Identification of Musical Instruments by means of the Hough-Transformation

Christian Röver[1], Frank Klefenz[2], and Claus Weihs[1]

[1] University of Dortmund*
Department of Statistics
44221 Dortmund, Germany
roever@statistik.uni-dortmund.de
[2] Fraunhofer-Institut für Digitale Medientechnologie
Langewiesener Straße 22
98693 Ilmenau, Germany

Abstract. In order to distinguish between the sounds of different musical instruments, certain instrument-specific sound features have to be extracted from the time series representing a given recorded sound.
The Hough Transform is a pattern recognition procedure that is usually applied to detect specific curves or shapes in digital pictures (Shapiro, 1978). Due to some similarity between pattern recognition and statistical curve fitting problems, it may as well be applied to sound data (as a special case of time series data).
The transformation is parameterized to detect sinusoidal curve sections in a digitized sound, the motivation being that certain sounds might be identified by certain oscillation patterns. The returned (transformed) data is the timepoints and amplitudes of detected sinusoids, so the result of the transformation is another '*condensed*' time series.
This specific Hough Transform is then applied to sounds played by different musical instruments. The generated data is investigated for features that are specific for the musical instrument that played the sound. Several classification methods are tried out to distinguish between the instruments and it turns out that RDA (a hybrid method combining LDA and QDA) (Friedman, 1989) performs best. The resulting error rate is better than those achieved by humans (Bruderer, 2003).

1 The Hough-transform

The Hough-transform was originally developed to detect straight lines in (noisy) digital images, and was then later generalized to arbitrary lines or shapes. The procedure has similarities to regression methods, the common problem being to derive line parameters from points lying on that line. The Hough-transform is very robust to outliers, points that are not on the line have little influence on the estimation. It is even possible to fit several different lines independently at the same time (Shapiro (1978)).
Here the Hough-transform is applied to digitized sounds — as a special case

* The work of Christian Röver and Claus Weihs has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475.

of time series data — the question being, whether this yields a useful sound characterization. We will check this by trying to identify musical instruments by the sounds they play.
The motivation to apply the Hough-transform to sounds is that recently a computer chip has been developed that is able to perform the numerically expensive algorithm in real-time.

2 Application to sound data

2.1 Digital sounds

A sound is a periodic oscillation over time, as shown in Fig. 1. In this case the sound frequency (pitch) is 440 Hz, so the oscillation period is $\frac{1}{440} = 0.0023$ seconds, as indicated by the bar in the upper left. A digital sound

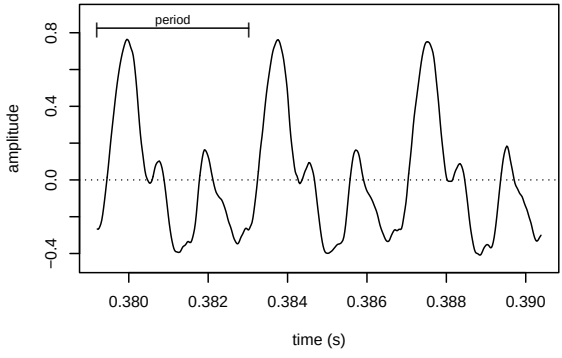

Fig. 1. Periodic oscillation of a sound.

recording is a discrete approximation of the original sound. The recording quality is determined by the resolution of this approximation: CD-tracks are recorded with a *sampling rate* of 44.1 kHz and a *resolution* of 16 bit, so the approximating step function has 44100 steps per second and each step height may take one out of $2^{16} \approx 65000$ values between 1 and -1.
So, statistically spoken, a digital sound is an equidistant time series.

2.2 Motivation: signal edges

The motivation to apply the Hough-transform is that a sound might have a specific oscillation pattern by which it can be identified. In order to catch the pattern features, we concentrate on the so-called *signal edges*, that is, the ascending oscillation sections rising from the time axis as indicated in Fig. 2. We will try to detect these signal edges by fitting appropriate curves

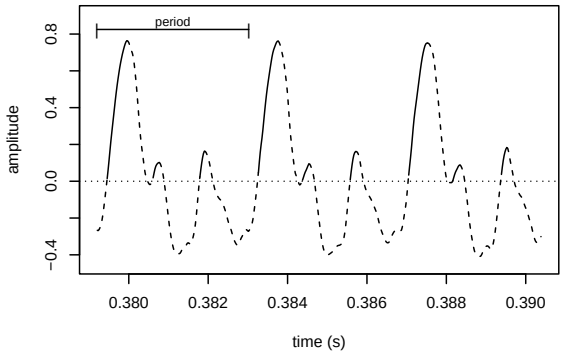

Fig. 2. Signal edges of a sound.

to the sound samples, and then see whether a sound can be classified by the generated sequence of signal edges.

