Musical Instrument Classification through Model of Auditory Periphery and Neural Network

Ladislava Jankø, Lenka LhotskÆ
Department of Cybernetics, Faculty of Electrical Engineering
Czech Technical University
TechnickÆ 2, Praha 6, 166 27
CZECH REPUBLIC

Abstract: - This paper deals with the problem of automatic classification of music instruments through model of auditory periphery and neural network. Early recognition and identification of sound source can highly improve the rate of successfulness of computational auditory scene analysis systems and also reduce the computing time. Monaural sound source classification is described in this paper. Digitized sound inputs the front-end component consisting of the model of outer/midle ear transduction, cochlea filter bank, and the model of inner hair cell transduction. Then the tones in signal in each cochlear channel are detected using onsets. For pitch extraction, we use the algorithm based on summary autocorrelation functions. Then the sounds in all cochlear channels are transposed in the ratio given by the original tone fundamental frequency and normalized fundamental frequency (440 Hz). Time domain envelope for each cochlear channel (time-frequency envelope) is extracted and cut to obtain 5000 samples. These samples inputs backpropagation based artificial neural network.

Key-Words: - music instrument classification, auditory scene analysis, artificial neural network, sound processing

1 Introduction

This study is motivated by an interest in the properties of sound signals and modelling of them with application to sound scene analysis and understanding in machine listening. The practical application, which is formulated to this problem, is automatic searching in large music databases containing music data in signal representation or in compressed format. As two key issues appear two tasks: task of computer melody recognition and a task of computer musical instrument classification. Such searching systems could offer to each user a wide range of features, for instance searching of music pieces containing sounds from user specified musical instruments. For this reason, these searching systems could be very user-friendly. Another application of musical instruments classification is an application in the line of audio indexing.

Investigations on the automatic music instrument classification involve mostly application of statistical pattern recognition, artificial intelligence, soft computing or neural nets: for example an application of hidden Markov models [14], which are widely used to speech recognition or the Gaussian mixture model, which was involved as a successful method of automatic speaker identification. The application to musical instrument identification was proposed in [19]. Another approach is related to the fuzzy preferences modelling and application of fuzzy decision mechanism, which involves construction of fuzzy classifier on basis of aggregation of fuzzy preferences. This approach requires feature extraction. Features that appear to be important for musical instrument recognition include: amplitude modulation, amplitude envelope, inharmonicity, spectral centroid, pitch, frequency modulation, onset asynchrony, etc. To extract these features, the channel outputs from cochlear model both in time and frequency domain are analyzed.

Also approach presented in this paper involves physiologically motivated sound pre-processing and feature extraction. It combines model of auditory periphery with the backpropagation based supervised feedforward artificial neural network.

2 Computational Auditory Scene Analysis

Auditory scene analysis covers the entire hearing process beginning with the reception of sound signal, through several stages of auditory processing, and culminating in the formation and separation of auditory sources [1]. This process is complex and poorly understood. It occurs in two conceptually distinct stages. In the first stage, sound is decomposed into a collection of sensory elements. In the second stage, elements that are likely to have arisen from the same sound source are grouped to form a perceptual whole [20].

Computational auditory scene analysis covers the area of systems based on principles of physiologically inspired grouping heuristics and learned schemata described by Bregman [7]. Numerous approaches to the problem of source separation have been developed. Early work on audio separation was done by Stockham, who used homomorfic signal processing to separate Caruso s voice from background noise and accompaniment in recordings made in 1908 [21]. Stockham was a pioneer, but the real research in the area of sound source separation started in the middle of 1980 s. Schneider and von der Malsburg s neural netwok model [11], Weintraub s statedependent model, Brown and Cooke s data-driven model [2,3], Ellis prediction-driven model [4] implemented on the basis of the blackboard architecture of the IPUS system, Nakatami and Okuno s multi-agnet system models for auditory scene analysis [12], or another approach based on oscillatory correlation [18]. This neural network for auditory pattern segmentation consists of laterally coupled two-dimensional neural oscillators with a global inhibitor. The application of neural oscillators was suggested by Schneider and von der Malsburg [11]. They described an idea of using neural oscillations for expressing segmentation, wherby a set of features froms the same segment if their corresponding oscillators oscillate in synchrony, and oscillator groups representing different segments desynchronize from each other. [18].

