Separating Musical Audio Signals[1]

Mark D Plumbley, Director of the Centre for Digital Music, Queen Mary University of London

Introduction

As consumers move increasingly to multichannel and surround-sound reproduction of sound, and also perhaps wish to remix their music to suit their own tastes, there will be an increasing need for high quality automatic source separation to recover sound sources from legacy mono or 2-channel stereo recordings. In this Contribution, we will give an overview of some for audio source separation, and some of the remaining research challenges in this area.

Separating mixed sounds

A common problem that arises in musical audio is that of *source separation*, where we want to separate one sound from a mixture of many sounds. This is sometimes known as the *cocktail party problem* (Cherry, 1953): we might imagine we are at a party trying to listen to one conversation, while many other noises and other conversations are going on around us.

This full-scale problem, with multiple sources, delays and background noise, is much too difficult for current techniques. Let us therefore start with something much simpler: two source instantaneously mixed in different ways before being picked up by two microphones (Fig. 1). We express this mathematically in matrix/vector form as $\mathbf{x}$=$\mathbf{As}$ where $\mathbf{s} = [s_1, s_2]^T$ is the pair of sources, $\mathbf{x} = [x_1, x_2]^T$ is the pair of microphones, and $\mathbf{A}$ is the 2×2 matrix with elements a_{ij} representing how much of the jth source is picked up by the ith microphone.

[Fig 1 goes about here]

Our task is then to try to "unmix" this process to find the original sources $\mathbf{s}$, given only the signals $\mathbf{x}$ picked up at the microphones, but without knowing the mixing matrix $\mathbf{A}$. Separating the sources without knowing the mixing process is known as *blind* source separation.

This blind separation problem might seem to be a difficult problem, but we can see how we might try to solve it by turning into a game. Imagine that instead of sound sources s_1 and s_2 we have two dice: Amber (A) and Blue (B). Then we ask an opponent to think of a pair of secret formulae to mix these to give two observations X and Y. For example, they might come up with $X = (1/2)A + 3B$ and $Y = 2A + B$. They then roll the dice only tell us the numbers X and Y (for example, as in Table 1). Can we work out the secret formulae, and the numbers A and B that were rolled?

[Table 1 goes about here]

To proceed, let us plot the values that we get for X and Y as a scatter plot (Fig 2).

[Fig 2 goes about here]

We can see that the outline of the scatter plot forms a lozenge shape, which gives us a clue about how to recover the mixing process. The slopes of the lines formed by the sides of the lozenge tell us the numbers that appear in the two formulae, which tells us how the numbers on the dice A and B were mixed to give X and Y (Fig. 3). We can then solve these two simultaneous equations (i.e. multiply by the inverse mixing matrix) to get back to the original numbers A and B.

[1] This in the accepted version of an article that appeared in *Acoustics Bulletin* 39(6): 44-47, Nov/Dec 2014. The published version is online at http://www.ioa.org.uk/publications/acoustics-bulletin-past

[Fig. 3 goes about here]

In fact, there are two ambiguities that remain after we recover the formulae. The first is a *scaling ambiguity*: A slope of 2 in ½ is the same as a slope of 4 in 1 or 1 in ¼, so we can only discover the relative amounts of A mixed in X and Y. For a die roll we will know that A can only take integer values between 1 and 6, so we can work out which is the correct scaling, but for more general signals this scaling ambiguity will remain. The second is a *permutation ambiguity*: we cannot tell the labels (colours) of the original dice from the mixtures X and Y. So we cannot tell if the original mixing formulae were (a) $X = (1/2)A + 3B$ and $Y = 2A + B$ or $X = (1/2)B + 3A$ and $Y = 2B + A$. So, while we might recover the values for A and B, the values that we get back might be swapped and/or scaled from their original values.

What we have seen is a simple example of *independent component analysis* (ICA) (Hyvärinen et al, 2001). While we solved this unmixing problem visually, ICA uses the statistical independence of the sources to solve the unmixing process, searching for an unmixing matrix $\mathbf{B}$ such that the elements y_i in the vector $\mathbf{y} = \mathbf{Bx} = \mathbf{BAs}$ are independent. When they are, $\mathbf{B}$ will be an inverse matrix for $\mathbf{A}$, subject to a scaling and permutation ambiguity.

Separating *more* mixed sounds

The ICA method is very useful, but since it uses the inverse of the mixing matrix (or equivalently, the solution of a set of simultaneous equations), it is limited to the case where the number of sound sources is the same as the number of observations (microphones). If instead we had more sound sources than microphones, such as a set of three instruments picked up by a 2-channel stereo microphone pair, then we cannot use the ICA method.

Going back to our dice game again: if we have three dice (Amber, Blue, Cherry) but only two mixtures, we cannot see the lines on the scatter plot that would help us to recover the mixtures (Fig 4.). This is not a game we can win, so let us play with a new set of rules based on the concept of *sparsity*.

[Figure 4 goes about here.]

One feature of many audio signals is that, if represent the signal into the time-frequency domain instead of the time domain, its representation is dominated by a small number of time-frequency components. When a signal can be represented or using a small number of non-zero components in this way, we say that it has a *sparse* representation.

