Automatic source identification of monophonic musical instrument sounds

Conference Paper · December 1995

DOI: 10.1109/ICNN.1995.488091 · Source: IEEE Xplore

2 authors, including:

Andrzej Materka
Lodz University of Technology
173 PUBLICATIONS **3,366** CITATIONS

SEE PROFILE

AUTOMATIC SOURCE IDENTIFICATION OF MONOPHONIC MUSICAL INSTRUMENT SOUNDS

I. Kaminskyj * and A. Materka+
* Electrical & Computer Systems Engineering, Monash University, Caulfield Campus, Vic 3145, Australia
+Politechnika Lodzka, Instytut Elektroniki, Stefanowskiego 18, 90-924 Lodz, Poland

1. ABSTRACT

A system has been developed which automatically identifies the source of monophonic musical instrument sounds. Preprocessing of sound recordings includes calculation of the short term RMS energy envelope, Principal Component Analysis and Ratio/Product transformations of the resultant Principal Components. An Artificial Neural Network and a Nearest Neighbour Classifier were compared to determine which provided optimum classification ability. The system performance was tested on sounds recorded from four musical instruments chosen to represent each of the major musical instrument families and playing notes over the range of one octave under varying volume conditions. Classification accuracies in the range 93.8 - 100 % were achieved.

2. INTRODUCTION

Much research on automatic music transcription has focused on note identification during a musical performance [1], [6]. However, surprisingly little work has been done on identifying the source of each musical note, thereby making the transcription task difficult.

For example, in their paper, Chafe and Jaffe [2] describe a number of techniques for source separation in polyphonic music. These include periodicity estimation, source verification and source coherence. Chowning and Mont-Reynaud [4] elaborated somewhat on Chafe's proposals and suggested that with periodicity estimation technique, chords are separated into source hypotheses and groups of partials are then tracked in time. Each source hypothesis is compared to a model describing the distinct features of different instrument acoustics. In this way, identification of the musical instruments becomes a possibility. Unfortunately, few practical results of this work have been published, even for monophonic work. Therefore, to verify the usefulness of these techniques in any quantifiable way is still difficult.

This paper provides the results of some preliminary work carried out to explore the potential of artificial neural network (ANN) and nearest neighbour classifiers (NNC) to solve this problem with monophonic musical instrument sounds. Four instruments were studied, chosen to represent each of the major musical instrument families. Recordings made were preprocessed by calculating the short term RMS energy envelope, performing Principal Component Analysis and determining Ratio/Product transformations of the resultant Principal Components. The two classifiers were then presented with this information to determine which provided the best classification results.

3. DATA COLLECTION

The four instruments used cover the note range C4-C5, have adjustable volume and are non synthesised. They were the guitar, piano, marimba and accordion. The guitar represents a plucked string instrument, the piano a struck keyboard string instrument, the marimba a percussion instrument of definite pitch and the accordion a double mechanical reed wind instrument [11].

The note range used was C4-C5 of the equally tempered musical scale, with C4 defined as middle C on a piano, with a frequency of 261.6 Hz. In this scale, each musical octave comprises twelve notes. A one octave note range was chosen to limit the complexity of the system, to provide practical limits on the number of recordings required and to simplify the selection of instruments, remembering that all instruments needed to cover the same range of notes.

As the characteristics of sounds produced vary with the volume of the sound [10], the recordings were made at five different volumes, corresponding to an equivalent musical dynamic range from '*p*' (soft) to '*f*' (loud), or a variation of 20 dB [8], in order to make the system more robust in dealing with 'real' instruments.

0-7803-2768-3/95/$4.00

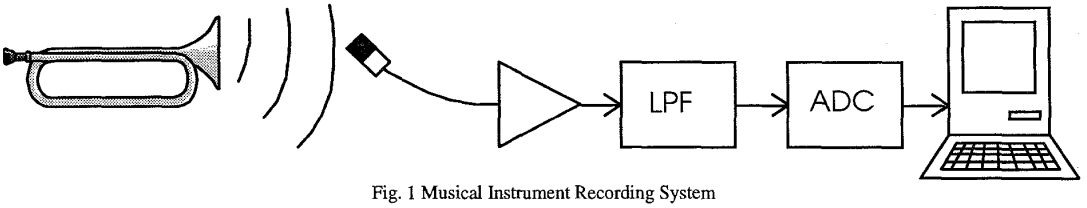

Fig. 1 Musical Instrument Recording System

In the recording system shown in Fig. 1, each instrument produces sounds that are converted into electrical signals by the microphone. The low amplitude signal (mV) is then amplified (V) and low pass filtered by a 5th order Bessel anti-aliasing filter with a nominal cutoff frequency of 10 kHz. The filtered signal is then converted to digital form by an A to D converter at a sampling rate of 32 kHz before being stored on disk for displaying, editing and further processing.

