Identification and separation of noises with spectro-temporal patterns

Roland Sottek
HEAD acoustics GmbH, Ebertstr. 30a, 52134 Herzogenrath, Germany, roland.sottek@head-acoustics.de

Abstract

Acoustic signals often contain perceptually detectable noise patterns with spectro-temporal structures, causing sensations like roughness (due to modulated signal components) and tonality. Technical sounds or environmental noises are often composed of several such components. It is assumed that perceptual evaluations of such complex scenarios show larger deviations because test participants concentrate on different components depending on their preference. Therefore, it is desirable to identify and possibly separate these components allowing for an investigation of each individual noise pattern. The goal is to recognize the composition of all components corresponding to their pitch and modulation rate. Such information could be used for further development and improvement of calculation methods for psychoacoustic parameters. This paper presents different approaches based on time-frequency analyses as well as on the hearing model of Sottek evaluating a three-dimensional autocorrelation analysis as a function of time, lag, and frequency band. The extension to the third dimension allows for a better consideration of modulated signals.

Keywords: Blind source separation, time-frequency analyses, hearing model
Identification and separation of noises with spectro-temporal patterns

1 Introduction

Acoustic signals often contain noise components with spectro-temporal structures like modulated sounds. These perceptually detectable noise patterns may cause sensations, for instance roughness and tonality. Technical sounds or environmental noises are often composed of several such components. Under these circumstances, preliminary listening test results show more variations compared to tests with single modulated components, indicating that test subjects might be focusing on different components [1]. Thus, it would be desirable to simulate this behavior by an automatic procedure to identify and evaluate each component separately, and, if possible, even to separate the sound of interest into the different components as time signals for further processing, e.g., for listening. It is often assumed that the overall sound can be represented as a linear combination of unknown components with unknown weights.

There are several methods available for such a blind source separation using different constraints. For example, the Independent Subspace Analysis (ISA) [2] estimates the different components using the spectrogram (magnitude of Short-Time Fourier Transform: STFT) of the mixture assuming statistically independent sources. An extended approach takes the complex spectrogram (magnitude and phase information, CISA) into account [3]. Another common method for blind source separation is the Nonnegative Matrix Factorization (NMF) using a time-frequency analysis like the STFT as a two-dimensional array to be factorized in a product of two nonnegative matrices containing the sources and the weights under the assumption of different constraints like sparseness of the matrices [4].

However, in all the above approaches, details about modulations often remain hidden. Therefore, a new method based on the three-dimensional autocorrelation analysis (dimensions: time, lag, and frequency band) of the hearing model of Sottek [5] is taken for the separation using a Nonnegative Tensor Factorization (NTF). The extension to the third dimension allows for a better consideration of modulated signals.

2 Hearing model of Sottek

Figure 1 displays the basic hearing model structure for the separation of modulated components using a NTF.

2.1 Outer and middle ear filtering

The pre-processing consists of filtering with outer and middle ear transfer functions: the input signal $p(t)$ is processed by a filter corresponding to the ear’s transmission characteristic $a_0$ of Zwicker’s loudness model [6] in series with a high-pass filter of 1st-order (attenuation of 6.5 dB/decade for frequencies below 500 Hz). Thus, the outer and middle ear filter
approximates the inverted equal loudness contour at 100 phon. It considers, for instance, the loudness elevation due to the cavum conchae resonance of the ear near 4 kHz.

![Diagram of hearing model structure](image)

**Figure 1:** Basic hearing model structure including the determination of the autocorrelation function as function of time, lag, and frequency band (see also Figure 2) for the separation of modulated components using a NTF

### 2.2 Auditory filtering bank

An auditory filter bank consisting of overlapping asymmetric filters\(^1\) models the frequency-dependent critical bandwidths and the tuning curves of the frequency-to-place transform of the inner ear, which mediates the firing of the auditory hair cells as the traveling wave from an incoming sound event progresses along the basilar membrane. The inconstant ratio of bandwidth versus frequency of the auditory filter bank conveys a high frequency resolution at low frequencies and a high time resolution at high frequencies, with a very small product of time and frequency resolution at all frequencies, which empowers, for example, human hearing’s recognition of short-duration low-frequency events.

The impulse responses of the auditory filters are chosen as modulated low-pass filters \((F\) is the center frequency of the corresponding critical band)

\[
h_F(t) = 2 \cdot \varepsilon(t) \cdot \frac{1}{(n - 1)!} \cdot \left(\frac{t}{\tau(F)}\right)^{n-1} \cdot \exp\left(-\frac{t}{\tau(F)}\right) \cdot \exp(j2\pi F t) \quad (1)
\]

where \(n\) is the filter order \((n = 5\) is recommended) and \(\tau(F)\) is a time constant (inversely proportional to the bandwidth of the critical band filter).