For our dice game, one equivalent might be to suppose that we only score something on each of our 3 dice A, B, C if we first roll a six, otherwise we score nothing. For example, 4 scores 0, 2 scores 0, but 6 followed by 4 scores 4. This means that most (5 out of 6) of our scores will be zero. If we now plot the scatter plot of the two secret formulae results X and Y, we can clearly see the mixture lines (Fig. 5).

[Fig. 5 goes about here]

The high proportion of zeros in the sparse sources has led to a concentration of the scatter plots corresponding to the three cases $B = C = 0$, $A = C = 0$, and $A = B = 0$, where only A, B and C respectively are non-zero. So we can now recover the mixing formulae. We can also determine the values of A, B and C where they are each the only non-zero value, but other cases are more difficult: the points on the scatter plot that are “between” lines could be caused by either two or three non-zero values, so we may not be able to recover them uniquely.

A corresponding situation arises for separation of audio sources. Suppose we have a pan-potted (instantaneous) 2-channel stereo mixture of several sources, where each source is panned to a particular L-R direction by mixing level only (no delays). If we transform the 2-channel mixture into the time-frequency domain (a spectrogram), and plot the resulting scatter plot, we might see something like that in Fig. 6.

[Fig. 6 goes about here]

We can see three clear signals that show as bidirectional "rods" sticking out from the centre in the scatter plot (a fourth signal is also just visible in this example, at an angle of about 80°/260° from vertical). Each spike corresponds to a sound source coming from a different direction from L to R, which has a different relative weighting of L and R. If we measure the relative L/R weighting for each time-frequency box in the spectrogram, and "colour" the spectrogram according to the nearest source, we can then "tease apart" the spectrogram using *masking* to give its separate component spectrograms (Fig. 7). By transforming those spectrograms back into the time domain, we recover an estimate of the original separated signals. This approach, known as *time-frequency masking*, forms the basis of the Degenerate Unmixing Estimation Technique (DUET) (Yilmaz & Rickard, 2004) and the Azimuth Discrimination and Resynthesis (ADRess) method (Barry et al., 2004).

Separating from a single microphone

The separation problem is even more challenging with only a single microphone, since we no longer have any direction-of-arrival information to estimate which source is present in each time-frequency box. However, if the sound sources are sparse in the time-frequency (spectrogram) domain, they will be approximately disjoint, and we can use the technique of non-negative matrix factorization (NMF) (Lee & Seung, 2001) to try to decompose the sources.

For NMF, we assume that each note j has a non-negative spectral profile, so our set of notes can be represented by a frequency x note matrix $\mathbf{W} = [w_{ij}]$, and that the non-negative note activities can be represented by a note x time matrix $\mathbf{H} = [h_{jk}]$. If the frequency profiles of the notes are approximately disjoint, the elements of the "frequency x time" mixture spectrogram $\mathbf{V} = [v_{ik}]$ is approximately given by adding up the contributions due to the frequency profile of each note in the ith frequency bin, multiplied by its activity at kth time frame, i.e.

$$v_{ik} \approx \sum_{j} w_{ij} h_{jk}$$

or in matrix notation, $\mathbf{V} \approx \mathbf{WH}$. So if we start with the spectrogram $\mathbf{V}$, we can use the NMF approach to decompose into the separation note spectral profiles and note activities (Fig. 8), resulting in a transcription of the musical notes (Smaragdis & Brown, 2003). While the standard NMF algorithm uses an Euclidean distance, Févotte et al (2009) suggest that approximating using the Itakura-Saito divergence is more suitable for audio sources.

To separate different musical instruments, we then need to group together the notes from the different sound sources. For a simple case we can do this by hand (Wang and Plumbley, 2005) or use convolutive methods (Virtanen, 2004).

We can therefore break our decomposition into sub-matrices corresponding to each of our sources, so our decomposition is split into $\mathbf{V} \approx \mathbf{WH} = \mathbf{W}_1\mathbf{H}_1 + \mathbf{W}_2\mathbf{H}_2 + \cdots + \mathbf{W}_N\mathbf{H}_N = \mathbf{V}_1 + \mathbf{V}_2 + \cdots + \mathbf{V}_N$ where $\mathbf{V}_n = \mathbf{W}_n\mathbf{H}_n$ is the estimate of the spectrogram due to the nth source. We then recover the estimated source by transforming back to the time domain.

Research challenges

For time-frequency masking and NMF to work well, the sources must each occupy different boxes in the transform domain, a property that is sometimes known as W-disjoint orthogonality (Yilmaz & Rickard, 2004). In reality, this orthogonality property does not completely hold, so different types of artefacts can be heard in the separated sources. For example, if two source occupy the same time-frequency box, some of one signal will be mis-allocated to the estimate of a different signal, and we can hear parts of one source signal “bleeding through” to the other.

Another artefact, often called *musical noise* (or *birdies*) arises due to filtering of background noise. In areas where the energy from all sources is low, the value in a time-frequency box may be mostly due to background noise rather than one of the sources. However, the basic algorithm does not detect the difference between noise and signal, and will allocate that noise component to the source with the nearest corresponding L/R ratio. Since each time-frequency box represents a brief burst of energy near the centre frequency for that box, this can be heard as a set of very quiet “bip” sounds at different frequencies, which combine to give a very noticeable “twittering” effect in the background.