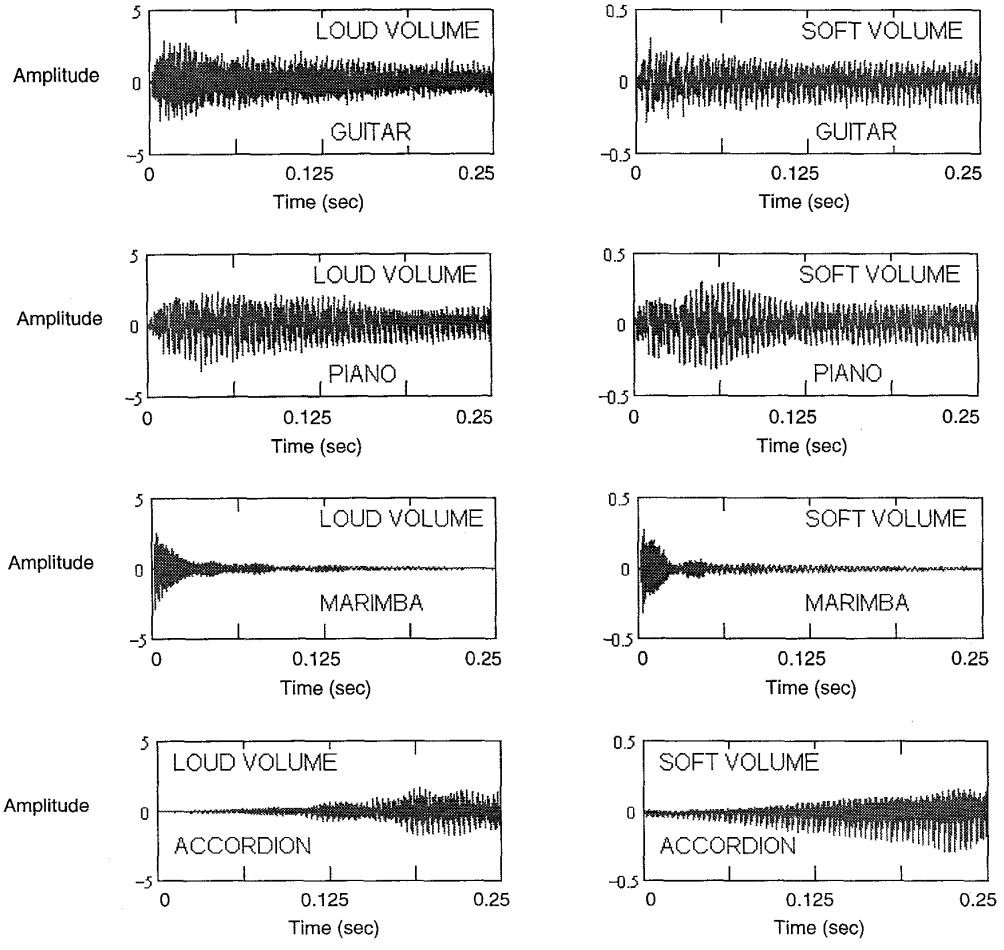

Fig. 2 Typical Waveforms for Loud/Soft Volumes Note C4

Fig. 2 shows the first 0.25 second of typical waveforms recorded for each of the four instruments. Notice the differences in the waveform attack, sustain and decay characteristics. Notice also the expected similarities between piano and guitar recordings, since both are stringed instruments.

4. DATA PREPROCESSING

Once the raw musical instrument recordings have been edited, they need to be preprocessed before classification by the ANN/NNC.

4.1 Short Term RMS Energy

The calculation of the short-term RMS energy for each instrument waveform provides important information as to its temporal evolution. Each short-term RMS energy calculation is performed over 400 samples at a time. This covers at least 3 periods of the fundamental of the waveform.

Fig. 3 shows the short-term RMS energy vs time for the first 0.25 second of sound.