2.3 Parametrization

The Hough-transform was then set to detect sinusoidal signal edges, that is, curves of the form

$$f(t) = A \cdot \sin(2\pi c \cdot t - \phi) \qquad (\phi \leq t \leq \phi + \tfrac{1}{4c})$$

are fit to the sound samples. Variable parameters are amplitude A (≥ 1) and phase difference ϕ (≥ 0); the center frequency c is fixed. The function is sketched in Fig. 3: A stretches the signal edge in the direction of the y-axis and so controls amplitude and slope, while ϕ places the edge along the time axis. Due to the transform procedure, both parameters take only discrete values: amplitudes are divided into 32 bins, and the phase difference resolution is defined by the sound sampling rate (44.1 kHz).
The transformation was then applied to the first 0.7 seconds of each sound, so for longer sound samples not the complete sound is captured in the transformed data.

2.4 Resulting data format

The result of transforming a digitized sound is another time series of amplitudes (A) and phase differences (ϕ); an example is given in Tab. 1: phase differences may be expressed in seconds or sample-indices, and the amplitude can be given in absolute values or bin-numbers. Note that low bin-numbers refer to high amplitudes (steep signal edges) and vice versa.
Fig. 4 shows the transformed data (amplitudes vs. time) for 4 different sounds, the left two played by a piano, and the right ones played on a trumpet. You

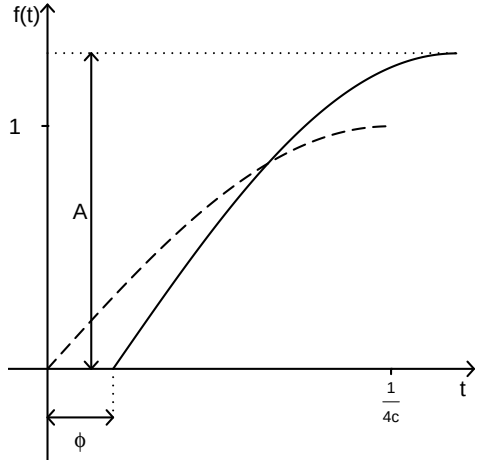

Fig. 3. The fitted signal edge.

Table 1. Data format after transformation.

	phase difference ϕ		amplitude A	
Nr.	sample	seconds	class-nr.	value
⋮	⋮	⋮	⋮	⋮
104	16731	0.3793881	28	1.163636
105	16838	0.3818141	31	1.049180
106	16894	0.3830841	22	1.488372
107	19896	0.3831291	25	1.306122
108	17004	0.3855781	30	1.084746
109	17065	0.3869611	27	1.207547
110	17173	0.3894101	31	1.049180
⋮	⋮	⋮	⋮	⋮

can clearly see similarities within the same instrument and differences across different instruments.
The next problem is now to derive characteristics from these time series that allow for classification of sounds.

3 Classification

3.1 Approaches

In general, two approaches were tried out to summarize the transformed data. The first question was whether the (overall) frequencies of amplitudes may yield a sufficient 'spectrum-like' sound characterization. The second approach

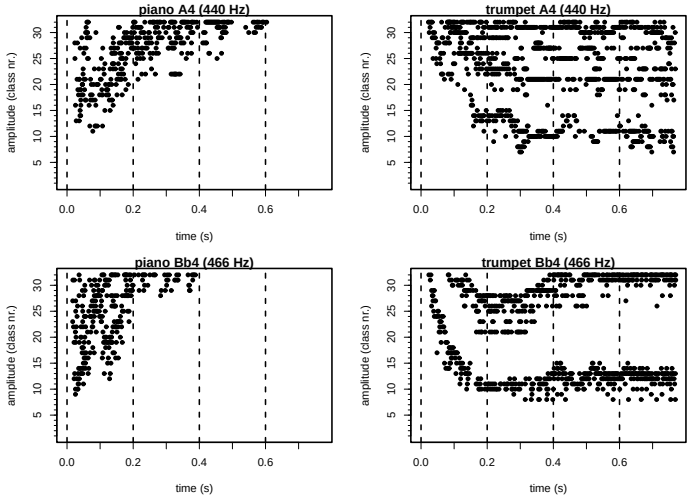

Fig. 4. Different instruments playing at same pitches; left: piano, right: trumpet.

was to derive other characteristics from the transformed time series (not only from amplitudes, but also from frequencies $f_i := \frac{1}{\phi_i - \phi_{i-1}}$).
The first approach uses 33 variables for classification (32 amplitude bins plus pitch), for the second approach 62 potential discriminators were derived from the transformed sound (for examples see results in section 3.5).

