CLASSIFICATION OF PIANO SOUNDS USING TIME-FREQUENCY SIGNAL ANALYSIS

Christoph Delfs and Friedrich Jondral

Nachrichtensysteme, Universität Karlsruhe
D-76128 Karlsruhe, Germany

ABSTRACT

A topical task is the classification of burst-like signals, e.g. in signal detection. Piano sounds are used here as an example. Different time-frequency methods including wavelet processing are used alternatively for feature extraction. A classifier checks whether the generated features are sufficient to identify the correct piano. Results of the real data signal processing are presented and discussed.

1 INTRODUCTION

This paper compares different time-frequency-analysis techniques as feature generators for the classification of piano sounds. Fig. 1 shows the flow chart of the signal processing associated with this task. The resulting sequences $\{x(l)\}$ and $\{x_a(l)\}$ are subject to a feature extraction by means of the short-time Fourier transform, the dyadic orthogonal wavelet transform, the wavelet packet transform and the windowed Wigner-Ville distribution.

Each extraction technique provides a time-frequency representation $\{F(n,m)\}$ of the input sequence and an index set G. The sequence $\{F(n,m)\}$ can be visualized before and after the feature selection is performed. The selected features $\{v_{n,i}(l)\}$ are either saved in a data base during adaptation or classified using the data base.

Section 2 explains the signal preprocessing unit. Section 3 presents the feature extraction techniques used. The following section 4 gives details of the feature selection. Section 5 provides information about the adaptation and classifier. Finally, simulations results are displayed and discussed in section 6.

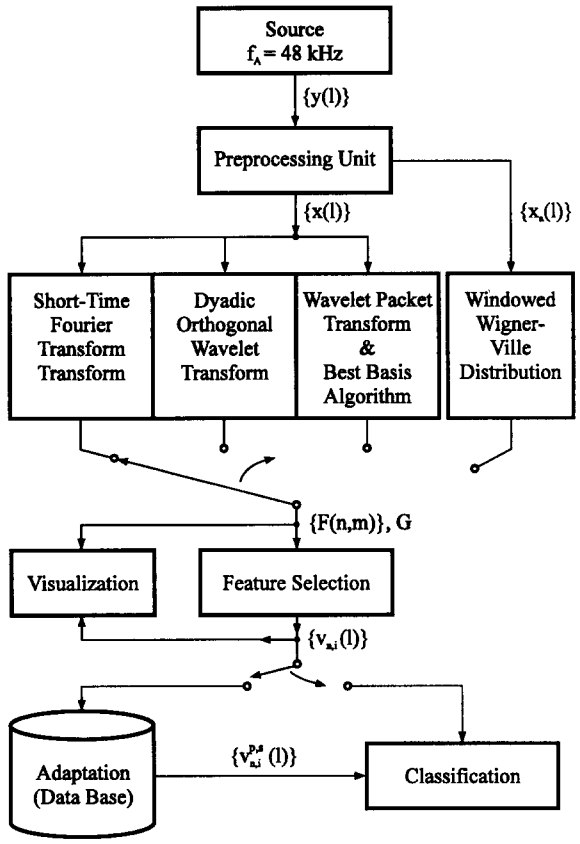

Figure 1: Flow chart of the signal processing for classification

2 PREPROCESSING UNIT

Fig. 2 shows the signal flow inside the preprocessing unit.

The piano sound is sampled at a rate of $f_a = 48kHz$. The maximum value of a sample is here normalized to 1. The preprocessing unit detects the start and the stop sample of a piano sound by measuring the average power within a window of length W starting at index M:

$$P_M^W = 10\log_{10}\left\{\frac{1}{W}\sum_{l=M}^{M+W-1}|y(l)|^2\right\} \quad .$$

Two different window lengths are used for determining the start and stop samples of a piano sound. In order to detect the start of the piano sound as precisely as possible, the length of the window is set to $W_{start} = 30$. The window length for determining the stop sample is chosen to be $W_{stop} = 7200$.

The start sample M_{start} is the first sample fulfilling:
$P^{30}_{M_{start}} \geq P_{start} = -22.5dB$.

The stop sample $M_{stop} > M_{start}$ is the first sample which obeys:
$P^{7200}_{M_{stop}} \leq P_{stop} = -50dB$.

The resulting sequence $\{y^{\#}(l)\}_{l=0}^{M_{stop}-M_{start}-1}$ corresponds to a duration of 4-5s in practice.

Removing low frequency noise is performed by a bandpass filter BP which, in addition, allows sampling rate reduction. The decimated sequence is $\{x(l)\}$, the corresponding analytical sequence is $\{x_a(l)\}$.

