Science*, 11 (5), pp. 341-356, 1982.
[42] Pawlak, Z., "Rough set elements," in Polkowski, L. and Skowron, A. (eds.) *Rough Sets in Knowledge Discovery* Heidelberg: Physica-Verlag, 1998.
[43] Czyzewski, A., "Soft processing of audio signals," in Polkowski, L. and Skowron, A. (eds.) *Rough Sets in Knowledge Discovery* Heidelberg: Physica Verlag, 1998, pp. 147-165.
[44] Polkowski, L. and Skowron, A., *Rough Sets in Knowledge Discovery* Heidelberg: Physica-Verlag, 1998.
[45] Kostek, B., "Soft computing-based recognition of musical sounds," in Polkowski, L. and Skowron, A. (eds.) *Rough Sets in Knowledge Discovery* Heidelberg: Physica-Verlag, 1998.
[46] Kostek, B. and Wieczorkowska, A., "Parametric representation of musical sounds," *Archives of Acoustics*, 22 (1), pp. 3-26, 1997.
[47] Wieczorkowska, A., "Rough sets as a tool for audio signal classification," in Ras, Z. W. and Skowron, A. (eds.) *Foundations of Intelligent Systems: Proc. of the 11th International Symposium on Foundations of Intelligent Systems (ISMIS-99)* Berlin: Springer-Verlag, 1999, pp. 367-375.
[48] Scheirer, E. D., "Music-Listening Systems." Ph.D. thesis. MIT. Cambridge, MA. 2000.
[49] Kashino, K. and Murase, H. Music recognition using note transition context. Proc. of the 1998 IEEE ICASSP. Seattle. 1998.
[50] Cooke, M., *Modelling auditory processing and organisation* Cambridge: Cambridge University Press, 1993.
[51] Elder IV, J. F. and Ridgeway, G. Combining estimators to improve performance. 1999. Proc. of the 5th International Conference on Knowledge Discovery and Data Mining. 1999.
[52] Ellis, D. P. W. Improved recognition by combining different features and different systems. To appear in Proc. of the AVIOS-2000, San Jose, CA. May, 2000.
[53] Rabiner, L. R. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proc. of the IEEE, 77, pp. 257-286. 1989.
[54] Zhang, T. and Jay Kuo, C.-C. Heuristic approach for generic audio data segmentation and annotation. ACM Multimedia Conference, pp. 67-76. Orlando, FLA. 1999.
[55] Michie, D., Spiegelhalter, D. J., and Taylor, C. C., *Machine Learning, Neural and Statistical Classification.* Chichester: Ellis Horwood; 1994.