Sound source identification cannot be observed as a process which is totally separated from the process of auditory scene analysis. The remarkably successful method of using short- time prediction to infer masked information presented by Ellis [4] appears as a essential in modern computational auditory scene analysis systems. Due to this fact an early recognition and identification of sound source can highly improve the rate of successfulness of such systems and also reduce the computing time. This approach is also supported by the results of the research in psychoacoustics (see Fig.1).

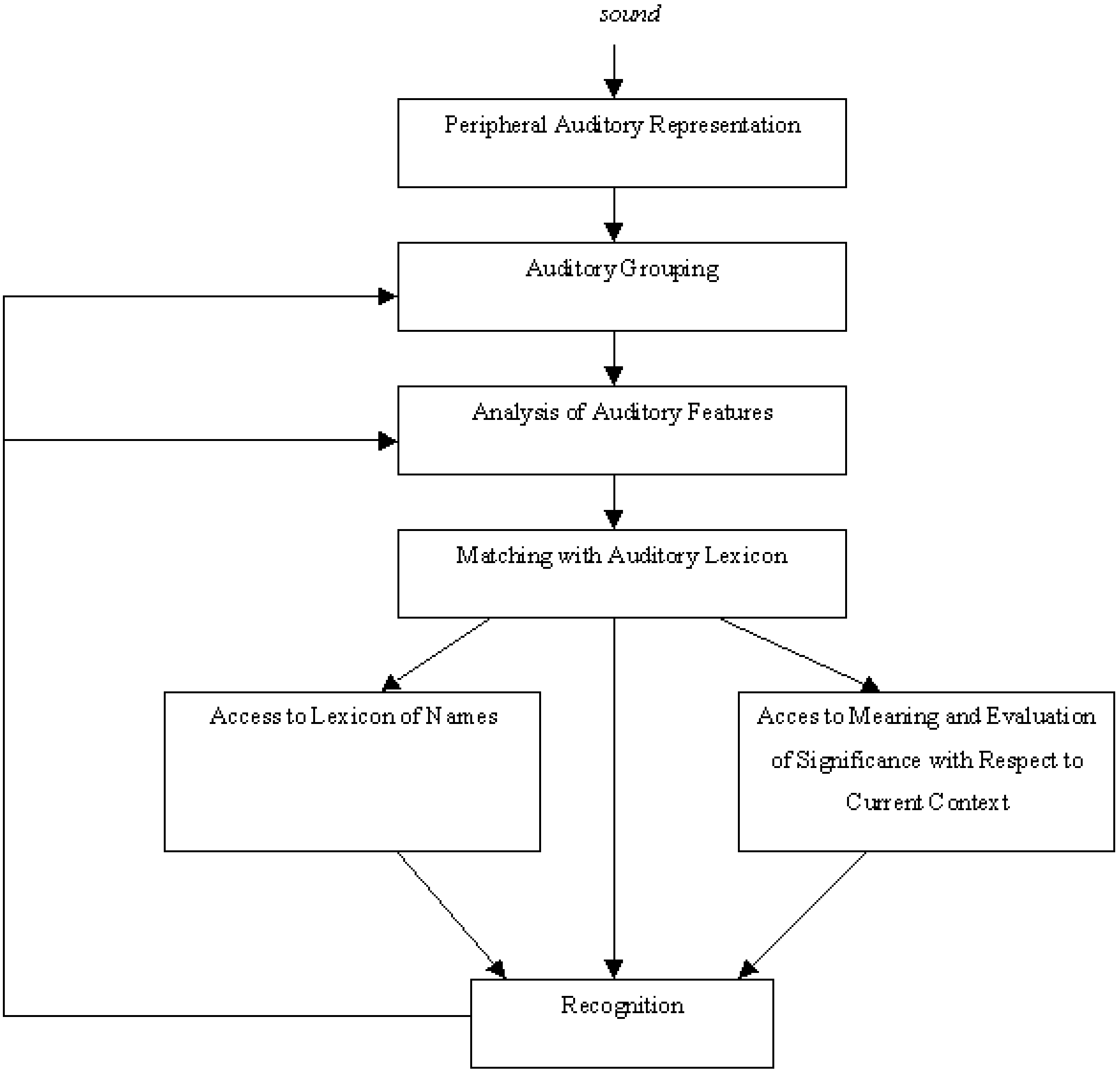
sound
Peripheral Auditory Representation
Auditory Grouping
Analysis of Auditory Features
Matching with Auditory Lexicon
Access to Lexicon of Names
Acces to Meaning and Evaluation of Significance with Respect to Current Context
Recognition

Figure 1.1: Block diagram of the stages of processing involved in recognition and identification.

