Applications of Soft Computing for Musical Instrument Classification

Daniel Piccoli, Mark Abernethy, Shri Rai, and Shamim Khan

Murdoch University, Western Australia
{dpiccoli,mark.abernethy,smr,s.khan}@murdoch.edu.au

Abstract. In this paper, a method for pitch independent musical instrument recognition using artificial neural networks is presented. Spectral features including FFT coefficients, harmonic envelopes and cepstral coefficients are used to represent the musical instrument sounds for classification. The effectiveness of these features are compared by testing the performance of ANNs trained with each feature. Multi-layer perceptrons are also compared with Time-delay neural networks. The testing and training sets both consist of fifteen note samples per musical instrument within the chromatic scale from C3 to C6. Both sets consist of nine instruments from the string, brass and woodwind families. Best results were achieved with cepstrum coefficients with a classification accuracy of 88 percent using a time-delay neural network, which is on par with recent results using several different features.

Keywords: neural networks, musical instrument recognition

1 Introduction

With the advent of digital multimedia, there is an increasing need to be able to catalogue audio data in much the same way that books are catalogued. Most digital audio formats in use today such as MP3 and WAV contain limited meta-data about the actual recordings that they contain [1]. However, the MPEG-7 specification requires that meta-data, such as the types of musical instruments in a recording, should be stored in the file with the recording to enable effective cataloguing of files [2]. The classification and identification of important features of musical instruments in digital audio will be a step towards such a cataloguing system. The process of classification based on a set of features is often referred to as 'Content-Based Classification [1].'

Meta-data generated by such a system may be used by a search engine to allow users to find specific styles of music. For example, a music teacher may be interested in searching for audio files with certain instruments playing or music from a certain genre. Currently, musical search systems only have the ability to classify their music based on filenames.

The model discussed in this paper concentrates on identifying musical instruments in sound recordings whilst assessing the usefulness of different features

T.D. Gedeon and L.C.C. Fung (Eds.): AI 2003, LNAI 2903, pp. 878–889, 2003.

that can represent musical instruments for the purpose of classification. Meta-data gathered by an effective instrument classification system can be used to build databases based on MPEG-7 meta-data. Another possible use of such an audio classification system include the ability to automatically generate meta-data about music files and to fight the distribution of copyright audio material on the Internet.

The main aim of this research was to compare the usefulness of spectrum and cepstrum based features using an artificial neural network. Also, the performance of the time delay neural network and the multi-layer perceptron were compared. The results may lead to an improved cataloguing system that allows people to search through music databases more effectively.

2 Previous Work

Two major characteristics of an audio classification system are the features that are used to distinguish between the sounds and method in which the classifier is 'trained' to recognise the sounds. A trained human ear can easily identify musical instruments that are playing in a sound mix, even if each of these instruments share a similar note range. This is because each instrument has a set of auditory features that distinguishes it from other instruments. For example, many musical instruments have harmonics or overtones that can be heard at multiples of the fundamental frequency. These harmonics colour the sound making each instrument sound different. The features that enable one to distinguish musical instruments can be described as the timbre of the sound. At present, computerised audio classifiers lack the accuracy of human classifiers. As a result, research is being conducted in improving the features that are fed to audio classifiers as well as the audio classification engines themselves (eg. ANNs).

Audio Classification Techniques

Herrera [2] provides a survey of different techniques that have been used to classify musical instruments in monophonic (where one instrument plays only one note at any given time) sounds. Techniques discussed include K-nearest neighbours, Bayesian classifiers, binary trees, support vector machines and neural networks. Comparability between these techniques is difficult due to differing experimentation approaches and sound samples used. Also, many of these experiments have been based on a limited set of musical instruments. However, the attractiveness of the ANN comes from its ability to generalise after being trained with a finite set of sound samples. Herrera also alludes to the difficulty involved with using the same algorithms for identifying musical instruments in sound mixes.