3.2 Data set

The investigated data set consisted of 1987 sounds played by different instruments and with pitches of each sound given. There were 62 sound sequences at subsequent pitches; different instruments covered different frequency bands, overall these spanned a range from A0 to C8 (27.5 to 4186 Hz). Sequences played by the same or very similar instruments were grouped together, like piano at different volumes or bassoon and contrabassoon. Finally, the set consisted of 25 instrument classes (Opolko and Wapnick (1987)).

3.3 Methods

The classification methods applied were:

- LDA: Linear Discriminant Analysis
- QDA: Quadratic Discriminant Analysis
- naive Bayes
- RDA: Regularized Discriminant Analysis
- Support Vector Machine
- Classification Tree

- k-NN: k-Nearest-Neighbour

Most methods should be well known except for RDA, which may require some explanation (for Classification Trees see Venables and Ripley (2002), for other methods see Hastie et al. (2001)).

Regularized Discriminant Analysis (RDA) is a hybrid method including LDA and QDA and was proposed by Friedman (1989). Assumptions and procedure are as in QDA, that is, group distributions are conditionally normal and the groups differ by their means and (co-)variances. But instead of using the usual groupwise covariance estimates, the covariance is manipulated using two parameters (λ and γ); first a convex combination is computed:

$$\hat{\Sigma}_k^{\mathrm{RDA}} = \lambda\hat{\Sigma}^{\mathrm{LDA}} + (1-\lambda)\hat{\Sigma}_k^{\mathrm{QDA}} \qquad (0 \le \lambda \le 1)$$

So the covariance estimate is a combination of the pooled ($\hat{\Sigma}^{\mathrm{LDA}}$) and the individual group covariances ($\hat{\Sigma}_k^{\mathrm{QDA}}$); for $\lambda = 1$ it is equal to LDA, and for $\lambda = 0$ it equals QDA. The second parameter γ then allows to shift the estimate towards an identity matrix, but this turned out not to improve error rates, so we restricted ourselves to using λ only and set γ to zero. Thus the covariance estimate simplifies to the above formula.

3.4 Variable selection

Variable selection is necessary for the second approach (characterizing variables), but not appropriate for the first (amplitude frequencies only). Also, classification trees select variables themselves.
For all other methods, variables were then selected applying the same principle (analogous to stepwise regression): Variables were selected step-by-step starting with pitch only and then in each step including the variable that improves the error rate (estimated by cross-validation) most.

3.5 Results

The best classification was achieved using 11 characterizing variables and applying RDA, which resulted in a misclassification rate of 26.1%. Using just the amplitude frequencies, the best error rate was only 66%, using k-Nearest-Neighbour.
The 11 discriminating features leading to the final error rate (26.1%) were:

- pitch
- waiting time for first edge and sound duration
- signal edge rate (per second)
- mean, variance and shape of amplitude distribution

- trend of amplitudes
- mean and variance of frequency distribution
- correlation of amplitude and frequency

The error rates are shown in detail in the confusion matrix (Table 2): each

Table 2. Confusion matrix for RDA using 11 variables (percentages).