\(^1\) The degree of overlapping depends on the application: e.g., for roughness calculation an overlap of half a critical bandwidth was used, for loudness calculation even no overlap is required, and for a spectral deconvolution algorithm several hundreds of filters were used [5].
2.3 Formation of envelope

Subsequent rectification accounts for the fact that nerves fire only when the basilar membrane vibrates in a specific direction. The firing rates of the nerve cells are limited to a maximum frequency. This is modeled applying 3rd-order low-pass filters with a cutoff frequency of 120 Hz, and leads to smooth envelope signals. The sampling rate of the envelope signals can be decimated by a factor of 16, yielding a reduced rate of 3 kHz for signals sampled at 48 kHz.

2.4 Autocorrelation function (ACF)

In early publications, Licklider assumed that human pitch perception is based on both spectral and temporal cues [7]. According to Licklider, the neuronal processing in human hearing applies a running autocorrelation analysis of the critical band signals. Under this assumption, psychoacoustic tonality phenomena like difference-tone perception or the missing fundamental phenomenon can be explained. This work inspired the idea to use the sliding autocorrelation function as a processing block in the hearing model for the calculation of roughness and fluctuation strength [5] and later for other psychoacoustic quantities like tonality [8].

\[
\hat{\phi}_{ss}(t, \tau, i) = \int_{-\infty}^{\infty} s_i(\tilde{t}) \cdot s_i(\tilde{t} - \tau) \cdot \tilde{w}_i(\tilde{t} - \tau) d\tilde{t}
\]

(2)

where \( \tau \) is the lag, and \( s_i \) is the band-pass signal of the \( i^{th} \) critical band. The weighting function \( \tilde{w}_i(t) \) determines how the past signal values contribute to the actual value of the autocorrelation function. Defining \( w_i(t) = \tilde{w}_i(-t) \), equation (2) can be expressed advantageously as a convolution product; \( w_i(t) \) is taken as an impulse response of a 3rd-order low-pass filter:

\[
\hat{\phi}_{ss}(t, \tau, i) = [s_i(t) \cdot s_i(t - \tau)] \ast w_i(t)
\]

(3)

Similarly, the envelope can be calculated as the magnitude of the analytic input signal using the Hilbert transform.
The calculation of the ACF according to equation (3) is highly efficient compared to the integral formulation in equation (2), but still time consuming. Therefore, the sliding ACF is calculated block-wise using the DFT to shorten computing time. In this case, the weighting of the past signal values is a little different. First, there is a low-pass effect due to averaging over the block length. Second, additional 3rd-order low-pass filters are applied to the ACF time curve for each frequency band and lag of interest. The equivalent rectangular duration of the 3rd-order low-pass filters amounts to 5 ms. For the evaluation of envelope fluctuations a DFT length of 512 is chosen (corresponding to a duration of approximately 170.7 ms at a sampling rate of 3 kHz). The calculation is performed for temporally-overlapping blocks: An overlap of about 75% leads to approximately 24 ACF calculations per second.

2.5 Consideration of specific loudness

The nonlinearity between specific loudness and sound pressure was reconsidered in the hearing model according to results of several listening tests [9]. Further improvements for higher levels above approximately 80 dB were achieved by introducing a nonlinearity function according to equation (4)

\[
y(\hat{p}) = \left( \frac{\hat{p}}{\hat{p}_0} \right)^{v_0} \prod_{i=1}^{N} \left( 1 + \left( \frac{\hat{p}}{\hat{p}_{i}} \right)^{\alpha_i} \right)^{-v_i/v_{i-1}}
\]

with \( v_0 = 1 \) and \( N \) further exponents \( v_i \). The thresholds \( \hat{p}_i \) were set to 15 dB, 25 dB, 35 dB, 45 dB, 55 dB, 65 dB, 75 dB, and 85 dB; all \( \alpha_i \) were set to 1.5. The \( N = 8 \) exponents (\( v_1 = 0.6602, v_2 = 0.0864, v_3 = 0.6384, v_4 = 0.0328, v_5 = 0.4068, v_6 = 0.2082, v_7 = 0.3994, \) and \( v_8 = 0.6434 \)) were achieved by applying a nonlinear-optimization procedure in order to minimize the mean squared error between the results of the loudness matching experiment and the results of the model calculation. The specific loudness is calculated by applying the nonlinearity to the square root of \( ACF(\tau = 0) \), because equation (4) is defined for pressure input signals, whereas \( ACF(\tau = 0) \) corresponds to the energy of the input signal.

The specific loudness in each band is set to zero if a frequency-dependent specific loudness threshold is not exceeded. These threshold values are chosen to be the specific loudness values corresponding to the lower threshold of hearing for pure tones minus a fixed value (for the first attempt 6 dB). This makes it possible to take into account loudness summation for complex sounds at threshold in combination with a lower threshold of total loudness. This simple but effective approach must be studied and perhaps optimized later together with the different weighting of the loudness of tonal and non-tonal components.