Experiments performed using ANNs to classify musical instrument sounds have found that good results can be achieved for monophonic instrument samples [3, 4]. Cemgil [4] provides a comparison between a multi-layer perceptron, a time delay neural network and a hybrid Self-Organising Map/Radial Basis

Function (RBF) for classifying instrument sounds. All network types are presented sound in the form of a set of harmonic envelopes. The number of instrument harmonics required for good neural network generalisation performance is also discussed. The results indicated that generalisation improves when more harmonics are presented to the neural networks. The number of time windows had less bearing on the performance of each of the models. It was found that the spectral content in the attack portion of the sound contained most of the information needed by the network to make a classification.

Cemgil's experiment found that the performance of each of the models in order of success were: the Time Delay Neural Network with classification success of up to 100%; the Multi-layer perceptron with a classification success of up to 97%, and lastly the hybrid Self Organising Map (SOM)/Radial Basis Function with a classification success of up to 94%. However, the results shown by the SOM are promising because they show how the instruments are organised according to timbre. Unfortunately, these results are optimistic for the classification of real life sounds. The sound samples included a limited range of notes (one octave) and limited instrument articulation (eg. a cello string can be plucked or bowed).

Experiments detailed in this paper test the adequacy of the standard back-propagation Multi-Layer Perceptron (MLP) the Time Delay Neural Network (TDNN) for detecting the presence of a musical instrument in a monophonic source regardless of the note played or its volume. Each ANN is fed a series of spectral parameters based on the Fourier Transform.

Selection of Auditory Features

A digital audio signal must be converted into a suitable form before classification becomes possible. Auditory features can be classified as spectral or temporal. Spectral features are parameters that can be extracted from the frequency spectrum of a sound whereas temporal features relate to the timing of events over the duration of a note.

When using multiple features for classifying musical instrument, it is possible for one bad feature to destroy the classification results. It is therefore important to determine whether a certain feature allows musical instruments to be distinguished. Kostek's [3] work goes some way to identifying instrument distinguishing sound characteristics.

Brown [5] demonstrated the validity of cepstral coefficients for distinguishing between an oboe and a saxophone by using a K-means algorithm. Eronen [6] later verified their robustness in classifying a wider range of instruments.

Having surveyed the literature, it must be noted that different features may be suitable for classifying a small set of instruments. However, a universal set of parameters with the ability to distinguish between any musical instrument is non-existent. Most of the timbral features discussed are based on perceptual models of the ultimate sound classifier, namely the human being. However, the difficulty lies in the objective measurement of these properties using a computer [5]. The human auditory system is a highly non-linear device in which

some harmonic components produced by musical instruments may be masked by spectral energy at nearby frequencies.

Extensive research has been completed on defining timbre in terms of human perception. The definition of timbre in terms of spectral features was apparent as early as 1954, when Helmholtz claimed that the relative amplitudes of the harmonic partials that compose a periodic tone is the primary determinant of a tone's sound quality [7, 8]. Temporal features were recognised as important determinant of musical instrument sounds as early as 1910 when Stumpf recognised the importance of the onset of a musical note for distinguishing musical sounds [8]. For this reason, experiments detailed in this paper involve presenting the ANN with representations of sounds beginning at the onset of a note.

The field of computational auditory scene analysis [9] attempts to mimic the ability of humans to conceptualise music. Abdallah [10] discusses how to extract features that make up our perception of musical sounds. His idea is to create a 'single unifying description of all levels of musical cognition from the raw audio signal to abstract musical concepts.'

This research could lead to improvements in how a neural network is fed sound features in the form of reduced redundancy.
In the meantime, techniques such as Mel scaling (passing audio through a set of band pass filters based on experimental results on human hearing) are used to approximate the human auditory system. Future work in musical classification based on perceptual models may be fruitful.

3 Methodology

Sound samples from a variety or woodwind, brass and string instruments were collected from the University of Iowa music samples web page [11]. Piano samples were also obtained from this site. A synthesised guitar was included in the sample set (from general MIDI) in order to have a representation of another stringed instrument. The note samples, which ranged from C3 to C6 on the chromatic scale, were separated into two sets. One of these sets was used for training and the other for testing. The testing set was used to assess the ability of the ANN to generalise based on a finite number of examples that were presented during training (the training set).