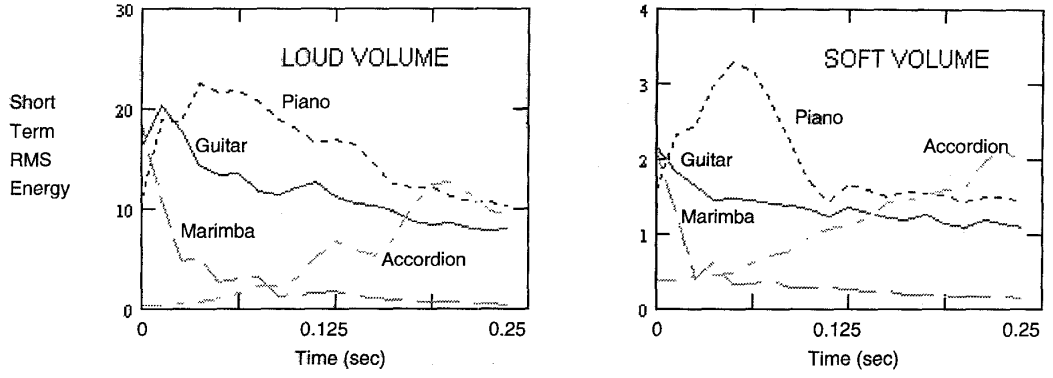

Fig. 3 Typical Short-term RMS Energy Waveforms for Loud/Soft Volumes, Note C4

Notice the marked variations in attack, steady state and decay portions for each instrument. Volume variation also has a significant effect and the distinctive features of the waveforms would make instrument recognition fairly straight forward.

Calculating the short-term RMS Energy for every 400 samples over the duration of 1 second (32000 samples at a sampling rate of 32 kHz) provides 80 short-term RMS Energy values for each musical signal.

4.2 Principal Component Analysis

The aim of principal component analysis (PCA) is to reduce the dimensionality of a data set which consists of a large number of interrelated variables, while retaining as much as possible the variation present in the data set [9]. This is achieved by transformation to a new set of variables, the principal components (PCs), which are uncorrelated and ordered so that the first few retain most of the variation present in all of the original variables.

If PCA were not utilised here, the ANN/NNC would need to analyse 80 short-term RMS Energy values for each musical signal. With 80 input values, the ANN would become relatively large, and subsequent training and testing laborious and slow.

The PCs for the 80 short-term RMS Energy values were calculated using the correlation matrix for the data set, **C**, with the following :

$$\mathbf{z} = \mathbf{A}^{\mathrm{T}}\mathbf{x}^{*} \qquad (1)$$

where : **z** is a column matrix containing the PCs themselves,
A is a matrix which has columns consisting of the eigenvectors of the correlation matrix, **C**, and
x* is a column matrix containing a standardised version of the data set vector **x**.

Using equation (1) to determine the PCs and then applying Kaiser's rule [9], only 3 PCs are retained for the short-term RMS Energy data. For these 3 PCs, the cumulative percentage of total variation represented was 88.8%. Hence the dimensionality of the input data vector to the ANN/NNC has been reduced from 80 to 3, a considerable saving.

To obtain some idea of the difficulty of discrimination of the four instruments based on these 3 PCs, XGOBI, an interactive dynamic graphics program for data visualisation developed at Bellcore, was used to plot a 3D scatter plot of the 3 PCs for the ANN/NNC training data set, shown in Fig. 4. Apart from some overlap between the guitar and piano clusters, fairly good separation exists between instruments. This means the approach chosen for the system has potential.

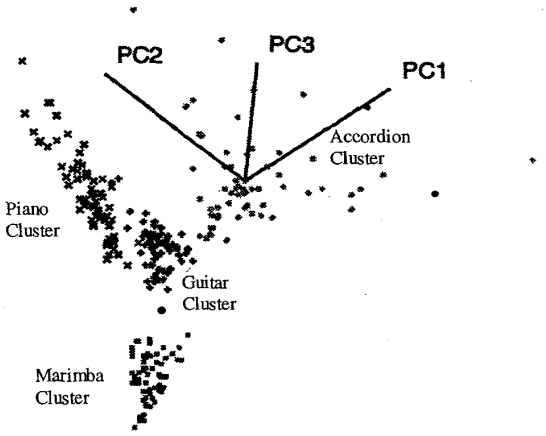

Fig. 4 3D Scatter Plot of 3 Short-term RMS Energy Principal Components

5. ARTIFICIAL NEURAL NETWORKS

A multilayer perceptron ANN with sigmoidal nonlinearity using the Back Propagation training algorithm was used as the classification paradigm. This was utilised primarily due to its widespread use, popularity and successful application by other researchers in performing similar classification tasks [7], [13]. Its architecture was designed using the guidelines suggested by Chen [3].

With four instruments being classified, four output layer units (M) were required. With three short term RMS Energy PCs, three input layer units (N) were necessary. A single hidden layer (HL) was utilised, with the initial number of units (h) being determined using Maren's suggestion, i.e., $h = \sqrt{(NxM)} \approx 4$ [3]. The optimum architecture however was finally determined using empirical results.

