From Local to Global Analysis of Music Time Series

Claus Weihs and Uwe Ligges

Fachbereich Statistik, Universität Dortmund, D-44221 Dortmund, Germany*

Abstract. Local and more and more global musical structure is analyzed from audio time series by time-series-event analysis with the aim of automatic sheet music production and comparison of singers. Note events are determined and classified based on local spectra, and rules of bar events are identified based on accentuation events related to local energy. In order to compare the performances of different singers global summary measures are defined characterizing the overall performance.

1 Introduction

Music has obviously a global structure. At least classical music is played from scores. Music has, however, also local structures, the most local structure being a period of time with a certain frequency, the most local structure relevant for scores is a note. Obvious more global structures are measures, indicated by bars, and musical motifs, phrases, etc. Such a hierarchy of more and more global structure might be revealed by means of automatic analysis of music time series. Such analysis is demonstrated by means of transcription of vocal time series into sheet music. With the performance of songs, however, more global structure can be identified. Apart from pitch correctness especially timbre gives a basis for comparison of different singers. For such comparison global characteristics of performances are derived.

The basic data was generated by an experiment where 17 singers, amateurs as well as professionals, all voice types, sung the classical song “Tochter Zion” (G.F. Händel), the piano accompaniment played back via headphones (Weihs et al., 2001). The transcription of these performances to sheet music was carried out by the analysis of the corresponding time series followed by classification using minimal background information about the piece of music and the singer in order to be able to automatically transcribe unknown music as well (Weihs and Ligges, 2003a).

The analysis is embedded in a more general concept of combination of time series and event analysis (cp. Figure 1; Morik, 2000). Events are derived from time series, event rules are derived from events, and time series models might be directly derived from time series or from event rules. The adequacy of this general

* The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, “Reduction of complexity in multivariate data structures”) is gratefully acknowledged.

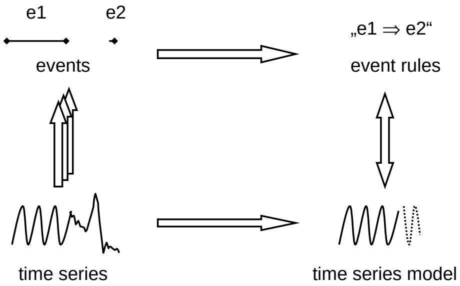

Fig. 1. Time-Series-Event Diagram

scheme for the indicated analysis of music time series is discussed. Steps from local to global analysis of music time series concerning automatic production of sheet music are:

1. Pitch estimation in local blocks of the time series.
2. Identification of tone change events and rests on the basis of pitch estimation, after smoothing off vibrato.
3. Classification of notes (**note events**) corresponding to different tones on the basis of constant tempo (**static quantization**).
4. Identification of local tempo, re-classification of notes (**dynamic quantization**).
5. Combination of notes to measures (**bar events**) by identification of meter via identification of high relative energy.
6. **Key identification** by comparing the identified notes with notes expected in keys.
7. Identification of **rhythm** by comparison with rhythm patterns.
8. Combination of notes to **motifs** by identification of repeated similar series of notes.
9. Combination of bars to **phrases** (e.g. 2 bars).
10. ...

For comparison of the performances of different singers characterizations of the performance of the song as a whole are defined, again based on the spectra of local blocks (Weihs and Ligges, 2003b).

Section 2 introduces the data the example analysis is based upon. Section 3 introduces blocking, the basic data preparation for finding local structure in music time series. Section 4 discusses automatic transcription into sheet music, and section 5 global comparison of singers. Section 6 gives a conclusion.

2 Data

The sheet music of “Tochter Zion” (G.F. Händel) can be found in Figure 2. Note the “ABA” structure of the song. This song was sung by 17 singers and