The note range from C3 to C6 was chosen because the instruments selected produce sounds that overlap in these frequencies. Notes above C6 were not included in either the training or test sets because the Nyquist theorem prevents the extraction of higher harmonics for higher notes. To enable the extraction of higher harmonics, a larger audio sampling rate must be used. However, due to processing time constraints, a sampling rate of 22 kHz was used.

It is important to use real life instrument samples for classification to ensure that the ANN can generalise regardless of how the instrument is played. The ANN training set contained a combination of soft notes, moderately loud notes and loud notes. The harmonic compositions of each note change dynamically (not proportionally) depending on how the note is articulated (see figure 1).

This makes the classification process more difficult for live instruments. For synthesized instruments the harmonic amplitudes are generally scaled in proportion to the loudness of the fundamental frequency. The training and testing sets contain examples of notes played vibrato (with modulation) and pizzicato (plucked) notes from the cello.

After the onset of each note is detected (by computer algorithm), 50 percent overlapping windows of 2048 Hann windowed samples were converted to the frequency domain using the fast Fourier transform (FFT). From the FFT coefficients, harmonic envelopes and cepstral coefficients were then calculated. The extracted features were either averaged over a number of frames or taken directly and the resultant feature vector was placed into a pattern file ready for presentation to the ANN. For the cepstral coefficients, only the first 128 coefficients were presented to the neural network. The lower coefficients correspond to the rough or coarse spectral shape. The rough spectral shape contains information of the formants or resonances of the instrument body [6], thus these were used as features. The calculation of cepstral coefficients generally involves taking the spectrum of a log spectrum [12].

These features were fed into both a multi-layer perceptron (MLP) and a time-delay neural network. A limitation of the multi-layer perceptron comes from the fact that the entire pattern has to be presented to the input layer at once. The nature of sound is that variations occur over time. For example, musical sounds have attack, sustain and decay transients that are difficult to represent. The time delay neural network is an extension of the MLP back propagation network developed by Alex Waibel [13] that has the advantage of being able to learn patterns that vary over time [14]. TDNNs have proven to be useful for musical instrument identification [4]. The time-delay neural network, like the perceptron, uses back-error propagation as its learning algorithm. However, the TDNN has a variety of time-delay components built into its architecture as described by Waibel [15, 13].

The convergence of each Neural Network towards the correct classification was assessed by monitoring the 'mean squared error' (MSE) of the training data and comparing it to the MSE of the test data after each epoch. Training was stopped once the MSE of the test no longer improved. The validity of the features (harmonic envelopes vs cepstral coefficients) was assessed by comparing the recognition accuracy of ANNs using each of these features.

4 Results

4.1 Instrument Classification with Neural Networks

MLPs Trained with Harmonic Amplitudes

An MLP was initially trained with examples of individual frames (no averaging over a number of frames) with 50% overlapping windows shifted across the entire sound file for each instrument. Therefore, the number of neurons in the input layer was equal to the number of harmonics presented to the neural network.

That is, the ANN was presented with the harmonic amplitudes of one frame at a time (not a harmonic envelope). This tests the ability of the ANN to recognise a musical instrument given just one frame.

Results were modest with a best classification score of 67%. However, these results suggest that the neural network struggled to identify musical instruments correctly given just one frame at a time. This may be due to the fact that harmonic content is not consistent across the entire length of a musical note, or across different notes or articulations produced by the same instrument. This makes it difficult for the ANN to generalise. Also, based on research regarding phenomena that influence timbre perception [16], frequency information in the attack transient of a musical note is important for classification. In the instances where the sliding window appears over a non-attack part of a note, instruments may be more difficult to distinguish.

There seems to be an indicative trend that by increasing the number of harmonics presented to the neural network, the ANN will be better able to distinguish between instrument sounds. Unfortunately, there were not enough test runs available to statistically verify this claim due to the Nyquist limitation (insufficient number of harmonics). However, this trend is consistent with findings indicated by Cemgil [4] and Kostek [3].