As mentioned above, recordings were made over a note range C4-C5, comprising thirteen notes. Each note was recorded at five different volume levels, with three recordings at each volume level for each note of each instrument. The training set comprised one example of each note recorded at each of five different volume levels for each instrument, while the test set comprised of a different example of each note recorded at each of five different volume levels for each instrument. As a result, both the training and test sets comprised 260 cases each.

During training of the ANN, pattern learning was used in preference to batch learning. A learning rate, $a = 0.25$ and a momentum term, $\eta = 0.15$ were used with less than 30,000 presentations to the ANN and with $a = 0.15$ and $\eta = 0.08$ beyond this value. All input data were scaled to the range 0.1 - 0.9. During testing, the ANN outputs (x) were thresholded at a value of 0.5. For $0 < x < 0.5$, a zero result was assumed. Conversely, for $0.5 \leq x < 1.0$, a result of 1.0 was assumed.

6. NEAREST NEIGHBOUR CLASSIFIER

To measure the effectiveness of the ANN classifier, a NNC ($k = 1$) [5] was also implemented, using the Euclidean distance measure. Classification was obtained by determining the closest member in a reference data set to any unknown case. The instrument type into which this closest member falls is deemed the instrument type for the unknown case.

Initially, the training data set, described above for the ANN development, was used as the reference data set. Classification performance was determined using the testing data set. A technique described by Chen [3] for compressing the size of this reference data set was then attempted to determine its minimum size without markedly compromising classification performance. The minimum reference data set was determined by storing only those samples of the training data set that would not be otherwise correctly classified.

7. RESULTS

Table 1 shows the classification performance of different HL size ANNs. Three different types of data transformations were applied to the input data before application to the ANN. They included no transformation as well as ratios and products of the short term RMS Energy PCs. For the ratio and product transformations, logarithms were applied to limit the data dynamic range. For each type of data transformation, with each number of HL units, five different training runs were performed.

Table 1 Artificial Neural Network Classification Results

Data Transformation	Number Hidden Layer Units	Best Training Set Result (%)	Best Test Set Result (%)
None	5	93.1	95.8
"	4	94.6	94.2
"	3	94.6	93.5
"	2	70.4	69.6
Log Ratio	5	97.7	96.5
"	4	98.1	97.7
"	3	97.7	96.2
"	2	74.6	75.0
Log Product	5	94.6	93.8
"	4	93.5	91.9
"	3	91.5	87.3
"	2	74.2	72.7

Table 2 shows the classification performance of the NNC using different reference data sets. Correct classification results are shown both for individual instruments, each comprising 65 cases, as well as for all the instruments taken together. Table 3 shows the number of PC vectors from each instrument class used in the NNC Compressed Reference Data Set

Table 2 Nearest Neighbour Classifier Classification Results

REFERENCE DATA SET	GUITAR	PIANO	MARIMBA	ACCORDION	TOTAL (%)
Complete Training Data Set	65	65	65	65	100
Compressed Training Data Set	61	65	65	64	98.1

Table 3 Nearest Neighbour Classifier Compressed Reference Data Set

GUITAR	PIANO	MARIMBA	ACCORDION
9	10	2	5

8. DISCUSSION

Considering the ANN results obtained, Chen's design procedure appears to have worked well. Very good results were indeed achieved with relatively little effort. The best test set classification results (97.7% or 254/260) were obtained using the logarithmic ratio data transformation with four HL units. Results obtained with the alternate data transformations were nearly as good (95.8% with no data transformation and 93.8% for logarithmic product; both with five HL units). Maren's suggestion of four HL units also proved of value. Only marginal improvements were achieved by increasing the HL units beyond four units.

Perfect ANN classification accuracy is an unreasonable expectation as the 3D scatter plot shows clearly that there does exist some class overlap between the guitar and piano instruments. In fact, analysing the misclassification results for the best performing ANN shows that the only instrument which was misclassified was the guitar.

Considering the NNC results, slightly better results were achieved than with the ANN. Perfect classification accuracy was achieved using the complete training data set for the reference data set, while 98.1% classification accuracy was achieved using the compressed reference data set. In the latter case, misclassifications again occurred primarily with the guitar, for the same reasons as discussed for the ANN. The compression ratio achieved was 10:1 (260 cases -> 26 cases) with only marginal deterioration in performance.