3 Combination of Auditory Periphery Model and Neural Net

As we said above, the process of early classification of the sound source can improve whole the sound scene analysis process, because knowledge of sound source characteristic can improve the short-time prediction. This fact can be regarded as a motivation for doing presented research.

This study is related to the problem of music instrument classification, but not to the sound source classification in a complex environment. Sound which inputs the system is monaural.

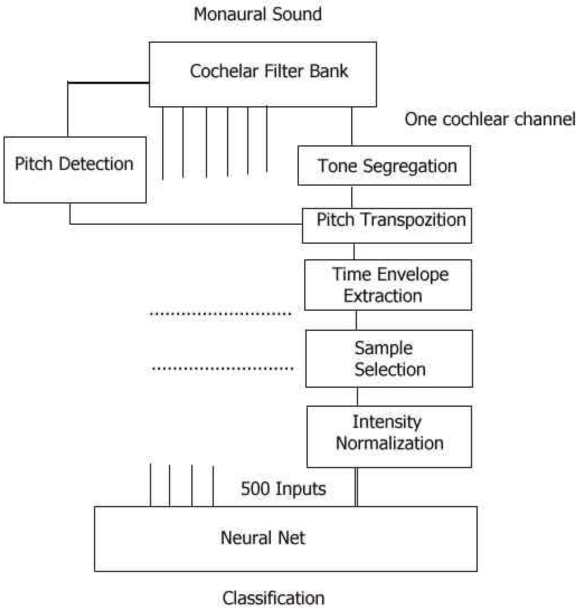

Fig.2.: Block diagram of the presented method combining the model of the auditory periphery and neural net.

Digitized sound inputs the front-end component consisting of the model of outer/midle ear transduction, cochlea filter bank, and the model of inner hair cell transduction. Then, the onsets are detected to recognize begins and ends of the tones. For pitch extraction, we use the algorithm based on summary autocorrelation functions. Then the sounds in all cochlear channels are transposed in the ratio given by the original tone fundamental frequency and normalized fundamental frequency (440 Hz). Time domain envelope for each cochlear channel (time-frequency envelope) is extracted and cut to obtain 500 samples. For details, see chapter 4.

Neural net has 50 x 500 neurons in input layer (2D layer). The number of the neurons in input layer corresponds both to the count of the cochlear channels and to the number of samples used for time-envelope describing multiplied by 10 because each neuron in input layer has 10 inputs.

Backpropagation based supervised feedforward artificial neural network was applied to perform classification. Both the neural network learning algorithm and classification process are based on recognition of the shape of time envelopes of cochlear channels.

4 Signal Pre-processing and an Extraction of Feature Vectors

Gamma-tone filter bank

Auditory periphery system could be regarded as a bank of bandpass filters with center frequencies spanning the audible range. In accordance with a research in the ability of the auditory system to separate the components in a complex sound, there are two fundamental properties of auditory filters - variation with center frequency and variation with level. At moderate sound levels the auditory filter is roughly symmetric on a linear frequency scale. At high sound levels the low-frequency side of the filter becomes less steep than the high frequency side. The shape of the auditory filter appears to depend mainly on the level at the input to the filter [9]. In our research, Patterson-Holdsworh   s simple gamma-tone filter bank consisting of filters with ERB

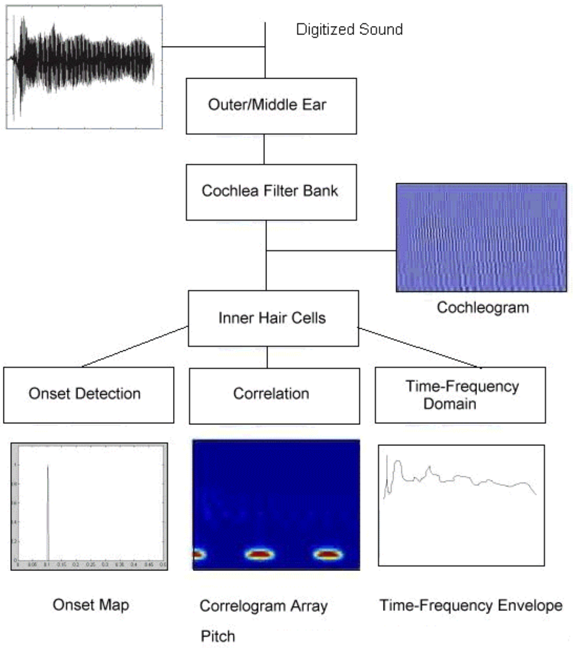

Fig.3.: Block diagram of the features extraction through model of auditory periphery

depending upon center frequency has been used. Full details of this model are given in [13].