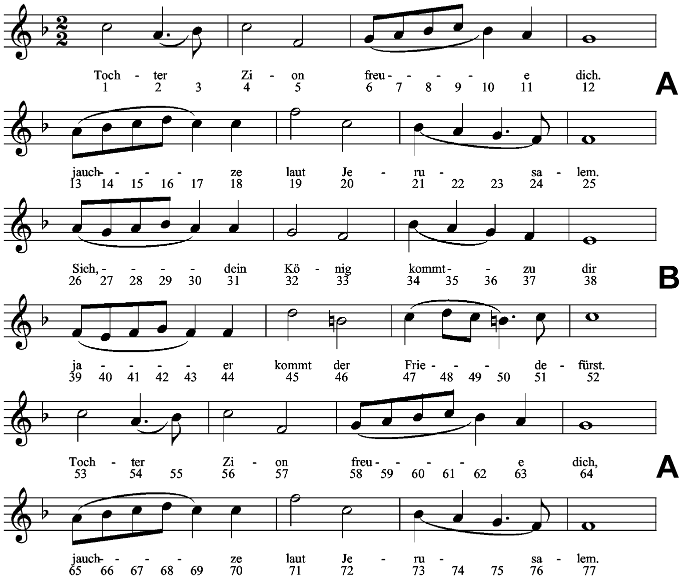
Toch - ter Zi - on freu - - - - e dich.
1 2 3 4 5 6 7 8 9 10 11 12
A
jauch- - - - ze laut Je - ru - sa - lem.
13 14 15 16 17 18 19 20 21 22 23 24 25
Sieh, - - - - dein Kö - nig kommt - - zu dir
26 27 28 29 30 31 32 33 34 35 36 37 38
B
ja - - - - er kommt der Frie - - - de - fürst.
39 40 41 42 43 44 45 46 47 48 49 50 51 52
Toch - ter Zi - on freu - - - - e dich,
53 54 55 56 57 58 59 60 61 62 63 64
A
jauch- - - - ze laut Je - ru - sa - lem.
65 66 67 68 69 70 71 72 73 74 75 76 77

Fig. 2. Sheet Music of "Tochter Zion" (G.F. Händel)

recorded in CD-quality (44100 Hz, 16 bit), but down sampled to 11025 Hz before use. Depending on the task, the corresponding time series, so-called wave, was transformed to spectra, e.g. for pitch estimation, and to energy, e.g. for tempo or meter analysis, both locally in blocks of 512 observations (see next section). For global comparisons of performances spectral characterizations of whole performances are derived from local spectra of 2048 observations. Typical waves and corresponding periodograms, here for the syllable "Zi" (c'' with 523.25 Hz), of an amateur and a professional singer look as in Figure 3, and Figure 4.

Energy is generated from the wave observations w_i by means of the formula:

$$\text{energy} = 20 \cdot \log_{10} \sum_{i=1}^{n} |w_i|, \tag{1}$$

with block size $n = 512$. Local energy is analyzed for the accompaniment (see Figure 5).

3 Local Analysis and Local Structure

Let us continue with some general arguments about local structure corresponding to local events in time series. In order to be able to identify such events, one has to analyze the time series in a granularity that allows for identification of

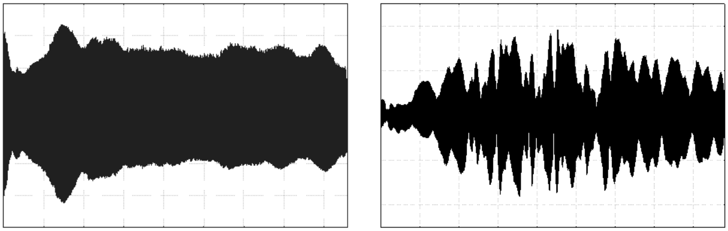

Fig. 3. Waves for syllable "Zi" (c'' with 523.25 Hz), amateur and professional

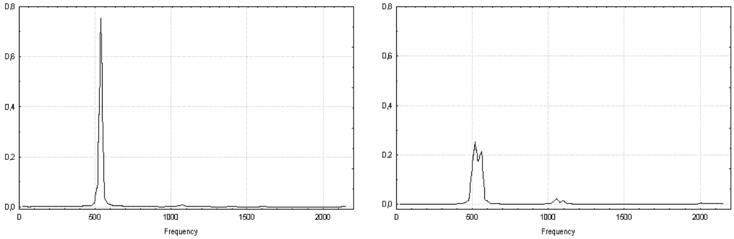

Fig. 4. Periodograms for syllable "Zi" (c'' with 523.25 Hz), amateur and professional