Solving these and other problems is a challenge for current researchers. For example, we have been exploring alternative time-frequency transforms to improve separation (Nesbit et al., 2007), the use of additional information such as user input (Smaragdis & Mysore, 2009), knowledge of musical scores (Fritsch & Plumbley, 2013; Ewert et al., 2014), and modelling the local phases and correlations in time-frequency transforms of signals (Badeau & Plumbley, 2014). Many other audio source separation methods are possible, see e.g. (Evangelista et al, 2011) for a review.

Conclusions

Musical audio source separation is an interesting and challenging area of research. Where we have the same number of microphones as sources to be separated, we can use methods based on *independent component analysis* (ICA). In many cases there are more sound sources than microphones, and for those cases we can use methods based on time-frequency masking, including the well-known non-negative matrix factorization (NMF) approach. However, this is still far from being a solved problem, particularly if we are looking for results with high perceived audio quality, suitable for upmixing or remixing applications, and currently may still require significant manual work to get the best results. Nevertheless, there are some promising areas of current research that suggest that progress is still possible, and it will be interesting to see how far we can go towards the target of fully automatic high quality audio source separation.

Acknowledgements

The author is funded by an EPSRC Leadership Fellowship EP/G007144/1, and parts of this work were funded by EPSRC grants EP/E045235/1, EP/H013059/1 and EP/H043101/1 and EU FET-Open Project FP7-ICT-225913 “SMALL”. Many others from the Centre for Digital Music and elsewhere contributed to this work, including Samer Abdallah, Emmanouil Benetos, Thomas Blumensath, Mike Davies, Dimitrios Giannoulis, Maria Jafari, Anssi Klapuri, Andrew Nesbit, Mark Sandler, Dan Stowell, and Beiming Wang.]

This Contribution is the result of a presentation at the IOA Workshop on Sound Recording Techniques, held at the University of Salford on 26 March 2014.

The Centre for Digital Music (C4DM) at Queen Mary University of London is a world-leading centre for research into digital technologies for new understanding and innovation in music and audio. The

Centre consists of over 50 people, including 12 Academic staff, around 35 PhD students and 12 postdoctoral researchers. C4DM has hosted several conferences in music and audio, including DAFx 2003, ISMIR 2005, and CMMR 2012. C4DM has many industry collaborations, including with Yamaha, FXpansion, I Like Music and Audio Analytic, and is a member of the BBC Audio Research Partnership.

References

[Cherry, E. Colin (1953). "Some Experiments on the Recognition of Speech, with One and with Two Ears". The Journal of the Acoustical Society of America 25 (5): 975–79.]

[Hyvärinen, A.; Karhunen, J.; Oja, E. (2001): Independent Component Analysis, New York: Wiley, ISBN 978-0-471-40540-5]

[Yilmaz, O., & Rickard, S. (2004). Blind separation of speech mixtures via time-frequency masking. Signal Processing, IEEE transactions on, 52(7), 1830-1847.]

[Barry, Dan; Coyle, Eugene; Lawlor, Bob (2004) Sound Source Separation: Azimuth Discrimination and Resynthesis, In Proc. 7th International Conference on Digital Audio Effects, DAFX 04, Naples, Italy.]

[D. D. Lee and H. S. Seung (1999) Learning the parts of objects by non-negative matrix factorization. Nature, 401:788–791]

[P. Smaragdis and J. C. Brown (2003) Non-negative matrix factorization for polyphonic music transcription. In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'03), pages 177–180.]

[C Févotte, N Bertin, JL Durrieu (2009) Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis, Neural computation 21 (3), 793-830.]

[B. Wang and M. D. Plumbley (2005) Musical audio stream separation by non-negative matrix factorization. In Proceedings of the DMRN Summer Conference, Glasgow.]

[T. Virtanen (2004) Separation of sound sources by convolutive sparse coding. In Proceedings of the ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing (SAPA 2004), Jeju, Korea.]

[A. Nesbit, M. D. Plumbley and M. E. Davies. (2007) Audio source separation with a signal-adaptive local cosine transform. Signal Processing 87(8), 1848-1858.]

[P. Smaragdis and G. J. Mysore (2009) Separation by humming: User guided sound extraction from monophonic mixtures," in Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, pp. 69-72.]

[J. Fritsch and M. D. Plumbley (2013) Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis. In: Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013), Vancouver, Canada, 26–31 May 2013, pp. 888-891.]

[S. Ewert, B. Pardo, M. Mueller and M. D. Plumbley (2014) Score-informed source separation for musical audio recordings: An overview. IEEE Signal Processing Magazine 31(3):116-124.]

[R. Badeau and M. D. Plumbley (2014) Multichannel high resolution NMF for modelling convolutive mixtures of non-stationary signals in the time-frequency domain. To appear in: IEEE Transactions on Audio, Speech and Language Processing.]

[G. Evangelista, S. Marchand, M. D. Plumbley and E. Vincent (2011) Sound source separation. In U. Zölzer (ed), DAFX: Digital Audio Effects, 2nd edition, Chapter 14, pp. 551-558. John Wiley & Sons. ISBN 978-0-470-66599-2.]

Figure 1: Mixing process of two sources into two microphones.