It is not surprising that within the compressed reference data set, so many examples should comprise the guitar and piano cases. These two instruments are the most comparable (Fig. 4) and thus, more reference cases are required to distinguish between them. As the marimba instrument provides the most consistent cases (Fig. 4) it is not surprising that the least number of reference cases are required for it. The accordion, on the other hand, provides the most variable cases (Fig. 4). As expected, it needs a number of reference cases which falls somewhere between that of the guitar/piano and the marimba.

In order to determine how similar the training and test set data were, the Kolmogorov-Smirnov (KS) Statistic was determined for the training/test data sets comparing each pair of Short-term RMS Energy PCs for each instrument [12]. The KS Statistics indicate that there is some similarity between the training and test data sets, but that they are by no means identical. For the marimba, similarity was highest, and this may suggest that for this instrument, the player has the least control over the sounds produced, particularly when the same hammer is used. The accordion, at the other extreme, shows the lowest similarity. This was attributable to its highly variable envelope due to the interference effects of the two slightly detuned reeds which are used. For the piano and guitar, some similarity exists for some of the short term RMS Energy PCs and not for others.

Comparing the results obtained using the ANN and NNC, it can be seen that slightly better results are obtained with the NNC (100% or 98.1% compared with 97.7%). For the complete reference data set NNC results, this is perhaps not surprising as considerably more information regarding the instrument short term RMS Energy PCs is stored than with the ANN (260 x 3 PC values (780) compared with h(NxM) + bias term weights (32)). With respect to the compressed reference data set NNC results, the NNC and ANN results are almost identical. The information utilised for the NNC comprises 78 values compared with 32 weights used for the ANN.

Given the results shown above, the NNC is the classifier of choice with regard to simplicity of implementation and ease of updating when more recordings and instruments are used. For real-time applications however, the ANN is the better choice, with its faster recall rate, assuming the sigmoidal function look-up table is used.

9. CONCLUSION

This paper has described some preliminary work using both ANN and NNC (k=1) classifiers for the automatic source identification of monophonic musical instrument sounds. The results achieved are encouraging although somewhat surprising given that only temporal, but not frequency, data was utilised. This may be due to the limited number of instruments used, the laboratory controlled conditions of the instrument recordings made and the good discrimination obtained with the four instruments chosen.

Further work is necessary to determine how robust the system performance is under a wider variety of recording conditions, with different types of and a greater number of stringed, percussion and wind instruments. Different ANN architectures could be evaluated, for eg. supervised Kohonen networks, to determine which provides optimum classification performance as well as which is the more amenable to electronic implementation. More work could also be carried out on feature selection, so as to better discriminate between sounds of instruments of the same type (example guitar and piano). Spectral features would perhaps provide the necessary information.

10. REFERENCES

1. Brown, J. C., "Musical Fundamental Frequency Tracking using Pattern Recognition Method", J. Acoustical Society America, Vol. 90, No. 3, Sept 1992, pp. 1394-1402
2. Chafe, C., and Jaffe, D., "Source Separation and Note Identification in Polyphonic Music", Proc. ICASSP 86, No. 25.6, 1986
3. Chen, L., "An Intelligent Hybrid System for Fault Diagnostic Problem Solving", Advances in Fault Diagnosis Problem Solving, Ed. Ntuen, C. A., CRC Press, 1994
4. Chowning, J., and Mont-Reynaud, B.,"Intelligent Analysis of Composite Acoustic Signals", CCRMA, Dept. of Music, Report No. STAN-M-36, Stanford University, May 1986
5. Duda, R. O., and Hart, P. E., "Pattern Classification and Scene Analysis", Wiley-Interscience, 1973
6. Foster, S., et al "Toward an Intelligent Editor of Digital Audio: Signal Processing Methods", Computer Music Journal, Vol. 6, No. 1, Fall 1982 pp. 42-51
7. Gorman, R., and Sejnowski, T., "Learned classification of sonar targets using a massively parallel network", IEEE Trans. ASSP, Vol. ASSP-36, 1988, pp. 1135-1140
8. Johnston, I., "Measured Tones, The interplay of physics and music", Adam Hilger, 1989, pp. 368-371.
9. Jollife, I. T., "Principal Component Analysis", Springer-Verlag, 1986.
10. Moorer, J., "On the Transcription of Musical Sound by Computer", CMJ, Vol. 1, No. 4, 1977, pp. 32-38
11. Olson, H. F., "Music, Physics and Engineering", 2nd Ed, Dover Publications, Inc., 1967, pp. 108-9
12. Press, W. H., et al, "Numerical Recipes in C", Cambridge University Press, 1988
13. Ramani, N., et al., "Fish-detection and classification using a neural-network based active sonar system - preliminary results", Proc. IEEE IJCNN, Vol. II, San Diego, 1989, pp. 527-530.