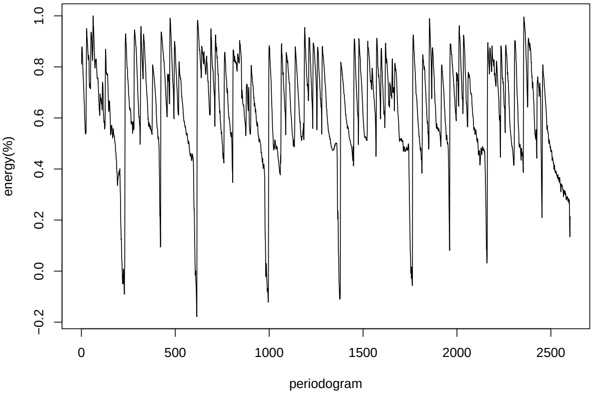

Fig. 5. Local energy of accompaniment

relevant events. Thus, the first task in the analysis is the identification of the size of blocks of observations in the time series to be analyzed. On the one hand, the smaller the blocks, the more exact the time period of an event can be identified. On the other hand, the smaller the number of observations in such a block, the more uncertain is the information on the event in the block. Moreover, it should be clear that event information has to be somewhat redundant to be able to be sufficiently certain about an event. For time series event analysis this can be interpreted as the requirement for ‘enough’ blocks ‘supporting’ the event. Obviously, however, more blocks will lead to smaller blocks and to more uncertain information. Overall, block size is a very important topic to be decided upon in the beginning of time-series-event analysis.

In our application, the most basic events are related to notes, which, again, are related to frequencies of signals. Frequencies are best derived from spectral densities. Spectral densities, however, are only observed at Fourier frequencies. The distance of Fourier frequencies is determined by the analyzed block size. E.g., in the case of a sampling rate of 11025 observations per second and a block size of 512 observations, the distance of Fourier frequencies is $11025/512 = 21.5$ Hz. If frequency estimates would be restricted to Fourier frequencies, then this distance would determine the precision of estimates. This would lead to unacceptably large time series blocks. If neighboring Fourier frequencies are too distant especially for identifying low tones, e.g. of a bass singer, it appears to be necessary to estimate the frequency of the realized tone between the observed frequencies. Our idea is to estimate the maximum of a quadratic model fitted to the Fourier frequency with maximum mass and its left and right neighbor. This leads to pitch estimated by interpolated peaking Fourier frequencies:

$$\text{pitch} := h + ((s - h)/2)\sqrt{ds/dh}, \tag{2}$$

where h = peaking Fourier frequency, s = peaking neighbor, dh, and ds are the corresponding density values. The quality of this method is tested with Midi tones corresponding to the tones sung by human voices. This resulted in a very acceptable maximum error lower than 2 Hz. Moreover, we use half overlapping blocks in our analysis. This leads to 12 blocks corresponding to an eighth for our application if constant tempo is assumed. This was assumed to be enough information for note identification for eighths which are the shortest note appearing in the analyzed song. Based on these arguments, a block size of 512 observations is used in the further analysis.

4 Transcription

Transcription of waves to sheet music can be divided into at least 5 steps:

- Separation of a single voice from other sound,
- segmentation of the sound of the selected voice into segments corresponding to notes, silence or noise,
- quantization, i.e. the derivation of relative lengths of notes,
- meter detection in order to separate notes by bars,
- key determination and
- final transcription into sheet music.

In our project, separation was already carried out by recording, i.e. the singing voice and the piano accompaniment were separated to different channels. Hyvärinen et al. (2001) propose *ICA* for polyphonic sound separation. See von Ameln (2001) for a music example. Segmentation was carried out by pitch estimation followed by classification to a corresponding note.

In our project, segmentation into notes is based on the pitch estimation described in the previous section. The segmentation procedure is described in detail in the next subsection. For alternatives in the literature cp., e.g., Cano et al. (1999) describing a Hidden Markov Model, and Dixon (1996) proposing a method using direct pitch estimation.

In quantization the relative lengths of notes (eighth notes, quarter notes, etc.) are derived from estimated absolute lengths. For this, global or local tempo is derived from accompaniment. In our project the sound is separated into eighths first, since we can assume that an eighth is the shortest note. For an alternative see Cemgil et al. (2000).

Meter identification is carried out by comparing the pattern of accentuation, i.e. the peaking energy distances, to standards corresponding to 4/4 and 3/4 (only, at the moment). Compare also Klapuri (2003). Key detection is postponed to the future, and final transcription into music notation is carried out by an Interface (Preusser et al., 2002) from R (R Development Core Team, 2004) to LilyPond (Nienhuys et al., 2002). This final step also comprises the combination of eighths of equal pitch to longer notes.

4.1 Segmentation Procedure

The task of segmentation is to identify so-called note events from the music time series corresponding to one voice. The procedure can be divided into the following steps:

- Passing through the vocal time series by sections of given size n ($n = 512$ appeared to be appropriate for a wave file sampled with 11kHz).
- Pitch estimation for each section by estimation of the spectral density by means of a periodogram, and the interpolation described above of frequencies of highest periodogram peaks.
- Note classification using estimated fundamental frequencies and the corresponding overtones, given the frequency of diapason a', which might have been estimated.
- Smoothing of classified notes because of vibrato. In our project a doubled *running median* with window width 9 is used. For an alternative vibrato analysis see, e.g., Rossignol et al. (1999).
- Segmentation, iff a change in the smoothed list of notes occurs.