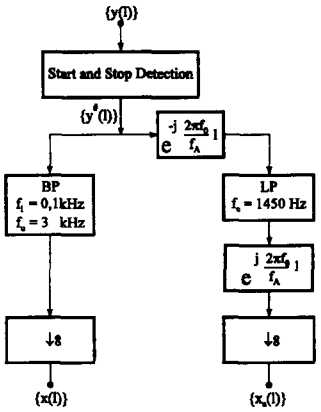

Figure 2: Flow chart of the preprocessing unit ($f_0 = 1.55kHz$)

3 FEATURE EXTRACTION

The techniques used for feature extraction are described for the discrete-time case only. In the following, details of the extraction methods are explained. They map the input sequence $\{x(l)\}$ on the sequence $\{F(n,m)\}$. The set G is the set of indices n in $\{F(n,m)\}$ to be used for feature selection.

3.1 The Short-Time Fourier transform (STFT)

In the short-time Fourier transform, the sequence $\{x(l)\}$ is multiplied with a sliding window sequence $\{w(l)\}$ and subjected to a DFT calculation:

$$F(n,m) = \sum_l x(l)w(l-Dm)e^{-j\frac{2\pi}{N}nl} \quad ,$$

$$n = 0,1,2,\ldots N-1 \quad , \quad m \in \mathcal{Z},$$
$$G = \{0,1,2,\ldots N-1\} \quad .$$

The windowing process can be interpreted as filtering process with the time-reversed sequence $\{h_0(l)\} = \{w(-l)\}$ with length L_0 and subsequent decimation by D. This approach leads to efficient implementations by means of a FFT filter bank [CR83].

3.2 The Dyadic Orthogonal Wavelet Transform (DOWT)

In multiresolution theory the sequence $\{x(l)\}$ is interpreted as a projection onto a set of orthogonal basis functions. $\{x(l)\}$ can be decomposed further using two orthogonal projectors, i.e. the filters $\{h(l)\}_{l=0}^{L-1}$ (highpass) and $\{g(l)\}_{l=0}^{L-1}$ (lowpass) [Mal89]. With the aid of the intermediate sequence $\{u_n(l)\}$:

$$u_0(l) = x(l) \quad ,$$
$$u_n(l) = \sum_{k=0}^{L-1} g(k)u_{n-1}(2m-k) \quad ,$$
$$n = 1,2,\ldots N-1 \quad , \quad m \in \mathcal{Z}$$

the DOWT can be expressed as follows:

$$F(n,m) = \sum_{l=0}^{L-1} h(l)u_n(2m-l) \quad ,$$
$$F(N-1,m) = u_{N-1}(m) \quad ,$$

$$n = 0,1,2,\ldots N-2 \quad , \quad m \in \mathcal{Z},$$
$$G = \{0,1,2,\ldots N-1\} \quad .$$

The transform is realized by a filter bank as shown in fig. 3.

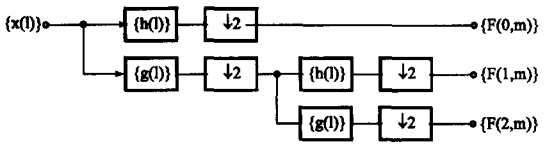

Figure 3: Realization of the dyadic orthogonal wavelet transform by a multirate filter bank with decomposition depth $N-1=2$.

3.3 The Dyadic Wavelet Packet Transform (DWPT)

The wavelet packet transform extends the filtering scheme, as presented in fig. 3, to a filter tree as in fig. 4 and allows multiple complete signal representations.

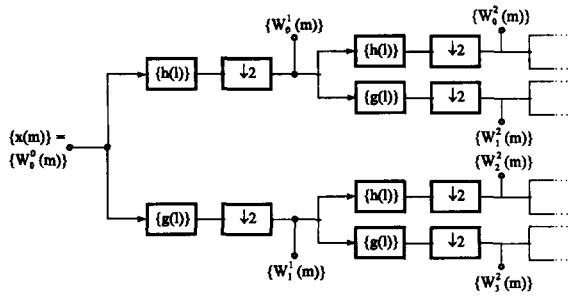

Figure 4: Filter bank according to the dyadic wavelet packet transform

The results of this filter bank can be described in an iterative manner using the double indexed sequences $\{W_l^k(j)\}$:

$$W_0^0(m) = x(m),$$

$$W_{2l}^{k+1}(m) = \sum_{l=0}^{L-1} h(l)W_l^k(2m-l),$$

$$W_{2l+1}^{k+1}(m) = \sum_{l=0}^{L-1} g(l)W_l^k(2m-l),$$

$$k = 0,1,2\ldots K-1 \quad , \quad m \in \mathcal{Z},$$
$$l = 0,1,2\ldots 2^k-1 \quad .$$

By using $n = 2^k + l$ the sequences $\{W_l^k(m)\}$ can be mapped onto the sequence $\{F(n,m)\}$:

$$F(2^k + l, m) = W_l^k(m).$$

The set G contains the indices n of the sequence $\{F(n,m)\}$, whose coefficients describe the input sequence completely and with the smallest entropy (best basis algorithm [CW92]).