Patterson-Holdsworth s cochlear model is based on an array of independent bandpass filters. The filters are tonotopically organized from high frequencies at the base of the cochlea to low frequencies at the apex. The bandwidth of each cochlear filter depends on center frequency. Our implementation of this filterbank follows an efficient algorithm presented in [16]. The fourth-order gammatone filter impulse response is implemented as a cascade of four second-order filters.

There is a short summarization of the most important properties of this bank of filters: fixed number of bandpass filters, bandwidth increasing with a center frequency, logarithmic spacing of the filter frequencies, no dependency on the input sound level, no abrupt high-frequency shift.

A Tone Extraction Algorithm

This algorithm separating the begin an the end of each tone in the monaural sound is based on onset detection. Onset Map is a physiologically inspired feature. Neurons that respond with a brief burst of activity at the onset of a tonal stimulus are found throughout the auditory nuclei [10]. Mechanism that accounts for this behavior could be described by two equations:

$$p(t) = p(t-1)c + E\,r(t) \quad I\,r(t-1) \qquad (1)$$

$$\begin{aligned} s(t) &= p(t),\ p(t) > \text{Treshold} \qquad (2) \\ &= 0 \text{ otherwise} \end{aligned}$$

Equation (1) describes the computation of membrane potencial p(t) at time t. E and I are the strengths of the excitatory and inhibitory inputs, r(t) is the hair cell response and c is a constant that determines a rate at which p(t) decays to its resting level. Brown and Cooke [3] recommend following parameters E=1.0, I=6.0, c=0.86, T=0.

A Pitch Extraction Algorithm

Applied pitch extraction algorithm is based on summary autocorrelation function [8]. To follow physiological approach, we have used the four stage model consisting of the cochlea filter bank, model of inner hair cells, auto-correlation block and the summary autocorrelation block. In the first stage, peripheral auditory processing is simulated by a bank of gammatone filters. This filter bank is followed by a bank of inner hair cell models that provide the input to an autocorrelation stage. These inner hair cell models are used as low-pass filters. Their function lies in reduction of heights of the peaks in the autocorrelation function in high-frequency cochlear channels. Then, all the autocorrelation functions are summed together. The pitch is estimated as a frequency corresponding to the highest peak in the summary autocorrelation function.

A Pitch Transposition Algorithm

Applied pitch transposition algorithm is based on well-known 12 semi-tone scale. Each tone is transposed to the base frequency, but only these music instrument are considered as possible sources, which are able to generated a sound with the original pitch.

Cochlear Channel Time Envelope Extraction

Time envelopes of the signals of all cochlear channels are extracted.

Signal Section Cutting

Neural net has 50 x 500 neurons in input layer, so the signal must be cut. The sampling period about 1 ms causes the time-frequency length corresponding to the time interval about 5 seconds.

Intensity Normalization

Music sound could be produced with the different intensity reflecting the emphasis of the playing tone. In We assume that this intensity has no influence to the shape of tone (shape of the time-frequency envelope). Both the neural network learning algorithm and classification process are based on recognition of the shape of time envelopes of cochlear channels.

6 Experiments & Results

The set of samples contains both percussive and sustained musical instrument tones. There were sounds of approximately 30s 90s duration.

There were four guitar sounds, four distortion guitar sounds, three violin sounds, two bass sounds, four alt sax sounds, three oboe sounds, three flute sounds, three pan flute sounds, three grand piano sounds (monophonic), and two clarinet sounds. These sounds represented musical instrument classes (e.g. piano class, clarinet class, guitar class, etc.). Some of these samples contain only tones with low fundamental frequencies, or high fundamental ones, The other appeared as good representations of whole fundamental frequencies range of the selected musical instrument.

Several training sets were selected from this set of samples. Till now, more than 200 experiments were performed. The average rate of successfulness of the described approach was about 60 percent.