4.2 Transcription by Example

Transcription is now demonstrated by means of an example: the last part A of the ABA scheme of “Tochter Zion” sung by a professional soprano singer. The pitch of each 512-section of the vocal time series is estimated on the basis of the estimated frequency of diapason a' of accompaniment equal to 443.5 Hz, and each corresponding note is classified.

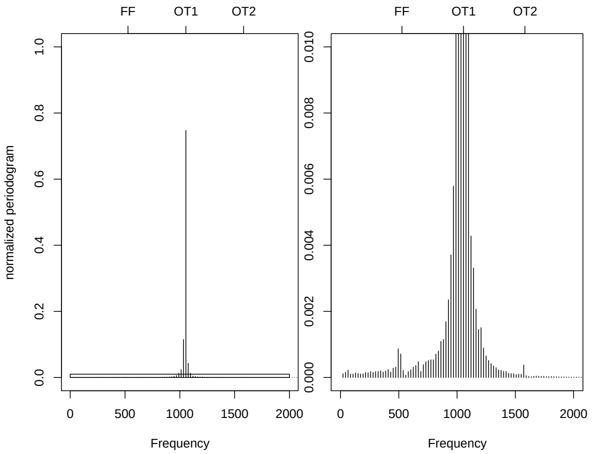

Fig. 6. c'' estimation one octave to high: standard periodogram, and zoomed

In Table 1 the raw classified sections of the first measure are given, where 0 corresponds to diapason a', other integers represent the distance of halftones from a', and silence and quiet noise is represented by `NA`. The singer has an intensive vibrato: classification switches rapidly between 2 (b'), 3 (c''), and 4 ($c\#''$) in the first 2 rows (changes marked by *). Smoothing does not smooth off the intensive vibrato completely (see Table 2), the second half of the note is classified one halftone flat. And moreover, the first sections are classified as c''' instead of c'' since only the first overtone is appearing in the spectrum (see Figure 6).

Table 1. Raw classified sections of the first bar

				*		*						*	*					*				*	
NA	NA	-12	NA	2	2	15	15	15	15	15	15	2	3	3	3	3	3	2	2	2	2	4	4
3	2	2	2	2	4	4	3	2	2	2	2	3	3	2	2	2	2	2	2	-30	NA	NA	NA
NA	NA	NA	-27	-14	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	-1	-1
0	0	0	0	-1	-1	-1	0	0	0	0	0	-1	0	1	1	1	1	1	1	1	0	NA	NA

Table 2. Smoothed classified sections of the first bar

15	15	15	15	15	15	15	15	15	15	15	15	3	3	3	3	3	3	3	3	3	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	NA	NA	NA
NA	NA	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	NA	NA

In Figure 7 a first impression of the sheet music is given. The line indicates the classified note events after smoothing without quantization. The indicated

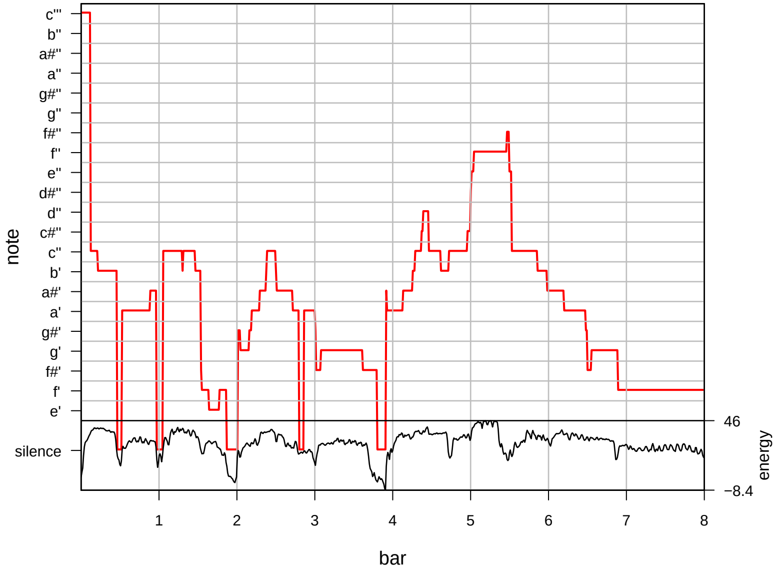

Fig. 7. Progression of note events and energy

bars are just showed for orientation. The progression of corresponding energy is shown as well, low energy reflecting breathing, silence, and strong consonants. Such parts will not be counted as errors in the following. For meter assessment accentuation events are derived from accompaniment. By smoothing constant energy in quarters is produced (cp. Figure 8), assuming global tempo (see below). Then, an *accentuation event* is defined as follows:

$$\text{Accentuation event } A_i, \text{ iff energy(turning point)} > 0.75, \tag{3}$$

i.e. if energy is high peaking, and $D_{i-1} := A_i - A_{i-1}$ indicate the *length of differences between accentuation events* in units of quarters. This can be used to establish rules for the different meters, e.g.