Afterwards, the entire ACF is normalized for each lag with the same factor, i.e., the ratio of the specific loudness and \( ACF(\tau = 0) \).

\[ ^3 \text{Recent investigations showed that the existing loudness procedures underestimate the loudness of tonal signals [8].} \]
2.6 Noise reduction

In general, $ACF(\tau = 0)$ corresponds to the overall signal energy, i.e., noise and tonal energy. The tonal energy can be estimated, e.g., as the 99%-percentile value of the ACF in a certain lag range, thus the noise energy can be taken as the difference between $ACF(\tau = 0)$ and the tonal energy [8]. In order to reduce the noise, which has a strong impact to the ACF mainly at very small lags, the ACF is attenuated by applying a symmetric window function around $\tau = 0$ (for $|\tau| \leq 4$ ms). The window function shows a minimum (tonal energy divided by $ACF(\tau = 0)$) at $\tau = 0$ and is increasing trigonometrically to a value of 1 at $|\tau| = 4$ ms.

2.7 Spectral transform and weighting

The ACF is an even function of the lag with a real-valued nonnegative spectrum. The spectrum of the ACF as a function of modulation rate is calculated by means of a DFT. Only half of the spectral data must be considered because of its symmetry: i.e., 256 values. The noise reduction described in section 2.6 may cause small negative values of the spectrum, which are set to zero.

In order to simulate the perception of roughness, the modulation spectrum can be weighted with respect to modulation rate in conjunction with a loudness-dependent weighting factor for each frequency band.

3 Nonnegative matrix and tensor factorization

3.1 Nonnegative matrix factorization

Many scientists have investigated NMF, but it became popular by the work of Lee and Seung [4] for the detection of patterns in two-dimensional data. The main purpose of NMF is to factorize a matrix $V$ (dimension $F \times N$) with nonnegative values into a product of two nonnegative matrices $W$ (dimension $F \times K$) and $H$ (dimension $K \times N$) of low rank (Figure 3) with $K$ components, (each matrix contains one vector per component). Figure 3 shows as an example a matrix $V$ (result of a time-frequency analysis) factorized into a product of two nonnegative matrices $W$ and $H$, containing the frequency pattern and the time pattern, respectively.

![Figure 3: NMF: factorization of a two-dimensional nonnegative matrix $V$ into a product of two nonnegative matrices; $W$: frequency pattern (activation in frequency bands), $H$: time pattern (activation in time) of low rank; representation of a component by two vectors](image)

In general, there is no unique solution and the factorization process leads to an approximation $\hat{V}$ of the initial matrix $V$:
\[ V \approx \hat{V} = W \cdot H \] (5)

The factorization can be achieved by an iterative procedure minimizing a cost function \( D \), e.g., a metric for the distance between \( V \) and \( \hat{V} \). The overall distance \( D \) is the sum of all scalar (element-wise) distances \( d \):

\[
D(V||\hat{V}) = \sum_{f=1}^{F} \sum_{n=1}^{N} d(V(f,n)||\hat{V}(f,n))
\] (6)

A typical measure for \( d \) is the \( \beta \)-divergence [10]-[12]:

\[
d_{\beta}(x||y) = \begin{cases} 
  x \cdot \frac{x^{\beta-1} - y^{\beta-1}}{\beta - 1} - \frac{x^{\beta} - y^{\beta}}{\beta} & \beta \in \mathbb{R} \setminus \{0,1\} \\
  x \cdot \log \left( \frac{x}{y} \right) - x + y & \beta = 1 \\
  \frac{x}{y} - \log \left( \frac{x}{y} \right) - 1 & \beta = 0
\end{cases}
\] (7)

The distance metric for \( d_{\beta} \) is known as Euclidian distance (scaled by 0.5) for \( \beta = 2 \), Kullback-Leibner distance for \( \beta = 1 \), and Itakura-Saito distance for \( \beta = 0 \). In case of the Euclidian distance small values of \( x \) and \( y \) lead to small values of \( d_{\beta=2} \). The Itakura-Saito distance metric is scale invariant \( (d_{\beta=0}(\lambda \cdot x||\lambda \cdot y) = d_{\beta=0}(x||y)) \), thus small and large values of \( x \) and \( y \) have the same relevance for \( d_{\beta=0} \). The Kullback-Leibner distance represents a compromise.