MLPs Trained with Harmonic Envelopes

The MLP was then presented with harmonic envelopes from several adjacent sliding windows starting from the onset of each musical note. In other words, the ANN was presented with harmonic amplitudes for several 50% overlapping frames beginning from the onset of each note. Each window was individually normalised (note that this is additional normalisation for each frame that was performed in addition to the normalisation of the entire note during pre-processing). It was hoped that a trend could be shown between the number of frames presented to the MLP and the ability of the MLP to generalise. However, the addition of new frames to the input vector failed to improve the classification results of the ANN. As shown in table 8 in appendix A, the best result attained by presenting multiple frames of harmonic amplitudes (harmonic envelope) was 65%. These results are well below those attained by Cemgil [4] using harmonic envelopes and are also inferior to the results obtained for this research for ANNs presented with harmonic amplitudes from only one frame.

These poor results indicate that the ANN failed to generalise to the general (test set) population given harmonic envelopes. There are two possible causes for this. Firstly, if the number of hidden layer nodes were excessive, the ANN is likely to learn the nuances of the training set rather than the major features that distinguish the instrument patterns. The second possible cause of the problem occurs when the training sample is not representative of the general population. For example, the trombone samples given in the training set may not be representative of all trombone sounds. Since the number of hidden layer nodes was carefully chosen, the former is unlikely to be the problem. Analysis of error during training revealed a disappointing relationship between test and training

data. Ideally, if the ANN was learning to distinguish musical instruments based on their harmonic content, then both training and testing errors would have decreased at a similar rate.

TDNN Trained with Harmonic Envelopes

When the Time-Delay Neural Network (TDNN) was presented with the harmonic envelope as an input vector, the classification success for the best performing TDNN increased to only 68%. This result is well below the 100% achieved by Cemgil (Cemgil et al., 1997). However, this may be due to the fact that every individual frame was normalised. To improve this experiment, the harmonic envelope should have been normalised as a whole. By normalising each individual frame, important information regarding change in volume associated with the onset of a note may be lost.

4.2 Supervised Neural Networks Using Cepstral Coefficients as Features

MLP Trained with Cepstral Coefficients

Results for MLPs, which were disappointing with a classification success of just 53 percent. It became apparent that a multi-layer perceptron is unable to distinguish musical instruments very well when trained with just one set of Cepstral coefficients per note. Thus, in order to reduce the effect of noise when measuring the cepstrum of a note, the average of several frames over each individual note were taken. The best result attained from averaging the Cepstrum over a number of frames is depicted in table 1.

Improvements to the classification ability of the MLP as a consequence were excellent. After interpretation of the results, it became evident that the ANN is better able to generalise when the Cepstrum is averaged over an increasing number of frames. This implies that the average of a number of cepstral coefficients at the onset of each note is more representative of a particular instrument than cepstral coefficients in isolation. More research is needed to examine whether this phenomenon continues beyond 7 frames. However, classification results of up to 85% for the MLP using cepstral coefficients averaged over 12 frames (at the onset of each note) indicate that 128 cepstral coefficients are a valid and robust input vector option for neural networks.

TDNN Trained with Cepstral Coefficients

The TDNN presented with cepstral coefficients (the first 128 coefficients) proved to be the most successful architecture for classifying musical instruments. Each TDNN was configured to accept 10 fifty percent overlapping delayed frames containing 128 cepstral coefficients. The TDNN was given 8 such examples per note. As depicted in table 2, the TDNN successfully classified up to 88% of instruments correctly.