$$\text{no.}(D_{i-1} = 4) > \text{no.}(D_{i-1} = 3) \quad \Rightarrow \quad 4/4 \text{ meter.} \tag{4}$$

In the next step to sheet music, static quantization is carried out, assuming unit = eighth, and no.(eighths) are known. Global tempo is then characterized by length(eighth) = length(series)/no.(eighths). Note events are now related to eighths, and re-classified as the statistical mode of the 12 classified sections of each eighth note in Table 2. Note that each row of Table 2 corresponds to the 24 blocks of one quarter, i.e. two succeeding eighths. Since 4/4 meter was identified, bar events are placed after each 8 eighths, assuming to be known that singing starts in the first measure with the first eighth. The result of this static quantization together with bar placement is shown in the Figure 9 which can

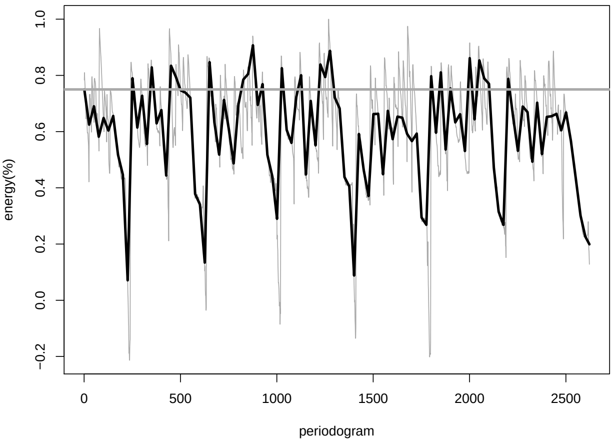

Fig. 8. Smoothed energy indicating meter

"directly" be transcribed into notes. For comparison, true notes are shown as grey horizontal bars.

For the final transcription into sheet music eighths with equal pitch are combined and music symbols for *rests* are used to transcribe silence and low energy noise. The result (cp. Figure 10) is judged by error rates calculated as follows:

$$\text{error rate} := \frac{\text{no.(erroneously classified eighth notes, without counting rests)}}{\text{no.(all eighth notes)} - \text{no.(eighth rests)}}$$

Note that in our example there are 64 eighth notes in 8 bars to be classified. Obviously, there are 9 erroneously classified eighth notes, and 2 eighth rests, thus the error rate is $9/62 = 15\%$. The transcriptions of the other singers' performances gave error rates from 4% to 26%. To assess our outcomes, one should, on the one hand, recognize that we used some a-priori information like knowledge about the shortest note length (eighth note, important for smoothers), the overall number of eighths (for global tempo), and the begin of first sung eighth in the first bar (for bar setting). On the other hand, one should note that the error rate is the sum of various kinds of errors:

- errors of the transcription algorithm, but also
- errors of the singer's performances,
- esp. errors from inaccurate timings of singer,
- errors from static quantization: local tempo of accompaniment was ignored.

This might lead to an analysis of local tempo of the singer or of the accompaniment. At least in our experiment, however, local tempo of singers was very irregular, and local tempo of accompaniment was only followed very roughly by the singer so that utilizing of local tempo did not improve quantization.