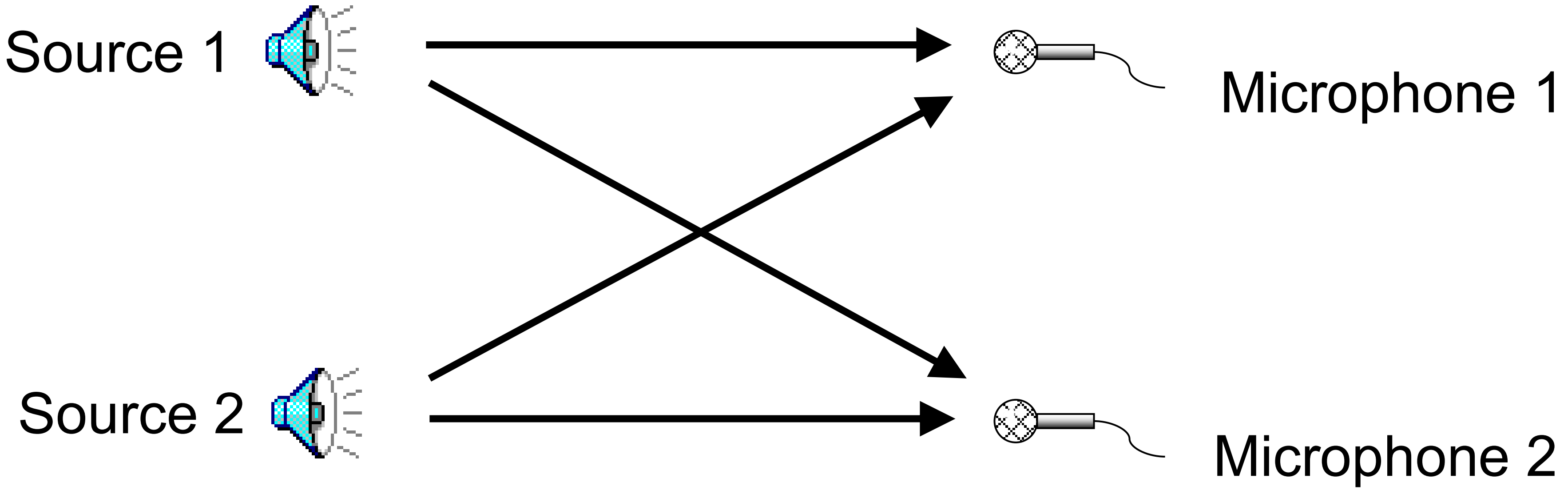
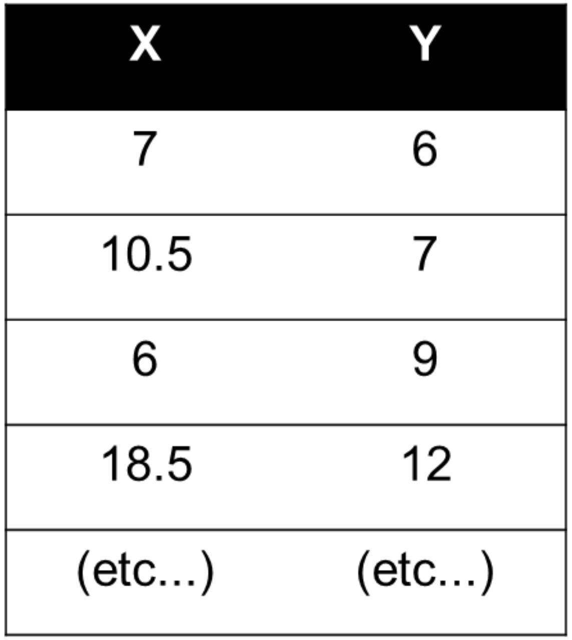


Table 1: Example results of applying the secret formulae to die rolls.



X	Y
7	6
10.5	7
6	9
18.5	12
(etc...)	(etc...)

Figure 2: Scatter plot of secret formula values Y against X, highlighting the point X=7, Y=6.