Table 1. Matrix of 9 instruments classified with an MLP using the first 128 cepstrum coefficients averaged over 12 frame(s) at the onset of each note

Actual Class/Target	Bass.	Flute	Clar.	Trom.	Horn	Oboe	Piano	Guit.	Cello
Bassoon	14	0	0	0	0	0	0	0	0
Flute	0	14	1	3	0	0	0	0	0
Clarinet	0	1	14	0	0	2	0	0	0
B. Tr	0	0	0	8	0	1	0	0	0
Horn	1	0	0	0	15	0	0	0	0
Oboe	0	0	0	4	0	12	0	0	0
Piano	0	0	0	0	0	0	8	0	0
Guitar	0	0	0	0	0	0	2	15	0
Cello	0	0	0	0	0	0	5	0	15
Total	15	15	15	15	15	15	15	15	15
Percentage Correct	**93**	**93**	**93**	**53**	**100**	**80**	**53**	**100**	**100**
Total Correct	**115**								
Total % Correct	**85**								

Table 2. Matrix of 9 instruments classified with a TDNN Using the first 128 cepstrum coefficients delayed over 10 frames at the onset of each note (best performance)

Actual Class/Target	Bass.	Flute	Clari.	Trom.	Horn	Oboe	Piano	Guit.	Cello
Bassoon	105	0	0	0	0	0	0	12	0
Flute	0	102	7	0	0	0	0	0	0
Clarinet	0	1	87	0	0	10	0	0	0
B. Tr	0	0	0	103	4	0	2	0	0
Horn	0	0	0	2	98	0	0	0	0
Oboe	0	0	0	0	0	95	0	25	0
Piano	0	0	0	0	0	0	85	15	0
Guitar	0	0	0	0	0	0	14	54	2
Cello	0	2	4	0	3	0	4	0	103
Total	**105**	**105**	**105**	**105**	**105**	**105**	**105**	**105**	**105**
Percentage Correct	**100**	**97**	**83**	**98**	**93**	**90**	**81**	**51**	**98**
Total Correct	**88**								
Total % Correct	**832**								

4.3 Discussion of Results

Comparison of Feature Vectors

From the above analysis, it appears that averaged cepstral coefficients (with classification success up to 88%) are a much more useful input vector to a classification algorithm (such as neural networks) than harmonic envelopes (with

classification success of up to 68%). This may be indicative of the fact that harmonic information alone is insufficient for distinguishing between musical instruments.

For classification with harmonic amplitudes to be possible, the relationship between the harmonics must be consistent for a given instrument class (eg. bassoon) regardless of the note played or its articulation. Figure 1 below depicts the harmonic structure of two different bassoon notes used in the experiments:

