Chapter 8
Audio Source Separation

I just wondered how things were put together.
—Claude Elwood Shannon

In order to enhance the (audio) signal of interest in the case of added audio sources, one can aim at their separation. Albeit being very demanding, Audio Source Separation of audio signals has many interesting applications: In Music Information Retrieval (MIR), it allows for polyphonic transcription or recognition of lyrics in singing after decomposing the original recording into voices and/or instruments such as drums or guitars, or vocals, e.g., for 'query by humming' [1]. In ASR, the separation of the target speaker from others, background noises or music [2] may help to improve the accuracy. Given multiple microphone tracks, ICA [3] is usually among the first choices. Traditional ICA, however, limits the number of sources that can be separated to the number of available input channels, which makes basic ICA unsuitable for many audio recognition and retrieval applications where only mono- or stereophonic audio is available. To improve performance of ICA in challenging scenarios, source localisation information can be integrated as a constraint, which is promising for ASR in hands-free human-machine interaction [4]. However, this also requires knowledge about the localisation of the microphones used for recording, which is again not given in typical audio mining and retrieval tasks.

On the other hand, fully blind separation of multiple sources from mono- or stereophonic signals is considered infeasible as of today. To summarise, in most Intelligent Audio Analysis applications, prior knowledge has to be exploited in audio source separation, as will be detailed in the following section. A general framework for such 'informed' source separation has recently been presented in [5]. In the light of Intelligent Audio Analysis, such informed methods are particularly interesting, as they leverage machine intelligence for the highly challenging problem of underdetermined source separation. Among the most promising approaches towards separation of monophonic sources are those centred around NMF [6–11], which will be the focus of this chapter. NMF can also be applied in different places along the Intelligent Audio Analysis processing chain, e.g., for audio feature extraction and

B. W. Schuller, *Intelligent Audio Analysis*, Signals and Communication Technology,
DOI: 10.1007/978-3-642-36806-6_8, © Springer-Verlag Berlin Heidelberg 2013

classification such as in noisy conditions [13–15]. This will be introduced towards the end of this section.

Let us now introduce the theoretical foundations of NMF. For clarity, the following notation will be used: For a matrix $\underline{A}$, the notation $\underline{A}_{i,:}$ denotes the i-th row of $\underline{A}$ (as a row vector), and let us analogously define $\underline{A}_{:,j}$ for the j-th column of $\underline{A}$ (as a column vector). Further, let $\underline{A} \otimes \underline{B}$ denote the elementwise product of matrices $\underline{A}$ and $\underline{B}$. The division of matrices is always to be understood as elementwise.

8.1 Methodology

Let us now discuss in detail how NMF can be used for source separation and NMF-based feature provision. The basic procedure is the extraction of an arbitrary number of sources—the 'components'— from audio by non-negative factorisation of a spectrogram in matrix representation $\underline{V} \in \mathbb{R}_{+}^{M \times N}$ into a spectral basis $\underline{W} \in \mathbb{R}_{+}^{M \times R}$ and activation matrix $\underline{H} \in \mathbb{R}_{+}^{R \times N}$:

$$\underline{V} = \underline{W}\,\underline{H}. \tag{8.1}$$

This yields R component spectrograms $\underline{V}^{(j)}$, $j = 1, \ldots, R$ either by multiplication of each basis vector $\underline{w}^{(j)} := \underline{W}_{:,j}$ with its activations $\underline{h}^{(j)} := \underline{H}_{j,:}$, as in [7], or by a more advanced 'Wiener filter' approach, as described in [6, 10]:

$$\underline{V}^{(j)} = \underline{V} \otimes \frac{\underline{w}^{(j)}\underline{h}^{(j)}}{\underline{W}\,\underline{H}}. \tag{8.2}$$

The spectrograms can be obtained from short-time Fourier transformation (STFT) and subsequent transformation to magnitude, power or Mel-scale spectrograms. Each $\underline{V}^{(j)}$ is then transformed back to the time domain by inverse STFT, using the original phase.