3.4 The Windowed Wigner-Ville Distribution (WWVD)

The windowed Wigner-Ville distribution of the analytical sequence $\{x_a(l)\}$ is defined as follows:

$$WWVD(n,m) = 2 \sum_{l=-P}^{P} h_1(l) x_a(m+l) x_a^*(m-l) e^{-j\frac{4\pi}{N}nl},$$

using the window sequence $\{h_1(l)\}_{l=-P}^{P}$. As the transform is periodic with period π, in order to reduce the amount of data, a modified version of the windowed Wigner-Ville distribution is used by introducing a decimation D:

$$\begin{aligned} F(n,m) &= 2 \sum_{l=-P}^{P} h_1(l) x_a(Dm+l) x_a^*(Dm-l) e^{-j\frac{2\pi}{N}nl}, \\ n &= 0, 1, \ldots N-1 \quad , \quad m \in \mathcal{Z} \quad , \\ G &= \{0, 1, 2, \ldots N-1\} \quad . \end{aligned}$$

Fast implementations using the FFT of length N are given in [BB87].

4 FEATURE SELECTION

The feature selection segments the sequence $\{F(n,m)\}$ with respect to m and a given segment length $V \in \mathcal{N}$ following the procedure:

$$\begin{aligned} v_{n,i}^{\#}(l) &= \begin{cases} F(n, iV+l) & \text{if } \sum_{l=0}^{V-1} |F(n, iV+l)|^2 > S, \\ & \qquad n \in G \\ 0 & \text{otherwise} \end{cases} \\ l &= 0, 1, 2 \ldots V-1 \quad , \quad i \in \mathcal{Z} \quad . \end{aligned}$$

Those segments $\{v_{n,i}^{\#}(l)\}$, whose energies exceed a threshold S, are chosen. Segments $\{v_{n,i}^{\#}(l)\}$, which are not selected or whose index $n \notin G$, are set to 0. This technique can be interpreted as selecting the greatest values of the time-frequency-representation $\{F(n,m)\}$.

Finally, the segments $\{v_{n,i}^{\#}(l)\}$ are normalized by:

$$v_{n,i}(l) = \frac{|v_{n,i}^{\#}(l)|}{\sqrt{\sum_{n,i,l} |v_{n,i}^{\#}(l)|^2}} \quad .$$

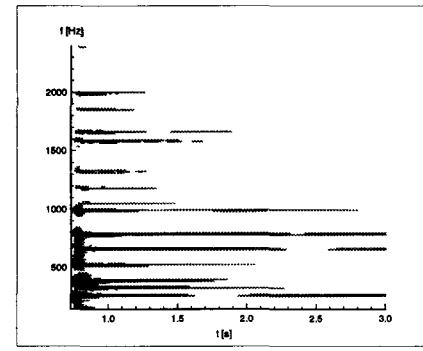

Figure 5: Time-Frequency-Representation of a piano sound analyzed with STFT

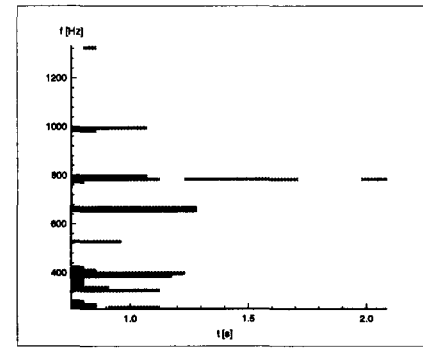

Figure 6: Time-Frequency-Representation of a piano sound analyzed with STFT after feature selection

Fig. 5 shows the time-frequency representation obtained using the short-time Fourier transform before the feature selection takes place. Fig. 6 shows the signal representation of the same signal after feature selection.

In simulations, the threshold S is chosen depending on $\{F(n,m)\}$ so as to obtain a fixed number B of segments $v_{n,i}(l) \neq 0$ for each extraction technique. The feature extraction and selection map the input sequence $\{x(l)\}$ onto a unit length vector with elements $\{v_{n,i}(l)\}$. The segments $\{v_{n,i}(l)\}$ form a pattern of the sound.

5 ADAPTION AND CLASSIFICATION

During the adaptation phase the segments $\{v_{n,i}(l)\}$ of each piano sound are stored in a data base with a class index p and a sound index s respectively. The sound s of piano p is therefore stored as pattern $\{v_{n,i}^{p,s}(l)\}$.