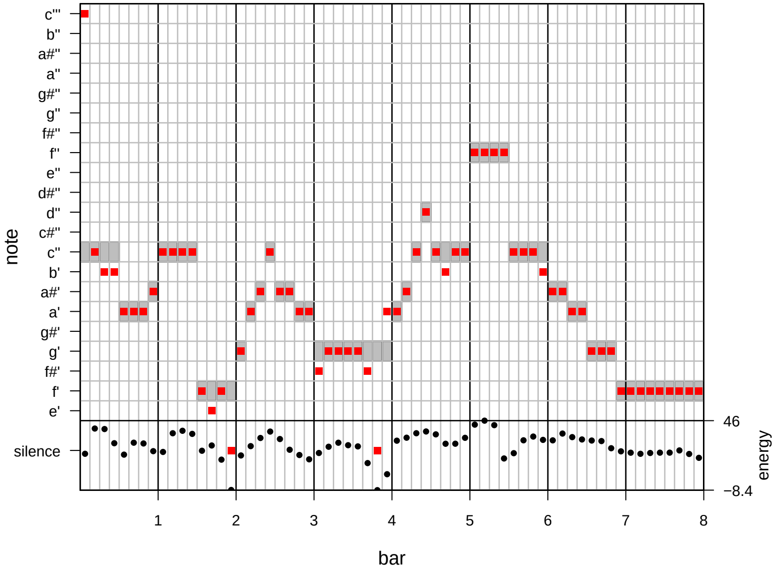

Fig. 9. Progression of notes after quantization

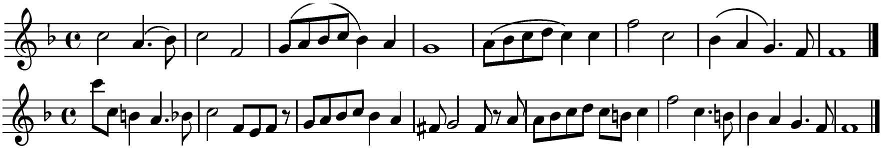

Fig. 10. Original, and estimated sheet music of "Tochter Zion"

In the future we will try to overcome the need of a-priori information, and moreover, we will try to improve our transcription by modelling of vibrato aiming at an improvement of the note classification.

5 Global analysis: Comparison of Singers

Up to now we have analyzed pitch related information in the audio time series locally and more and more globally. This way, we were aiming at automatic transcription into scores. What we have nearly totally ignored until now is the fact that different singers produce different performances of the song. In order to compare such performances it is necessary to define global summary measures characterizing the overall performance. One possible such measure is the size of pitch errors compared to the given score to be reproduced. A possible ratio scale is 'parts of half tone' pht. After assessment of pitch correctness by pht we con-

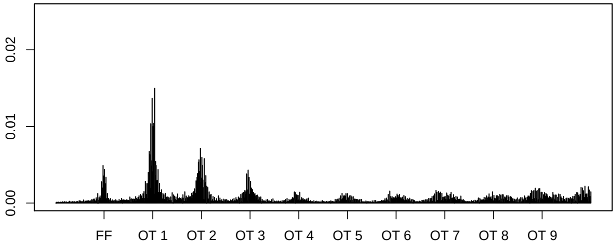

Fig. 11. Pitch independent periodogram

centrate on the information in the spectrum residual to pitch information. This is realized by elimination of pitch from the spectrum. The Fourier Frequencies are linearly rescaled, so that the frequency corresponding to the fundamental is mapped to 1, the frequencies corresponding to the first overtones are mapped to 2, 3, etc. Overlaying and averaging the spectra of different blocks lead to, what we call, the pitch independent spectrum (cp. Figure 11). This time we used half overlaying blocks of $n = 2048$ observations as a basis for periodograms. In pitch independent spectra the size and the shape of the first 13 partials, i.e. the fundamental frequency (FF) and the first 12 overtones (OT1, ..., OT12), are used as characterizations of the residual information in the spectrum after pitch elimination, which is said to be related to the individual timbre of the performance. In order to measure the size of the peaks in the spectrum, the mass (weight) of the peaks of the partials are determined as the sum of the percentage shares of those parts of the corresponding peak in the spectrum which are higher than a pre-specified threshold. The shape of a peak cannot easily be described. Therefore, we only use one simple characteristic of the shape, namely the width of the peak of the partials. The width of a peak is measured by the half tone distance between the smallest and the biggest frequency of the peak with a spectral height above a pre-specified threshold. Mass is measured as a percentage (%), whereas width is measured in parts of halftones (pht). For details on the computation of the measures see Güttner (2001). Based on music theory as a last voice characteristic formant intensity is chosen. This gives the part of mass lying in what one calls the singer's formant lying between 2000 and 3500 Hz individual for the voice types (Soprano, Alto, Tenor, Bass). A large singer's formant characterizes the ability to dominate an orchestra. Overall, every singer is characterized by the above 28 characteristics as a basis for comparison. Figure 12 illustrates the voice print corresponding to the whole song “Tochter Zion” for a particular singer. For masses and widths boxplots are indicating variation over the involved tones.

As an example for comparison let us consider the masses of professional and amateur bass and soprano singers. Figure 13 illustrates that professional singers (Becker and Hasse) have less mass at the fundamental frequency and more mass on higher overtones. The latter is especially true the professional bass singer who has particularly large mass at the singer's formant (cp. Figure 14). For sopranos the singer's formant does not appear to be that pronounced in general.