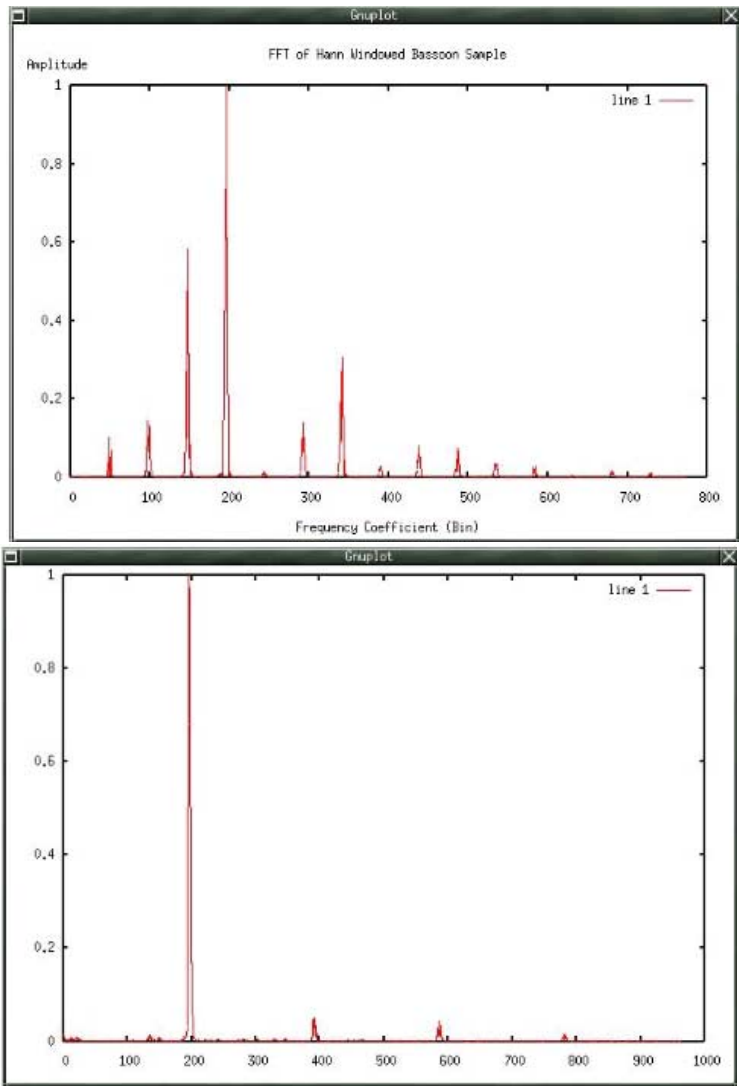

Fig. 1. Comparison of two bassoon notes with different articulation

The graphs in figure 1 depict harmonic peaks of a 2048-sample frame gathered from the onset of two bassoon notes. Each frame has been normalised to 1 so that the relationship between harmonic amplitudes for each note can easily be seen. From these diagrams, it is evident that the harmonic pattern is not consistent for each bassoon note. In the top diagram, the third harmonic has the highest amplitude whereas the second diagram shows the fundamental as having the highest amplitude. These inconsistencies may limit the ability of an ANN to distinguish between instruments using harmonic amplitudes exclusively. Future studies may be interested in testing whether the process of averaging harmonic envelopes improves the classification process as it did for cepstral coefficients.

The relatively poor results attained using harmonic envelopes are not comparable to results from Cemgil's [4] experiments. Differences in the training and test-set data may be the cause for these discrepancies. For example, it appears that Cemgil used sound samples from the standard AWE32 (sound card) set. This makes classification using harmonic envelopes alone feasible due to consistency in the harmonic envelopes of synthesised sounds. As stated by Kaminskyj [17], it is important for the sound classification research community to make their results as comparable as possible in future.

Comparison of Classifiers

In terms of the feature classifiers themselves, the TDNN consistently returned better results than the MLP. The TDNN had classification results of up to 68% using harmonic envelopes and 88% using cepstral coefficients while the MLP had classification results of up to 67% using harmonic envelopes and 85% using cepstral coefficients as features. This is indicative of the importance of temporal features for classifying musical sounds. Further research is needed to confirm this statistically.

One advantage of using the MLP as opposed to the TDNN came from the fact that it completed training much more quickly. However, this was influenced by the experiment technique used. For the MLP, cepstral coefficients were averaged over several frames. Therefore, the number of input layer nodes was equal to the window size (or 128 for the ANNs designed to accept the first 128 cepstrum coefficients). However, for the TDNN, 10 individual frames (delays) of 128 cepstral coefficients were applied to the input layer, totalling 1280 input layer nodes. This slowed the training process for the TDNN.

Identification of Timbral Families

One of the aims of these experiments was to identify timbral families and relationships between the musical instruments. This can reveal whether or not the computer 'perceives' musical instruments in a similar way to the human auditory system. Table 3 provides totals of the most common misclassifications with the TDNN architecture using cepstral coefficients as features:

This data can be used to reveal relationships inherent between the musical instruments. For example, the table reveals that the flute and the clarinet,

Table 3. Common Instrument Misclassifications

Instrument	Most Commonly Misclassified As	No. Of Instances
Bassoon	Bass Trombone	2
Flute	Clarinet	12
Clarinet	Flute	44
Trombone	Horn	90
Horn	B. Trombone	10
Oboe	Clarinet	29
Piano	Guitar	77
Guitar	Bassoon	59
Cello	B. Trombone	34

both woodwind instruments are commonly misclassified as one another. Also, the two brass instruments, namely the bass trombone and the horn are commonly misclassified. The only two instruments that significantly did not exhibit the expected misclassification pattern were the guitar (string), which was most commonly misclassified as a bassoon (woodwind), and the cello (string), which was most commonly misclassified as a trombone (brass). The fact that most instruments were misclassified as instruments within their own orchestral (timbral) family indicates that cepstral coefficients have some relationship with the perception of timbre by humans.