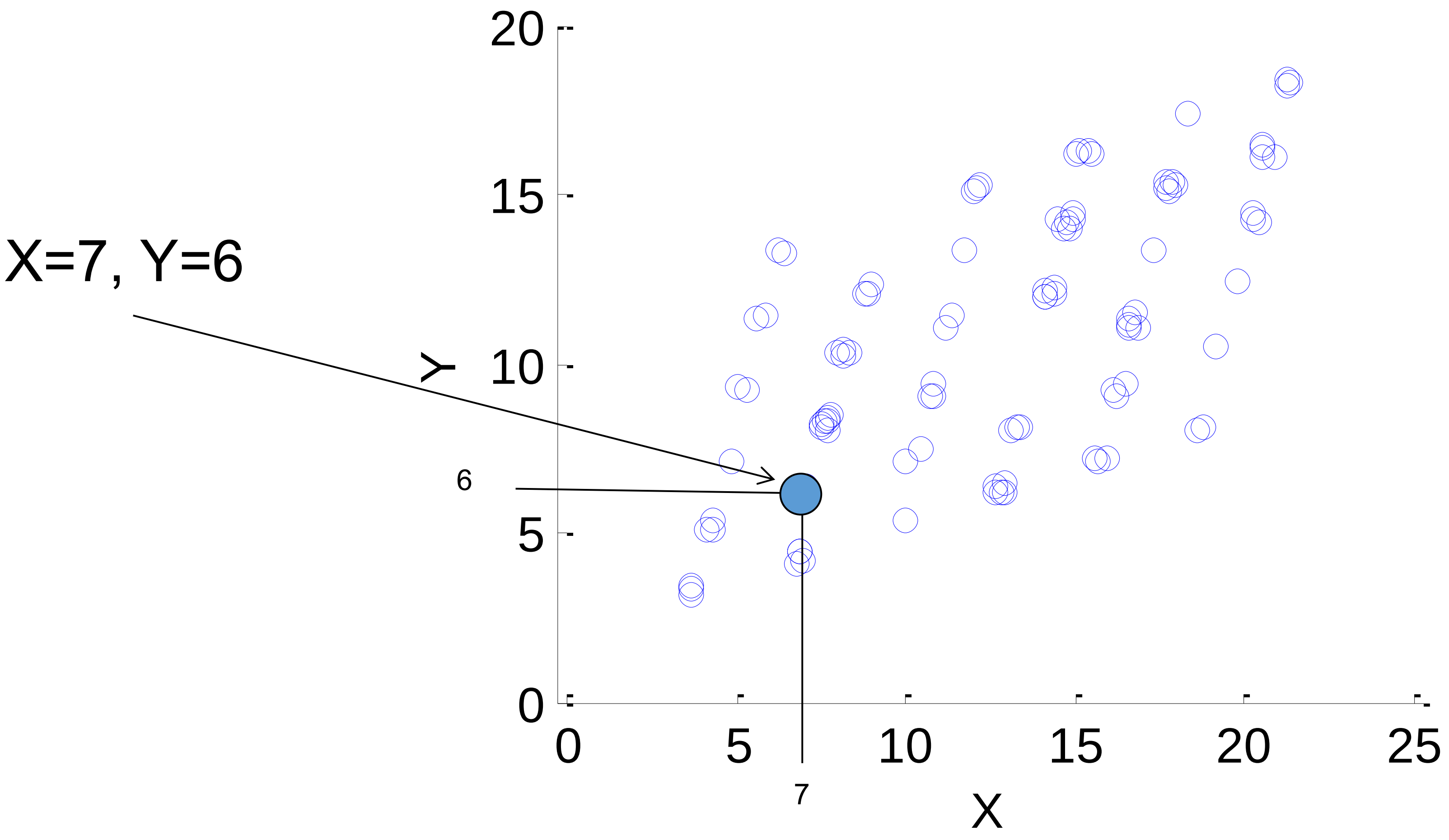


Figure 3: Reading off the mixtures from the scatter plot

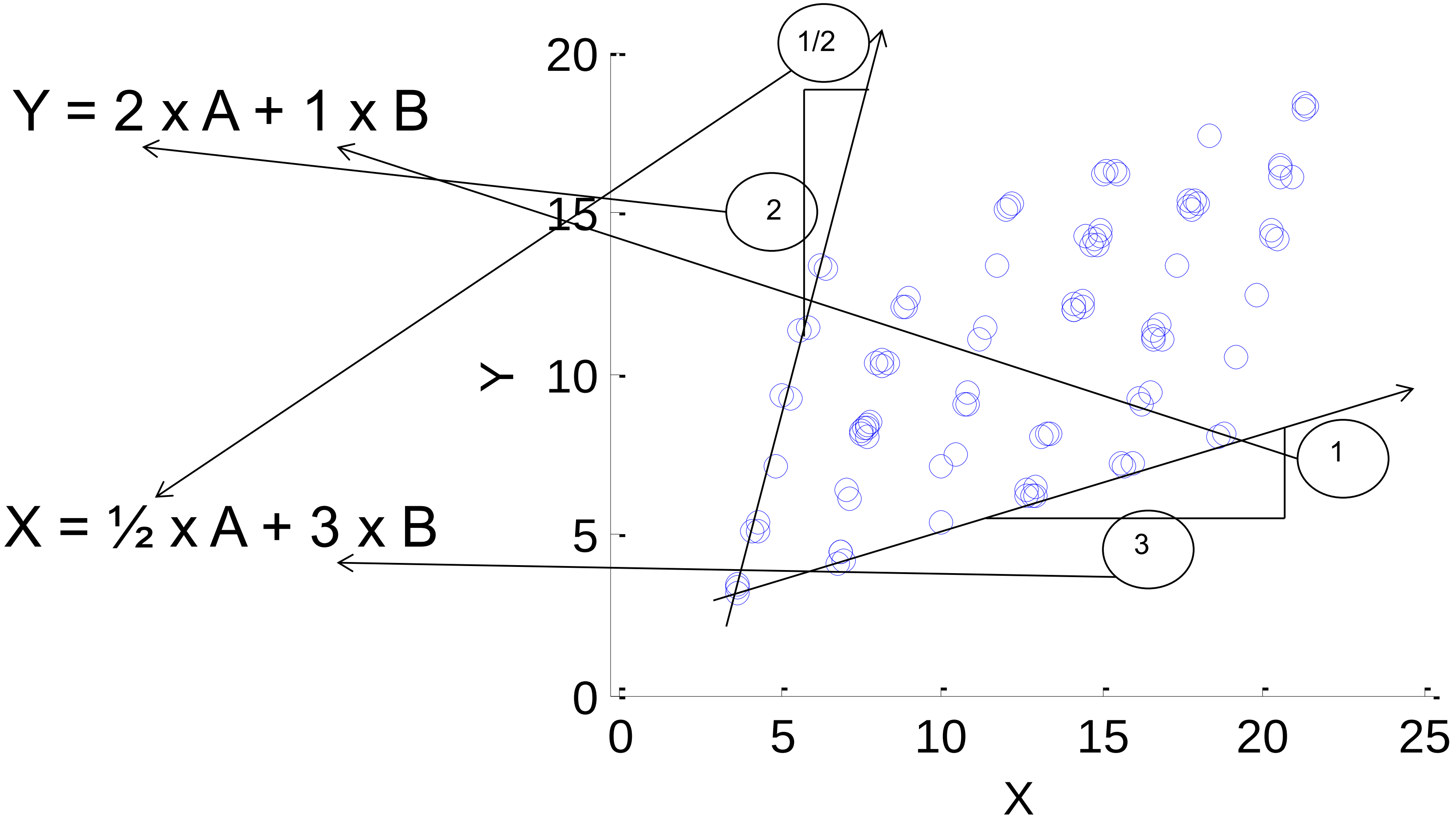


Figure 4: With three sources but only