Several NMF algorithms can be used for the factorisation according to (8.1). These minimise a distance function $d(\underline{V}|\underline{W}\,\underline{H})$ by multiplicative updates of the matrices. The starting point can be a random initialisation. $d(\underline{V}|\underline{W}\,\underline{H})$ can be chosen as the β-divergence or one of its special instances, the Itakura-Saito (IS) [10] divergence, Kullback-Leibler (KL) divergence, or squared Euclidean distance (ED) [16]. Further, to support overcomplete decomposition, i.e., choosing R such that $R(M + N) > MN$, sparse NMF variants [7] exist for either of the named distance functions, as well as the sparse Euclidean NMF variant used in [17]. In addition, non-negative matrix deconvolution (NMD) [6, 8] has been proposed as a context-sensitive NMF extension. In NMD, each component is characterised by a sequence of spectra, rather than by an instantaneous observation. Alternatively, sequences of spectral feature vectors can be modelled as 'supervectors' in a sliding window approach to use standard NMF for context-sensitive factorisation [13]. More precisely, the original spectrogram $\underline{V}$ is transformed to a matrix $\underline{V}'$ such that every column of $\underline{V}'$ is the row-wise concatenation of a sequence of short-time spectra (in the form of row vectors):

$$\underline{V}' := \begin{bmatrix} \underline{V}_{:,1} & \underline{V}_{:,2} & \cdots & \underline{V}_{:,N-T+1} \\ \vdots & \vdots & \ldots & \vdots \\ \underline{V}_{:,T} & \underline{V}_{:,T+1} & \cdots & \underline{V}_{:,N} \end{bmatrix}, \tag{8.3}$$

where T is the desired context length. That is, the columns of $\underline{V}'$ correspond to overlapping sequences of spectra in $\underline{V}$. If signal reconstruction in the time domain is desired, the above named spectrogram transformations, including Mel filtering and transformation according to (8.3), can be reversed.

The basic NMF method as explained above is entirely unsupervised. In many practical applications, such as speech or music separation, prior knowledge about the problem structure can be exploited. A simple yet very effective method to integrate a-priori knowledge into NMF-based source separation is to perform supervised or semi-supervised NMF. This means that parts of the first NMF factor are predefined as a set of spectra characteristic for the sources to be separated rather than choosing random initialisations of both factors. This can be useful in audio enhancement, e.g., in a 'cocktail party' situation with several simultaneous speakers [6, 17], or noise versus a speaker of interest [18]. The initialisation spectra may themselves stem from NMF decomposition of training material or can be based on simpler methods such as median filtering or simply random sampling of training spectrograms. This procedure is outlined in Fig. 8.1 as a flowchart. An alternative supervised NMF method, depicted in Fig. 8.2, is to assign components computed by unsupervised NMF to classes such as 'drums' and 'non-drums' by means of a supervisedly trained classifier as in [19]. This allows dealing with observations that cannot be described as a linear combination of pre-defined spectra, but assumes that unsupervised NMF by itself can extract meaningful units, such as notes of different instruments. Given an assignment of NMF components to sources as described above, it is straightforward to synthesise the audio signals of interest by overlaying component spectrograms.

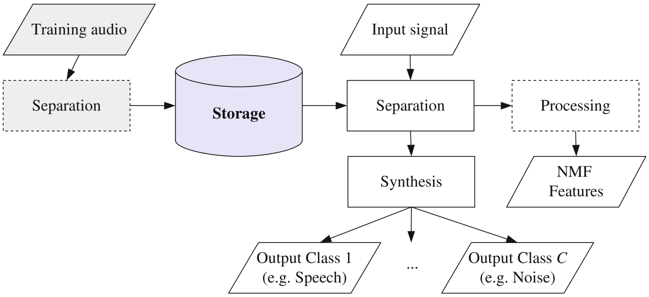

Fig. 8.1 Supervised NMF: A set of spectral components (which can themselves be computed by NMF from training audio) serve as constant basis for NMF; the activations can be exported as features or be used to synthesise audio signals for the sources [12]