5 Conclusions and Future Research

The initial aim of this research was to compare the usefulness of cepstral coefficients and harmonic envelopes as inputs to a neural network for the purpose of classifying musical instruments. Results have indicated that cepstral coefficients may be more useful than harmonic envelopes for the purpose of distinguishing between musical instruments. This research has also demonstrated the usefulness of the MLP and TDNN as classification tools.

Future work may involve combining Cepstral coefficients with spectrum related parameters such as spectral brightness and odd/even harmonic components in order to produce a better classification model. Temporal features must also be carefully analysed for their usefulness in classifying musical sounds. Further experiments involving the human perception of sound may also be fruitful for devising a better way to present a classifier with sound data.

The results attained for this research were limited to monophonic sounds. For classification models discussed in this paper to be useful, prior separation of sounds is required. The ultimate goal of audio classification is to identify musical instruments that exist in polyphonic sounds. This may be achieved in the future by using a technique known as source (or stream) separation. Future research in this area will improve the viability of classifying musical instruments in polyphonic sounds.

References

[1] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Classification, search and Rereval of audio," *IEEE Multimedia Magazine, vol. 3,* no. *3, pp. 27-36, 1996.* 878

[2] P. Herrera, X. Amatraiain, B. E., and X. Serra, "Towards instrument segmentation for music content description: A critical review of instrument classification techniques," in *Proceedings of the International Symposium On Music Information Retreival, 2000.* 878, 879

[3] B. Kostek and R. Krolikowski, "Application of artificial neural networks to the recognition of musical sounds," *Archives of Acoustics, vol. 22,* no. 1, pp. *27-50, 1997.* 879, 880, 883

[4] A. Cemgil and F. Gurgen, "Classification of musical instrument sounds using neural networks," *Proceedings of SUI97, 1997.* 879, 882, 883, 887

[5] J. Brown, "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features," *J. Acoust. Soc. Am., vol. 105, pp. 19331941, 1999.* 880

[6] A. Eronen and A. Klapuri, "Musical instrument recognition using cepstral coefficients and temporal features," *Proceedings of the IEEE International Conference an Acoustics, Speech and Signal Processing, 2000.* 880, 882

[7] H. Helmholtz, "On the sensations of tone as a physiological basis for the theory of music," *Dover, A. J. Ellis Trans, 1954.* 881

[8] K. Martin and Y. Kim, "Musical instrument identification: A pattern-recognition approach," in *136th Meeting of the Acoustical Society of America,* October 1998. 881

[9] A. S. Bregman, "Auditory scene analysis: The perceptual organisation of sound," *MIT Press, 1990.* 881

[10] S. Abdallah and M. Plumbley, "Unsupervised learning in neural networks for auditory perception and music cognition," *Cambridge Music Processing Colloquium,* Sept 1999. 881

[11] L. Fritts, "Musical instrument samples web page," tech. rep., University of Iowa, URL: http://theremin.music.niowa.edu/ web/sound/, 2002. 881

[12] R. Randall, *Frequency Analysis.* Naerum: Bruel and Kjaer, 1987. 882

[13] A. Waibel, T. Hanazawa, G. Hinton, K. Schikano, and K. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions in Acoustics, Speech and Signal Processing 37, 1989.* 882

[14] J. Dayhoff, *Neural Network Architectures - An Introduction.* New York: Van Nostrand Reinhold, 1990. 882

[15] A. Waibel, "Consonant recognition by modular construction of large pnonetic timedelay neural networks," *Neural Information Processing Systems, 215-223.* 882

[16] J. Grey, "Multidimensional perceptual scaling of musical timbres," *J. Acoust. Soc. Am. 61(5), pp. 1270-1271, 1977.* 883

[17] I. Kaminskyj, "Multi-feature musical instrument sound classifier w/user determined generalisation performance," in *Proceedings of ACMA02, 2002.* 887