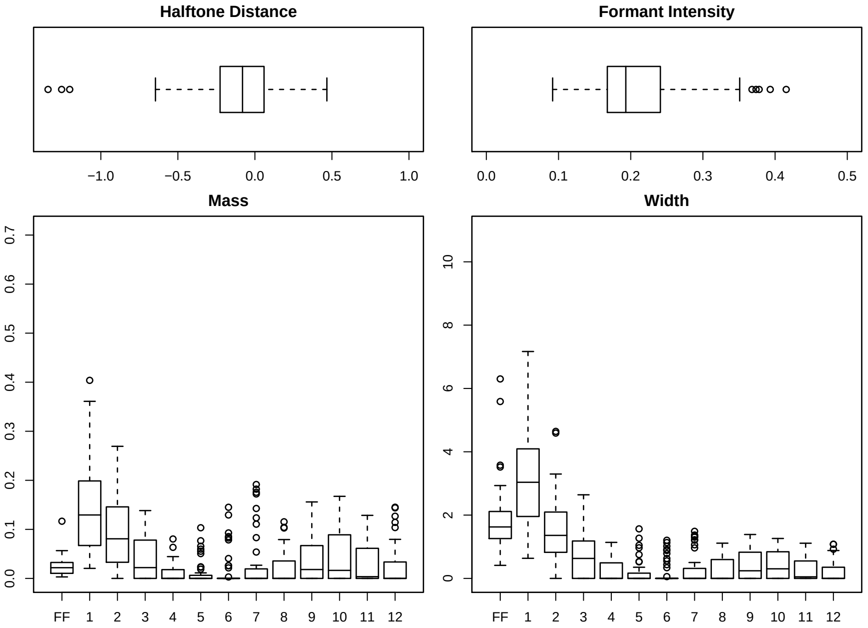

Fig. 12. Voice print: Professional Bass Singer

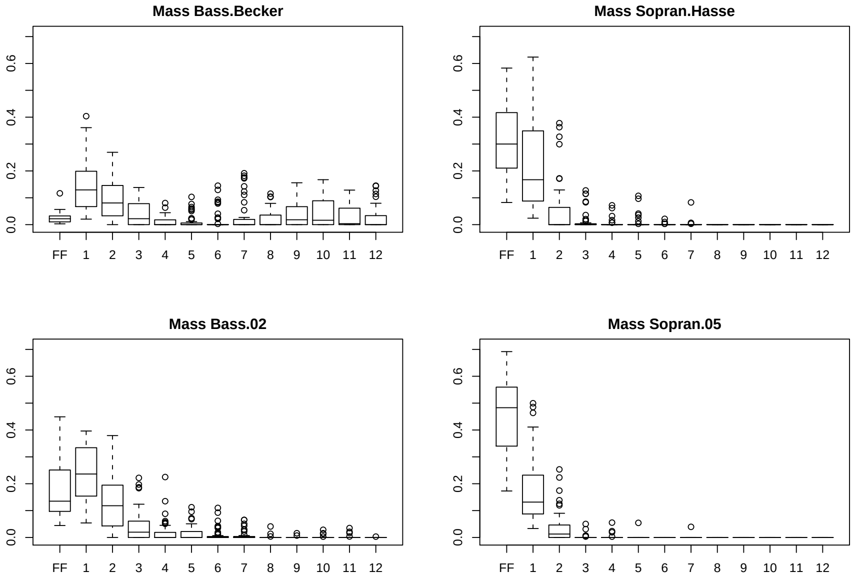

Fig. 13. Voice prints: Comparison of Masses

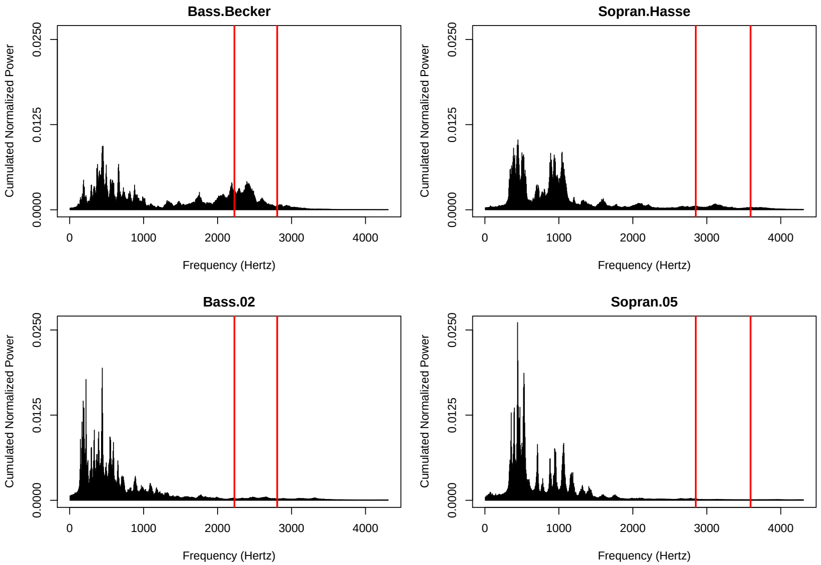

Fig. 14. Comparison of Formants: Formants indicated by vertical lines

6 Conclusion

Our analysis was embedded in a more general concept of combination of time series and event analysis (cp. Morik, 2000). We derived three different kinds of events, for other events event rules have to be derived:

- **Note events** are derived by pitch analysis of time series.
- **Accentuation events** are derived by energy analysis of time series.
- For **bar events** an **event rule** was derived by comparison of accentuation events with meter related accentuation patterns.
- **Rhythm rules** have to be derived by comparison of note lengths with rhythm patterns.
- **Higher structuring rules** for identification of motifs, and phrases of music pieces have to be derived from prescribed event rule types.

Note that we did not derive time series models corresponding to event rules to complete the time-series-event analysis. Concerning local and global analysis we investigated the local and somewhat more global structure of a piece of music:

- Very local pitch and energy estimation was the basis for identification of more global note events.
- Note events were basic for transcription: Identification was based on smoothing of local preliminary notes.
- Local energy estimation of quarters was the basis for accentuation events and meter identification.

– Bar events structure note events: Identification was based on accentuation events.

More global analyses would be key analysis, rhythm analysis, and motif and phrase analysis of music pieces. This was postponed to future research.

For the global comparison of different singers voice prints were developed. In particular, this lead to the identification of pitch independent spectral differences of basses and sopranos and of professionals and amateurs.

References

von Ameln, F.: Blind source separation in der Praxis. Diploma Thesis, Fachbereich Statistik, Universität Dortmund, Germany (2001)