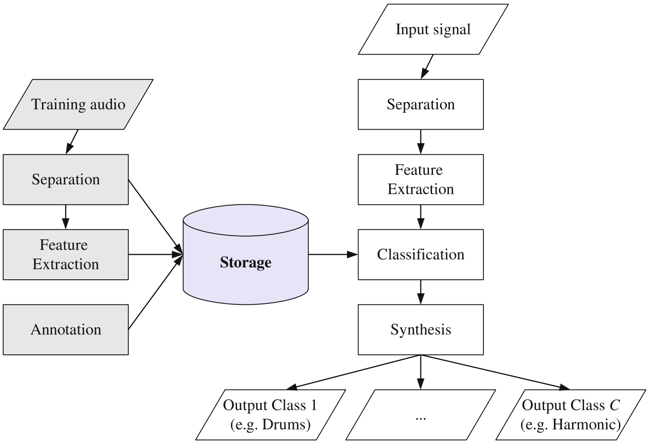

Fig. 8.2 Unsupervised NMF followed by supervised component classification, as in musical instrument separation: A classifier is built from labelled separated components. Steps required to train the classifier are *gray shaded* [12]

Besides using source separation as pre-processing for Intelligent Audio Analysis, the activations computed by NMF can be used directly for classification, as indicated by the flowchart in Fig. 8.1. This approach will be presented in more detail in Sect. 8.3.

8.2 Performance

To get an idea of the separation performance by basic NMF in a challenging task, let us consider the separation of two simultaneously speaking speakers from a monaural signal in the ongoing. Fig. 8.3 visualises the separation quality in terms of source-distortion ratio (SDR) depending on the targeted RTF. SDR, as introduced by [20], can be considered as the most popular objective metric for the evaluation of audio source separation as of today. In the considered scenario of speaker separation, it takes into account the suppression of the interfering speaker but also penalizes the introduction of artifacts due to signal separation, i.e., information loss in the target speech—note that perfect interference reduction can be trivially achieved by outputting a zero signal. These experiments are based on the procedure proposed in [6] and the results correspond to those reported in [12]. NMF is used over NMD based on the finding in [6] of no significant difference in separation quality by either of these two bases. The effect of using different numbers of iterations, DFT window sizes and the NMF cost function is assessed; the importance of these parameters on separation quality and computational complexity has been pointed out in [6, 12]. 12 pairs of male and female speakers—ensuring that the speech spectra do not fully overlap—were selected randomly from the TIMIT database (cf. also Sect. 10.4.3). Per pair, two

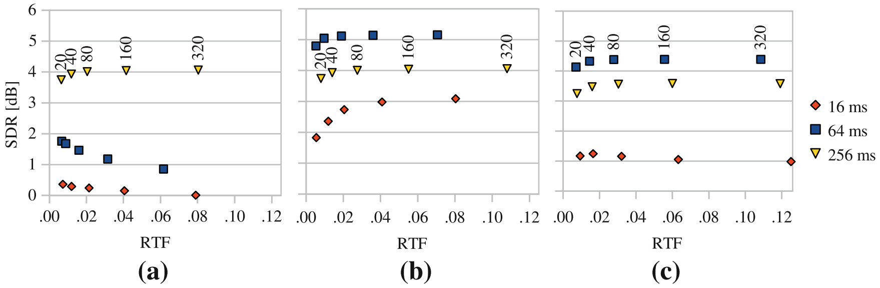

Fig. 8.3 Benchmark results for monaural speaker separation by supervised NMF, in terms of RTF and signal-to-distortion ratio (SDR) [20]. Mixed signals from pairs of male/female speakers (24 speakers total) from the TIMIT database. The open-source openBliSSART toolkit is used, and computation is performed on a consumer grade GPU (NVIDIA GeForce GTX 560). The number of NMF iterations (20–320), the DFT window size (16, 64, 256 ms) and the NMF cost function are adjusted. **a** Euclidean distance. **b** KL divergence. **c** Itakura-saito divergence