two mixtures, the scatter plot no longer allows us to find the mixtures.

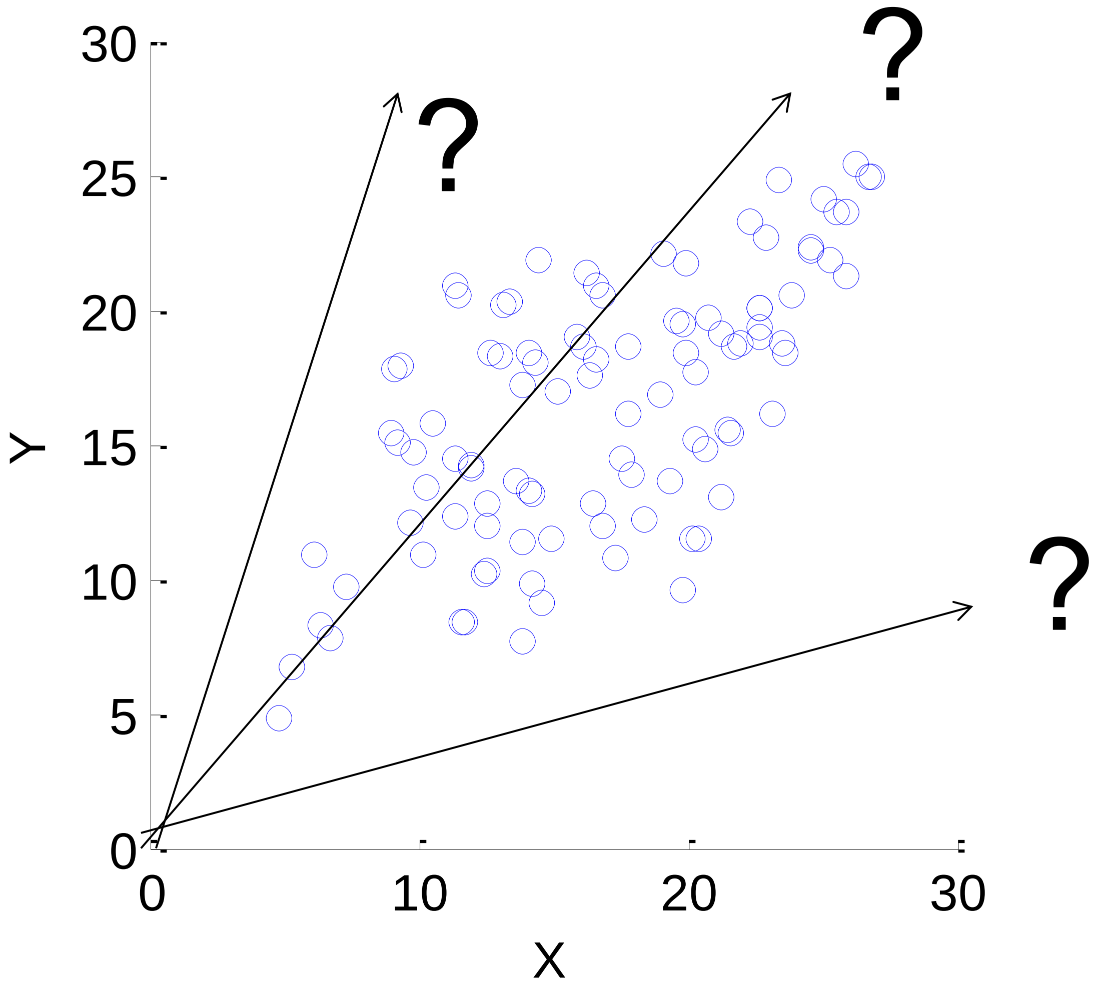


Figure 5: Scatter plot of two mixtures of three sparse sources.