Cano, P., Loscos, A., Bonada, J.: Score-Performance Matching using HMMs. In: Proceedings of the International Computer Music Conference, Beijing, China (1999)

Cemgil, T., Desain, P., Kappen, B.: Rhythm Quantization for Transcription. Computer Music Journal **24** (2000) 60–76

Dixon, S.: Multiphonic Note Identification. Australian Computer Science Communications **17** (1996) 318–323

Güttner, J. : Klassifikation von Gesangsdarbietungen. Diploma Thesis, Fachbereich Statistik, Universität Dortmund, Germany (2001)

Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley and Sons, New York (2001)

Klapuri, A.: Automatic Transcription of Music. In: Proceedings of the Stockholm Music Acoustics Conference, SMAC03 (2003)

Morik, K.: The Representation Race – Preprocessing for Handling Time Phenomena. In López de Mántaras, R., Plaza, E., eds.: Proceedings of the European Conference on Machine Learning 2000 (ECML 2000), Lecture Notes in Artificial Intelligence 1810, Berlin, Springer (2000)

Nienhuys, H.W., Nieuwenhuizen, J., et al.: GNU LilyPond – The Music Typesetter. Free Software Foundation. (2002) version 1.6.5

Preusser, A., Ligges, U., Weihs, C.: Ein R Exportfilter für das Notations- und Midi-Programm LilyPond. Arbeitsbericht 35, Fachbereich Statistik, Universität Dortmund, Germany (2002)

R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2004)

Rossignol, S., Depalle, P., Soumagne, J., Rodet, X., Collette, J.L.: Vibrato: Detection, Estimation, Extractiom, Modification. In: Proceedings 99 Digital Audio Effects Workshop (1999)

Weihs, C., Berghoff, S., Hasse-Becker, P., Ligges, U.: Assessment of Purity of Intonation in Singing Presentations by Discriminant Analysis. In Kunert, J., Trenkler, G., eds.: Mathematical Statistics and Biometrical Applications, Lohmar, Josef Eul Verlag (2001) 395–410

Weihs, C., Ligges, U.: Automatic transcription of singing performances. In: Bulletin of the International Statistical institute, 54th Session, Proceedings. Volume LX. (2003a) 507–510

Weihs, C., Ligges, U.: Voice Prints as a Tool for Automatic Classification of Vocal Performance. In: Kopiez, R., Lehmann, A.C., Wolther, I., Wolf, C., eds.: Proceedings of the 5th Triennial ESCOM Conference, Hanover University of Music and Drama, Germany, 8-13 September 2003 (2003b) 332–335