randomly selected sentences of roughly equal length were mixed, and a NMF basis $\underline{W}$ was computed from the other sentences spoken by each speaker. To this end, unsupervised NMF (250 iterations) was applied to the concatenated spectrograms of these sentences and only the first factor was kept. Separated signals for both speakers were obtained by supervised NMF with $\underline{W}$, by summing up component spectra corresponding to either speaker, and applying inverse STFT as discussed above. Computations base on a 2.4 GHz desktop PC with 4 GB of RAM, using a consumer grade GPU (NVIDIA GeForce GTX 560) with 336 CUDA cores. The NMF implementation from the open-source toolkit openBliSSART [12] is used. RTFs are computed by the elapsed GPU time over the length of the mixed signals. The number of separation iterations was chosen from {20, 40, 80, 160, 320} due to the quick saturation of the convergence of multiplicative update NMF algorithms in audio source separation [9]. The different DFT window sizes considered are powers of two, ranging from 2^6 to 2^{12}, or 8–256 ms assuming 16 kHz sample rate. From Fig. 8.3, it can be seen that the best average results are obtained by using the KL divergence as cost function. The Euclidean distance allows faster separation at the expense of quality, but here, reasonable results are only achieved for long window sizes (256 ms), which limits the practical applicability in contexts where real-time operation is required. Finally, the IS divergence enables robust separation, but is inferior to KL divergence both in terms of separation quality and RTF. Generally, it can be observed that in case of inadequate modeling of the sources (indicated by overall low SDR), more iterations do not necessarily improve separation quality, despite the fact that they linearly increase computational complexity; in fact, more iterations sometimes degrade quality, e.g., for the Euclidean cost function and 16 or 64 ms window size.

8.3 NMF Activation Features

Let us now move on to describe how NMF can be used directly for audio recognition, instead of performing signal pre-processing by audio source separation. The core idea is to use supervised or semi-supervised NMF (cf. above), and then directly exploit the matrix $\underline{H}$ for classification. In this case, NMF seeks a minimal-error representation of the signal (in terms of the cost-function) with only a set of given spectra. As outlined in Sect. 8.1, the $\underline{H}$ matrix measures the contribution of spectra to the original signal. Thus, by using a matrix $\underline{W}$ that contains spectra of different target classes, the rows of $\underline{H}$ provide information whether the original signal consists of components of these target classes. Furthermore, in this framework, additive noise can be modelled by simply introducing additional NMF components corresponding to noise.

For discrimination of C different audio signal classes $c \in \{1, \ldots, C\}$, the matrix $\underline{W}$ is built by column-wise concatenation:

$$\underline{W} := \underline{W}_1 | \underline{W}_2 | \cdots | \underline{W}_C | \underline{W}_N .$$

where each $\underline{W}_c$ contains 'characteristic' spectra of class c and the optional matrix $\underline{W}_N$ contains noise spectra. Similarly to the source separation application, there are a variety of methods for computing $\underline{W}_c$ and $\underline{W}_N$, such as base learning by NMF as in the supervised speaker separation example above, or simply by randomly sampling training spectrograms.

Based on this, NMF activation features can be derived from $\underline{H}$. In the example shown in Fig. 8.4, an exemplary scheme for static audio classification based on NMF activations is shown that delivered remarkable performance in discrimination of linguistic and non-linguistic vocalisations [15]. In this scheme, it is supposed that base learning by NMF is used. An activation feature vector $\underline{a} \in \mathbb{R}^R$ is calculated such that $\underline{a}_i$ is the Euclidean length of the i-th row of $\underline{H}$. For independence of the length and power of the signal, a_i is normalised such that $|\underline{a}|_1 = 1$. The 'NMF activation features' then are the components of the vector $\underline{a}$. This vector can be passed on to a suited classifier, or the activations per class can be summed up to derive class posteriors. In dynamic classification, e.g., the index of the most likely class per frame can be used as in [14, 21].

Let us now conclude the discussion of audio source separation and feature extraction by NMF by showing an exemplary application to keyword recognition in highly non-stationary noise [21]. This example is based on the CHiME (Computational Hearing in Multisource Environments) challenge task of recognising command words in a reverberated indoor domestic environment with multiple noise sources and interfering speakers [22].