The classifier compares the incoming pattern $\{v_{n,i}(l)\}$ with each pattern $\{v_{n,i}^{p,s}(l)\}_{p,s}$ of the data base by calculating the euclidian distance:

$$d_{p,s} = \sqrt{\sum_{n,i,l} |v_{n,i}(l) - v_{n,i}^{p,s}(l)|^2} \quad .$$

The nearest neighbour classifier is used for relating the incoming sound to a class p^*. Correct classifications are counted.

STFT				
Filter Type	:	Fourier approximation		
Channels N	:	128	256	512
Filter Length L_0	:	256	512	1024
Decimation D	:	128	256	512
DOWT				
Filter Type	:	Daubechies		
Filter Length L	:	10	10	20	20
Depth $N-1$	:	5	10	5	10
DWPT				
Filter Type	:	Daubechies		
Filter Length L	:	10	10	20	20
Depth $N-1$	:	5	10	5	10
WWVD				
Window Type	:	Hamming		
Window Length $2P+1$	:	127		511
FFT Length N	:	128		512
Decimation D	:	32		32

Table 1: Parameter choices for feature extraction

Feature Selection				
Number of Segments B	:	50	100	200
Length of Segments V	:	5	5	5

Table 2: Parameter choices for feature selection

Furthermore, a threshold R is introduced for rejecting insecure classifications. Therefore, if $d_{p\#,s\#} > R$ the pattern $\{v_{n,i}^{p\#,s\#}(l)\}$ is excluded from the classification process.

6 SIMULATION RESULTS

6.1 The Data Set and Parameter Choices for Feature Extraction and Selection

The data set consists of 18 grand pianos, 15 pianos and one keyboard, each with 20 sounds. Each sound is generated by touching the chord $C_4 - E_4 - G_4$ of the instrument until the sound has faded away.

The feature extraction techniques were parametrized as shown in table 1. Each parameter choice from table 1 was simulated with the feature selection parameters as given in table 2. 39 simulations were performed.

6.2 Classification Results and Discussion

The classification results of the simulations are presented in tab. 3 with respect to the number of segments B and threshold $R = 1$.

The perfect classification results of the STFT are independent from the the channel number N and the number of segments B. Tab. 3 shows the result for $N = 512$. The

Number of Segments B	:	50	100	200
STFT	:	100 %	100 %	100 %
DOWT	:	68.8 %	71.8 %	76.5 %
DPWT	:	98.8 %	99.1 %	99.1 %
WWVD	:	97.4 %	98.5 %	99.1 %

Table 3: Classification results

classification results of the DOWT are independent from L and depth $N-1$. The DWPT achieves the best results when parametrized with a long filter set and a long depth. The rates for both techniques refer to $L = 20$ and the depth $N - 1 = 10$ respectively.

The WWVD achieves its best classification results with a long window $2P + 1 = 511$ and a big number of segments $B = 200$. The entry in tab. 3 refers to the long window.

The bad results for the DOWT can be explained by the fact that the underlying filter bank does not fit the signal structure. Furthermore this transform is highly variant with respect to the shifts of the input sequence $\{x(l)\}$. The latter is also valid for the DWPT, but the ability of the DWPT to match the signal structure overcomes this disadvantage.

The performance of the WWVD is on a par with the DWPT and STFT. However, the expenditure of calculation time and computer memory when using the WWVD is greater by an order of magnitude.

In conclusion, wavelet packet techniques show good classification capabilities when applied to piano sounds. Nevertheless, the results show that traditional algorithms as the STFT are still very important. As comparisons between wavelet methods and other time-frequency methods for classification tasks are seldom, one should focus on the question under which circumstances wavelet techniques offer advantages in comparison with other methods.

7 ACKNOWLEDGEMENT

We would like to thank the Musikhaus Schlaile, Karlsruhe, Germany, for allowing the numerous recordings needed for our research.

References

[BB87] B. Boashash and P.J. Black. An efficient real-time implementation of the Wigner-Ville-distribution. *IEEE Trans. Acoust. Speech, Sig. Proc.*, vol.35(no.9):pp. 1518–1521, 1987.

[CR83] Ronald E. Crochiere and Lawrence R. Rabiner. *Multirate Digital Signal Processing.* Prentice-Hall Inc., Englewood Cliffs, 1983.

[CW92] R.R. Coifman and M.V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Trans. Inf. Theory*, vol.38(2):pp. 713–718, 1992.

[LP96] J. Liang and T.W. Parks. A translation-invariant representation algorithm with applications. *IEEE Trans. Signal Processing*, vol.44(2):pp. 225–232, 1996.

[Mal89] S. Mallat. A theory for multiresolution signal decomposition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.2(7):pp. 674–693, 1989.