If we train the neural net not to recognize each of the music instruments, but to recognize a group o them, the rate of successfulness was pretty higher (about 80 percent). We follow a well-known hierarchy of musical instrument sounds. At the highest level, instrument tones are classified as either percussive or sustained. Sustained sounds are further classified as blown or bowed, and the blown tones could be classified as brass and woodwind.

7 Conclusion

We first discussed the problem of sound source identification in the context of the sound scene analysis. Sound source identification cannot be observed as a process which is totally separated from the process of auditory scene analysis, because early recognition and identification of sound source can highly improve the rate of successfulness of computational auditory scene analysis systems and also reduce the computing time.

Then the method combining model of auditory periphery with the backpropagation based supervised feedforward artificial neural network was presented. Three layer 2D network has been applied to this pattern recognition problem. The number of neurons in input layer corresponds to the count of cochlear channels and samples describing time-domain envelope multiplied by 10.

This chapter presents a novel mechanism to sound source identification combining neural nets with biologically inspired auditory models. Future work will be focused on the neural net development to exclude necessity detect the tone onsets accurately. Another approach involves the application of the artificial neural oscillators (spiking neurons).

Acknowledgements

This research work has been supported by the research program Transdisciplinary Biomedical Engineering Research M MT:210000012 of the Czech Technical University in Prague (sponsored by the Ministry of Education, Youth and Sports of the Czech Republic).

References

[1] Bregman, A.: Auditory Scene Analysis: The Perceptual Organization of Sound. Camridge, MA, The MIT Press.

[2] Brown, G.J., Cooke M.: Computational Auditory Scene Analysis. Computer Speech and Language, 8, 297-336.

[3] Brown, G.J.: Computational Auditory Scene Analysis, A representational approach, Ph.D. Thesis CS-92-22, CS dept., University of Sheffiled, 1992

[4] Ellis, D.: Prediction Driven Computational Auditory Scene Analysis, Ph.D. thesis, MIT 1996

[5] Janku, L.: Several Approaches to Computational Auditory Scene Analysis, unpublished proposal for Ph.D. thesis, CTU in Prague, 2000

[6] Janku, L.: Sound Source Separation through Models of Auditory Processes and Fuzzy-Rule System, in: Proceedings of MOSIS 2001, TU Ostrava, 2001

[7] Martin, K. D.: Toward Automatic Sound Source Recognition: Identifying Musical Instruments, NATO Computational Hearing Advanced Study Institute, 1998

[8] Meddis, R., O Mard, L.: Psychophysically Faithfull Methods for Extracting Pitch, in Rosenthal, Okuno: Computational Auditory Scene Analysis, Lawrence Erlabaum Associates, Inc., 1998

[9] Moore, B.C.J.: Hearing, Academic Press, 1998.

[10] O Mard, L., Meddis, R. A computational Model of Non-Linear Auditory Frequency Selectivity.

[11] von der Malsburg,, Schneider, W.: A neural cocktail/party processor. Biological Cybernetics, 54, 1986

[12] Nakatani, T., Okuno, H. G.: Multiagent Based Binaural Sound Stream Segregation, in Rosenthal, Okuno: Computational Auditory Scene Analysis, Lawrence Erlabaum Associates, Inc., 1998

[13] Patterson, R.D., Robinsosn, K., Holdsworth, J. et al., Complex Sounds and Auditory Images, in Cazans, Demany, Horner (eds.): Auditory Physiology and Perception, Pergamon, Oxford 1992

[14] Rabiner, L.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of IEEE, 1989

[15] Robbinson, K., Patterson, R.D.: The Duration Required to Identify the Instrument, the Octave, the Pitch Chroma of a Musical Note, Music Perception, Vol.13, 1995

[16] Slaney, M. An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank. Apple Computer Technicl Report #3.

[17] Janku, L.: Sound Source Separation through Models of Auditory Processes, unpublished research report, FEL ¨VUT 2001.

[18] Wang, D.L.: Stream Segregation based on Oscillatory Correlation in Rosenthal, Okuno: Computational Auditory Scene Analysis, Lawrence Erlabaum Associates, Inc., 199

[19] Brown, J.C. Computer identification of musical instruments using pattern recognition with cepstral coefficients as features J. Acoust. Soc. Am. 105, 1933-1941.

[20] Mellinger, D.K., Mont/Reynaud, B.M, Scene Analysis, in Auditory Computation, 1996, Springer Verlag

[21] Smaragdis, P.J.: Information Theoretic Approaches to Source Separation, MIT, 1997