NMD bases are learnt for each of the 51 words in the vocabulary, and an additional NMD noise base is computed from a set of noise samples in the training data. Speech separation is performed in a procedure similar to the speaker separation example above. Additionally, NMF activation features are computed using a base matrix $\underline{W}$ assembled from spectrogram 'patches' in the training data, in a 'sliding window NMF' framework (cf. above) with $T = 20$. As each speech spectrogram patch is

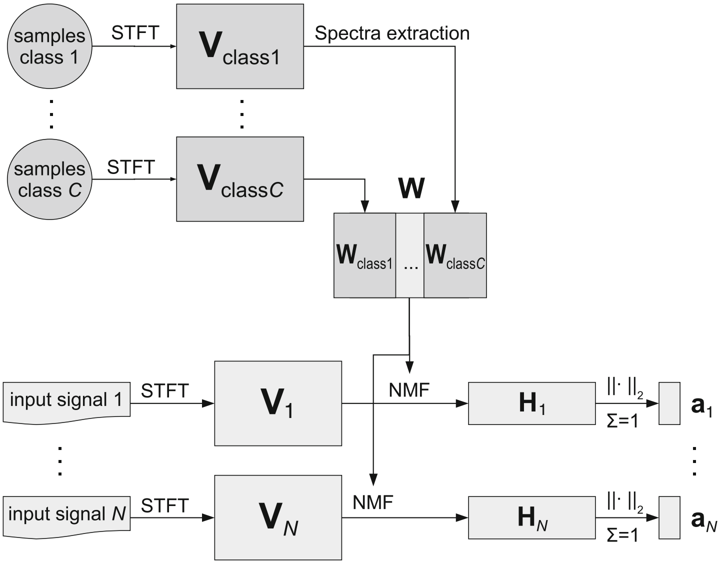

Fig. 8.4 Exemplary block diagram for extraction of NMF activation features for discrimination of C classes in N input signals [15]. Matrices denoted by $\underline{V}$ are spectrograms. The matrix $\underline{W}$ consists of spectra computed from training data for supervised NMF. Activation features are the resulting $\underline{H}$ (activation) matrices. $||\cdot||_2$ indicates that the Euclidean norm of each matrix row is computed, and $\sum = 1$ is a normalisation for the components of each vector $\underline{a}_i$ sum up to 1

associated with word likelihoods, the index of the most likely word per frame can be computed from the frame-wise activations of each spectrogram patch and used as a discrete feature. In this calculation of NMF activation features, $\underline{W}_N$ is pre-defined by training noise samples. Table 8.1 shows the WAs on the 35 keywords by SNR and on average, obtained by a baseline HMM recogniser adapted to noisy speech features, the results achieved by considering NMD speech separation as pre-processing, the results by usage of NMF activation features in HMM decoding, and combination of both. From the results, it is evident that both methods are complementary—the interested reader is referred to [21] for a more in-depth discussion.

Table 8.1 Effect of NMD speech separation and NMF activation features on speech recognition results (WA) reported in [21] on the Computational Hearing in Multisource Environments (CHiME) task [22]

WA [%]	SNR [dB]						Average
	−6	−3	0	3	6	9	
Baseline	54.5	61.1	72.8	81.7	86.8	91.3	74.7
NMD speech separation	75.6	79.2	84.1	87.7	88.3	90.6	84.2
NMF activation features	67.2	75.1	85.0	89.8	92.0	93.4	83.7
Combination	79.1	82.8	88.7	91.2	92.7	93.5	88.0