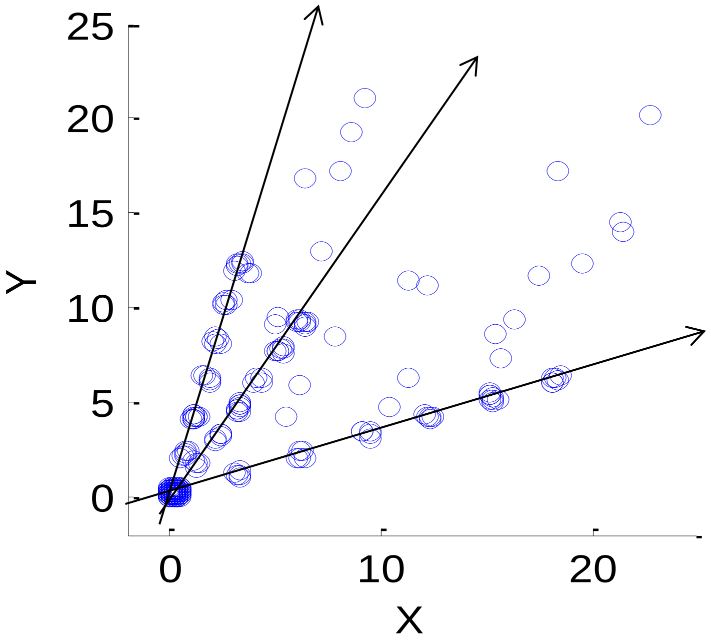


Figure 6: Scatter plot of the real values of a time-frequency transform of pan-potted stereo signal.