The iterative factorization starts with the initialization of \( W \) and \( H \), e.g., with nonnegative random numbers. Usually, different random initializations are tested, and the initial values showing the lowest value of \( D \) are selected as a starting point \( (i = 0) \) for the iterative procedure [13]. The update rules to minimize \( D \) are given by [11]:

\[
H_i = H_{i-1} \cdot \frac{W^T \cdot (W \cdot (H_{i-1})^\cdot (\beta - 2)) \cdot V}{W^T \cdot (W \cdot (H_{i-1})^\cdot (\beta - 1))}
\] (8)

\[
W_i = W_{i-1} \cdot \frac{((W_{i-1} \cdot H)^\cdot (\beta - 2)) \cdot V \cdot H^T}{((W_{i-1} \cdot H)^\cdot (\beta - 1)) \cdot H^T}
\]

\footnote{The cost function can be extended using additional constraints like sparseness.}
Note that \(^{\wedge}\), \(\ast\), and the fraction line denote element-wise exponentiation, multiplication, and division, respectively. The iteration stops if a certain criterion is fulfilled, e.g., if the difference of \(D\) between two iteration steps is below a given threshold. The factorization provides for each component a vector of \(W\) and a corresponding vector of \(H\) allowing for the generation of a submatrix (a spectrogram) for the component of interest. A disadvantage of NMF is that a-priori the number of relevant components must be known, but there are strategies available for an automatic detection of relevant components [14].

The separated components are often created using the Wiener reconstruction, an approach similar to Wiener filtering [11]. This method uses for all components the same phase information. Time signals can be resynthesized based on inverse transforms, e.g., in the case of a STFT by an ISTFT (Inverse Short-Time Fourier Transform). A Hann window is used for the synthesis in order to reduce artifacts due to the transition between successive blocks [1].

### 3.2 Nonnegative tensor factorization

The NTF is an extension of NMF for the detection of patterns in three-dimensional data. A tensor \(V\) (dimension \(F \times N \times I\)) with nonnegative values is factorized into a product of three nonnegative matrices \(W\) (dimension \(F \times K\)), \(H\) (dimension \(N \times K\)), and \(Q\) (dimension \(I \times K\)) of low rank (Figure 4) with \(K\) components (each matrix contains one vector per component).

**Figure 4: NTF: factorization of a three-dimensional nonnegative matrix \(V\) (tensor) into a product of three sparse nonnegative matrices; \(W\): frequency pattern (activation in frequency bands), \(H\): modulation pattern (activation in modulation rates), \(Q\): time pattern (activation in time)**

Similar to the NMF method, the NTF factorizes the tensor by minimizing a cost function using also the \(\beta\)-divergence as defined in equation (7). Additionally, the multiplicative update rules can be extended to the three-dimensional case. For details on the algorithm, please refer to [15].

Figure 4 shows a scheme factorizing a three-dimensional nonnegative matrix \(V\) (tensor) in three sparse nonnegative matrices. In case of the presented autocorrelation analysis of the hearing model of Sottek (Figure 2), the three dimensions are frequency bands (\(F = 24\) for non-overlapping critical bands or \(F = 47\) if an overlap of half a critical bandwidth is used), modulation rates (\(N = 256\) for a DFT length of 512), and time (\(I \approx 24 \cdot T\), \(T\) is the signal duration in s).

In principle, time signals can also be resynthesized based on a Wiener reconstruction in case of the presented NTF approach. The first results are promising [1].
4 Application example

The number of the separated components for the following example was determined automatically using the algorithm described in [14]. The components were determined by means of minimizing the distance metric $D$ based on the Kullback-Leibner distance $d_{\beta=1}$. Figure 5 shows results obtained by NTF using the presented hearing model for a sound recorded from a hard disk drive. The several modulated components could not be resolved in the same way by NMF [1].

![Figure 5: NTF: factorization of hard disk noise; top left picture shows the activation with respect to modulation rate for the five found components. The other pictures show the spectrogram (upper part) and the modulation spectrum vs. critical bands (lower part) for each component.](image)

Modulated signal shares in different frequency bands that show the same pattern regarding modulation rates are not resolved and are considered as one component. This was tested for a mixture of five modulated components (different frequency bands and various conditions of modulation rates (same/different). It is promising to see that the same phenomenon is reported by some participants of preliminary listening tests. They consider such uniform modulated signals covering more than one frequency band similarly as one sound object, but more elaborate tests are needed to assure this observation [1].

5 Conclusions

NMF and NTF are promising methods for blind source separation of audio signals, whereas NTF has the advantage to consider additional dimensions to frequency and time, e.g., in this paper the modulation rate of a signal component that is important from the perceptual point of view. Further investigations are ongoing to work on the separation of modulated as well as tonal components, especially for the challenge of tones with varying frequency. Additionally, investigations for improving the auralization of the modulation tensor components using the hearing model will be performed.
Acknowledgments
The author wishes to thank his former student Stephan Drewes for his investigations on algorithms for the detection of acoustic patterns [1].

References