References

1. Schuller, B., Rigoll, G., Lang, M: Hmm-based music retrieval using stereophonic feature information and framelength adaptation. In: Proceedings 4th IEEE International Conference on Multimedia and Expo, ICME 2003, vol. II, pp. 713–716. Baltimore, MD, July 2003 (IEEE, IEEE)
2. Weninger, F., Feliu, J., Schuller, B.: Supervised and semi-supervised supression of background music in monaural speech recordings. In: Proceedings 37th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2012, pp. 61–64, Kyoto, Japan, March 2012 (IEEE, IEEE)
3. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons Inc., New York (2001)
4. Maas, R., Schwarz, A., Zheng, Y., Reindl, K., Meier, S., Sehr, A., Kellermann, W.: A two-channel acoustic front-end for robust automatic speech recognition in noisy and reverberant environments. In: Proceedings of CHiME, pp. 41–46 (2011)
5. Ozerov, A., Vincent, E., Bimbot, F.: A general flexible framework for the handling of prior information in audio source separation. IEEE Trans. Audio Speech Lang. Process. **20**(4), 1118–1133 (2012)
6. Smaragdis, P.: Convolutive speech bases and their application to supervised speech separation. IEEE Trans. Audio Speech Lang. Process. **15**(1), 1–14 (2007)
7. Virtanen, T.: Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Audio Speech Lang. Process. **15**(3) (2007)
8. Wang, W., Cichocki, A., Chambers, J.A.: A multiplicative algorithm for convolutive non-negative matrix factorization based on squared Euclidean distance. IEEE Trans. Signal Process. **57**(7), 2858–2864 (2009)
9. Schuller, B., Lehmann, A., Weninger, F., Eyben, F., Rigoll, G.: Blind enhancement of the rhythmic and harmonic sections by nmf: Does it help? In: Proceedings International Conference on Acoustics including the 35th German Annual Conference on Acoustics, NAG/DAGA 2009, pp. 361–364, Rotterdam, The Netherlands: Acoustical Society of the Netherlands. DEGA, DEGA (2009)
10. Févotte, C., Bertin, N., Durrieu, J.-L.: Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis. Neural Comput. **21**(3), 793–830 (2009)
11. Duan, Z., Mysore, G.J., Smaragdis, P.: Speech enhancement by online non-negative spectrogram decomposition in non-stationary noise environments. In: Proceedings of Interspeech, Portland, OR, USA (2012)
12. Weninger, F., Schuller, B.: Optimization and parallelization of monaural source separation algorithms in the openblissart toolkit. J. Signal Process. Syst. **69**(3), 267–277 (2012)
13. Gemmeke, J.F., Virtanen, T.: Noise robust exemplar-based connected digit recognition. In: Proceedings of ICASSP, pp. 4546–4549, Dallas, TX, March 2010
14. Schuller, B., Weninger, F., Wöllmer, M., Sun, Y., Rigoll, G.: Non-negative matrix factorization as noise-robust feature extractor for speech recognition. In: Proceedings of 35th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, pp. 4562–4565, Dallas, TX, March 2010 (IEEE, IEEE)
15. Schuller, B., Weninger, F.: Discrimination of speech and non-linguistic vocalizations by non-negative matrix factorization. In: Proceedings of the 35th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, pp. 5054–5057, Dallas, TX, March 2010 (IEEE, IEEE)
16. Lee, D., Seung, H.: Algorithms for non-negative matrix factorization. In: Proceedings of NIPS, pp. 556–562, Vancouver, Canada (2001)
17. Schmidt, M.N., Olsson, R.K.: Single-channel speech separation using sparse non-negative matrix factorization. In: Proceedings of Interspeech, pp. 2–5, Pittsburgh, Pennsylvania (2006)
18. Ozerov, A., Févotte, C., Charbit M.: Factorial scaled hidden markov model for polyphonic audio representation and source separation. In: Proceedings of WASPAA, pp. 121–124, Mohonk, NY, United States (2009)

19. Helén, M., Virtanen, T.: Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In Proceedings of EUSIPCO, Antalya, Turkey (2005)
20. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1462–1469 (2006)
21. Weninger, F., Wöllmer, M., Geiger, J., Schuller, B., Gemmeke, J., Hurmalainen, A., Virtanen, T., Rigoll, G.: Non-negative matrix factorization for highly noise-robust asr: to enhance or to recognize? In: Proceedings 37th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2012, pp. 4681–4684, Kyoto, Japan, March 2012 (IEEE, IEEE)
22. Christensen, H., Barker, J., Ma, N., Green, P.: The CHiME corpus: a resource and a challenge for Computational Hearing in Multisource Environments. In: Proceedings of Interspeech, pp. 1918–1921, Makuhari, Japan (2010)