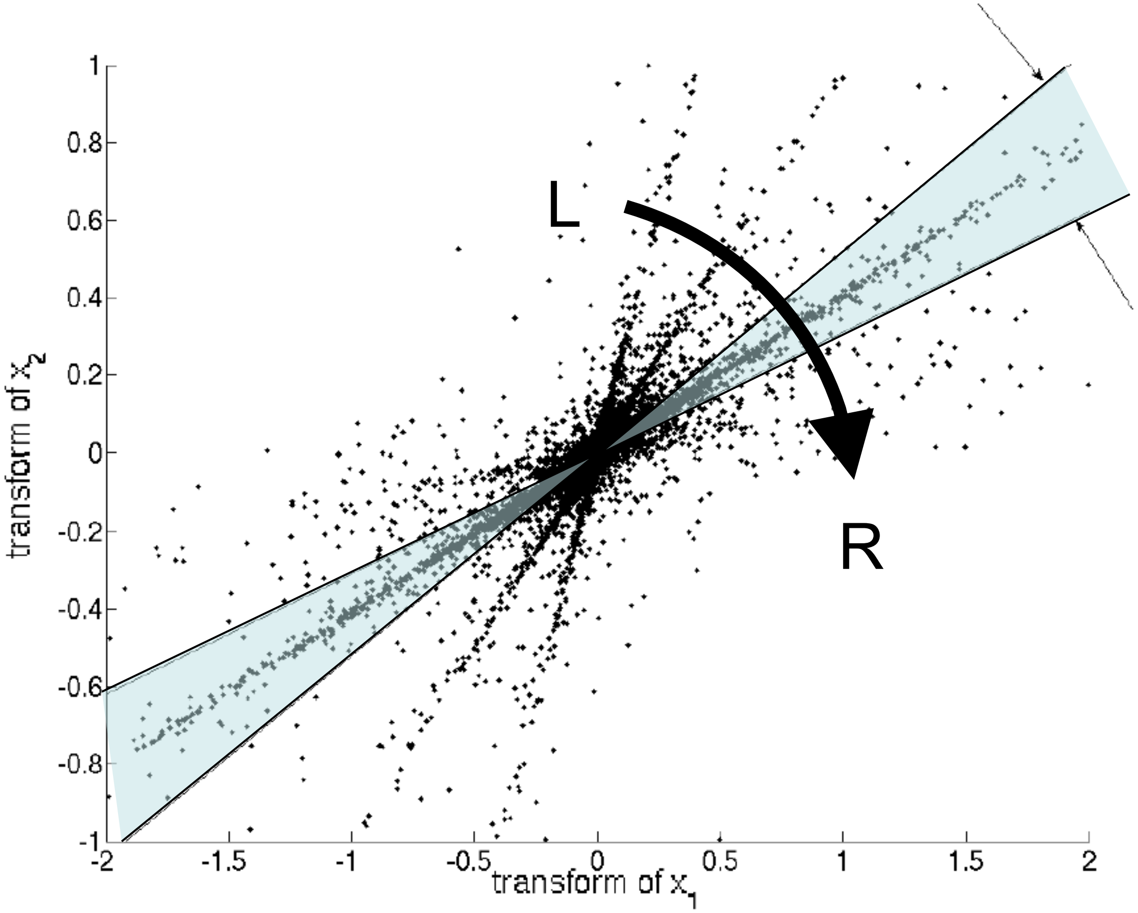


Figure 7: Separation by colour-coding the spectrogram

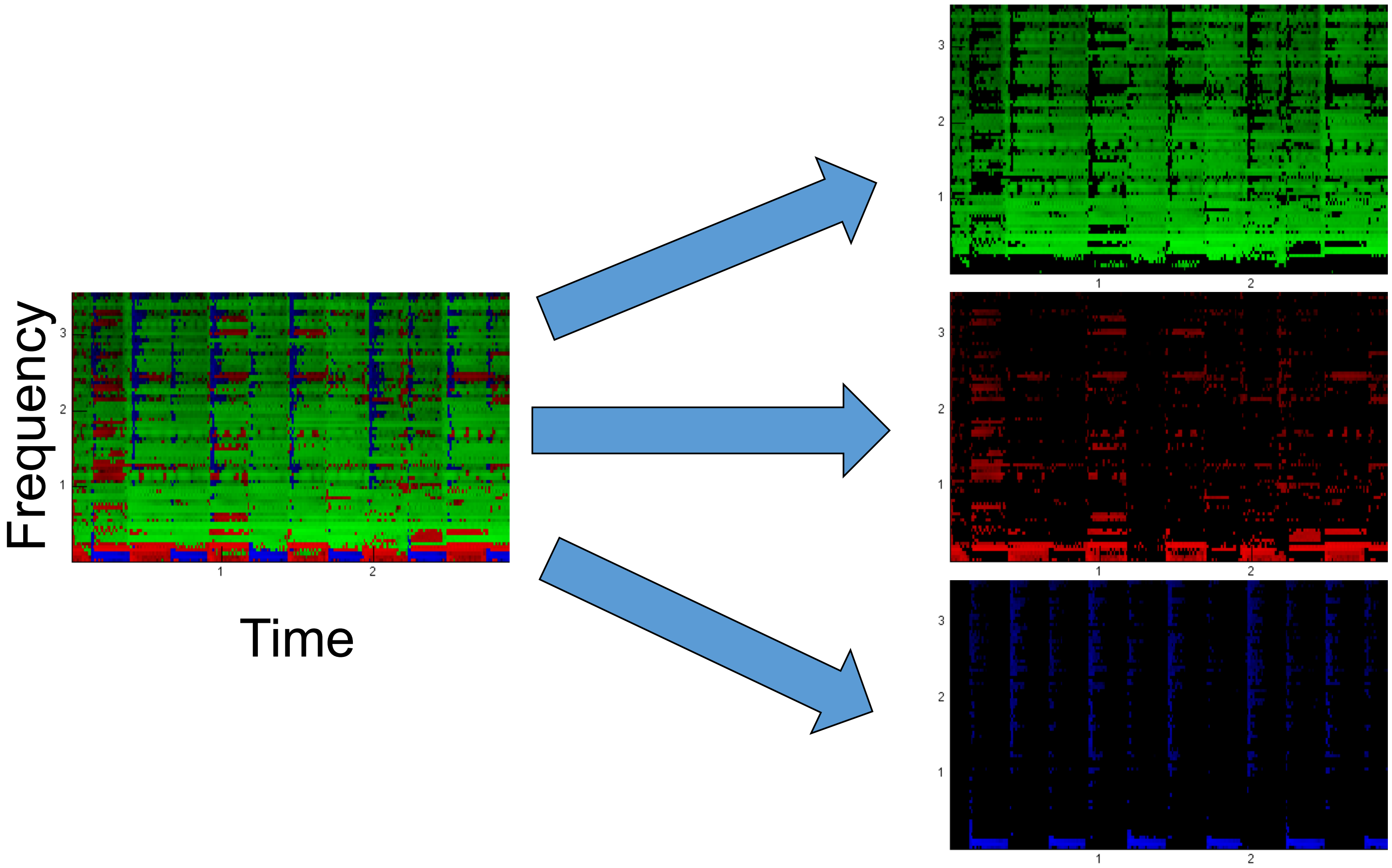
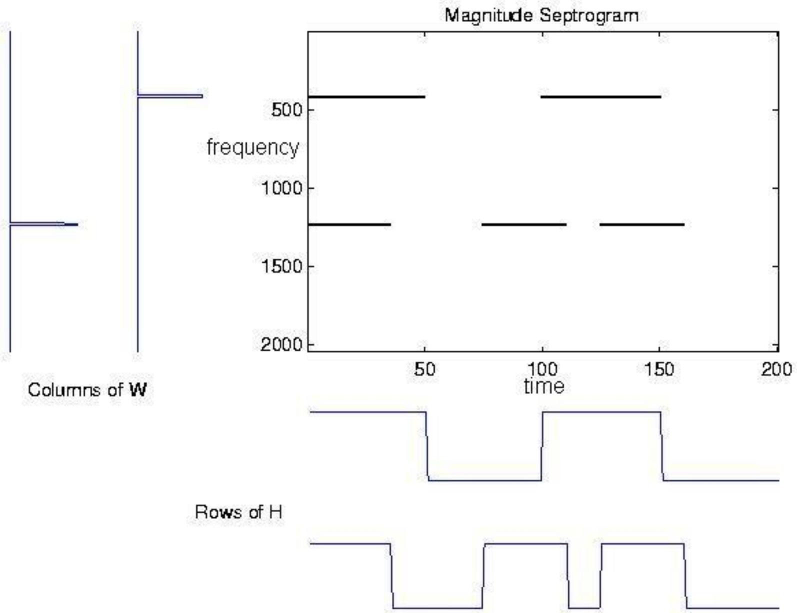


Fig. 8: Illustrative example of decomposition of a spectrogram by non-negative matrix factorization (NMF)


Magnitude Septrogram
frequency
time
Columns of W
Rows of H