%	ba	be	ce	cl	cr	eb	eg	ed	ef	fl	fr	gk	ma	ob	pi	sx	sy	tb	tp	tp	tu	vb	vp	vi	xy	Σ
bassoon	78	0	2	1	0	1	0	0	0	0	0	0	0	1	0	0	2	9	0	0	6	0	0	0	0	22
bells	0	95	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	5
cello	6	0	72	3	0	0	4	3	0	0	0	0	0	1	0	4	0	0	2	0	5	0	0	0	0	28
clarinet	2	0	3	52	0	0	0	8	0	2	1	0	0	7	0	10	0	3	7	0	1	1	0	3	0	48
crotales	0	0	0	0	97	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	3
elecbass	0	0	0	0	0	80	7	0	4	0	0	0	2	0	4	0	0	0	0	0	2	1	0	0	0	20
elecguitar	1	6	8	1	0	12	53	1	2	0	1	0	0	0	4	1	0	1	0	0	1	6	0	1	1	47
elecguitar-*distd.*	0	0	0	1	0	3	0	95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
elecguitar-*fh.*	0	0	0	0	0	12	3	0	73	0	0	0	0	0	3	0	0	0	0	0	0	1	8	0	0	27
flute	1	0	1	1	0	0	0	0	0	69	0	0	0	3	0	2	0	3	3	0	8	2	0	6	0	31
french horn	0	0	0	2	0	0	0	0	0	0	90	0	0	4	0	2	0	0	2	0	0	0	0	0	0	10
glockenspiel	0	0	0	0	11	0	0	0	0	0	0	83	0	0	1	0	0	0	2	0	0	0	0	0	2	17
marimba	0	0	0	0	0	8	0	0	0	0	0	0	61	0	1	0	0	0	0	0	0	0	3	0	26	39
oboe/enghorn	0	0	0	9	0	0	0	0	0	5	2	0	0	70	0	2	0	2	7	1	0	0	0	2	0	30
piano	6	1	1	0	0	7	3	0	1	0	0	2	10	0	55	0	0	0	0	0	0	4	2	0	8	45
saxophone	8	0	10	11	0	0	0	0	0	0	7	0	0	6	0	46	0	3	6	0	0	2	0	0	0	54
synthbass	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	98	0	0	0	0	0	0	0	0	2
trombone	4	0	0	7	0	0	0	1	0	3	0	0	0	3	0	0	0	73	7	0	0	0	0	1	0	27
trumpet	0	0	1	2	0	0	0	0	0	4	5	0	0	8	0	2	0	7	68	0	0	3	0	0	0	32
trumpet-*csto*	0	0	0	3	0	0	0	3	0	0	0	0	0	3	0	0	0	0	0	90	0	0	0	0	0	10
tuba	3	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	95	0	0	0	5
vibraphone	0	2	1	1	0	5	8	0	2	9	1	0	3	0	0	0	0	0	1	1	7	57	0	0	1	43
violin-*pizzicato*	0	0	0	0	0	2	0	0	5	0	0	1	6	0	2	0	0	0	0	0	0	0	84	0	0	16
violin/viola	2	0	2	6	0	0	0	0	2	7	1	0	3	16	0	1	0	7	1	2	1	0	1	48	1	52
xylophone	0	0	0	0	0	0	0	0	1	0	0	5	23	0	2	0	0	0	0	0	0	2	0	0	66	34

total misclassification rate: 26.1%

line corresponds to one instrument and shows how it was classified (in percentages); the main diagonal shows correct classifications, the off-diagonal elements show false classifications. The last column gives the total (instrument-wise) error rate.

For example, you can see that xylophone and marimba get confused with each other, and that there are certain instruments that are classified well (bells), while others are not clearly identified (saxophone).

Closer examination of the transformed data suggested that tuning of Hough-transformation settings might lead to further improvement of classification results. For further details see Röver (2003).

3.6 Comparing the results

The misclassification rate achieved by pure guessing would be $\frac{24}{25} = 96\%$. Error rates achieved by humans or other automatic classification approaches have previously been investigated in other experiments; in roughly comparable problem settings (with regards to number of instruments) rates for humans are quoted at 44%, and for automatic classification these range from 19–7.2% (Bruderer (2003)).

Note that in this study only the first 0.7 seconds of a sound were used,

whereas usually complete sounds are evaluated for recognition. Other approaches often use features like envelope characteristics or fourier frequencies for classification.

4 Conclusions

Application of the Hough-transform to digitized sounds yields a useful sound characterization; the generated data allows to distinguish between sounds played by different instruments. Classification of 25 instruments leads to an error rate of 26.1%.
The misclassification rate so far is better than those achieved by humans, but still worse than for other automatic approaches. Further tuning of transform settings and application to complete sounds (longer than 0.7 seconds) might still improve the procedure.

References

BRUDERER, M.J. (2003): *Automatic recognition of musical instruments*, Masters Thesis, Ecole Polytechnique Fédérale de Lausanne.

FRIEDMAN, J.H. (1989): Regularized Discriminant Analysis. *Journal of the American Statistical Association, 84, No. 405, 165–175.*

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. (2001): *The elements of statistical learning; data mining, inference, and prediction.* Springer-Verlag, New York.

OPOLKO, F., WAPNICK, J. (1987): McGill University Master Samples (*CD-Set*). See `http://www.music.mcgill.ca/resources/mums/html/`

RÖVER, C. (2003): *Musikinstrumentenerkennung mit Hilfe der Hough-Transformation*, Diploma Thesis, Universität Dortmund.

SHAPIRO, S.D. (1978): Feature Space Transforms for Curve Detection. *Pattern Recognition, 10, 129–143.*

VENABLES, W.N., RIPLEY, B.D. (2002): *Modern Applied Statistics with S.* Springer-Verlag, New York.