7 - 12 SEPTEMBER 2014, KRAKÓW
FORUM ACUSTICUM

Environmental Sound Classification in Realistic Situations

K. Haddad, W. Song
Brüel & Kjær Sound and Vibration Measurement A/S, Skodsborgvej 307, 2850 Nærum, Denmark.
X. Valero
La Salle, Universistat Ramon Llull, Quatre Camins 30, 08022 Barcelona, Spain.

Summary
There is an increasing interest for automatic classification of sounds in various applications. The processing consists in extracting some selected features from the raw data and applying a classifier algorithm to automatically estimate the class of source the input data belongs to. In order to perform this task, the system is first trained on a set of input data (in case of supervised learning), then applied on new data. The performances of classification depend on many factors: the selected features, the classifier, but also the data. For some practical applications, this approach is of interest to estimate the sound pressure level or others acoustic quantities attached to a specific class of sources in a complex environment. For example, there is a benefit to quantify the only contributions of airplanes, in the sound acquired by a noise monitoring system located nearby an airport. One practical difficulty is the presence of other sources. In the present study, we are looking at this type of scenario, when multiple sources may act at the same time or in presence of background noise. We consider here different environmental sound sources: aircrafts, car traffic among others. In a first step, different features in connection with a classifier are evaluated on 'clean' data, meaning with no mixture. In a second step, we mix artificially data from different sound sources at different rates of mixtures and calculate the sound pressure level contribution from each source based on the proposed classification algorithm. We consider as well mixtures with sounds sources that are not among the predefined classes. Finally the most robust classification configuration is evaluated in the case of real outdoor measurements.

PACS no. 43.50.Rq, 43.60.Lq

1. Introduction

Environmental sounds represent a quite broad range of acoustic signals generated by different types of sources: transportation, industrial activities noise, human activities, animals, natural phenomena (rain, storm…). In many cases, they are considered as noise pollution that we want to reduce the impact, such as for transportation or industrial noises. Nevertheless, there are more than one dominant source in general. Therefore, in order to reduce the noise pollution at a particular site, it is useful to be able to recognize different sources and to rank their contributions. As an example where such process is useful: around airport, citizens and authorities require a noise impact evaluation of aircrafts in the neighborhood. However often the measurement terminals are located in different areas where potentially other sources could contribute to the level estimation. In such case, a system that could estimate the presence and the contribution of only the aircrafts is desirable.

From another perspective, environmental sounds are not always considered as negative noises. Soundscape approach aims at understanding or eventually at shaping the sound environment at a selected site. Again, for this purpose, tools to recognize and rank sound contributions are also useful.

Different approaches are possible to perform these tasks. As examples of techniques, we can briefly mention as a non-exhaustive list, denoising techniques (spectral subtraction, Wiener filtering), blind source separation techniques (such Independent Component Analysis) or microphone array processing. Each one having assumptions, advantages and drawbacks, depending on the

signals type (stationary or not for example), the location of the measurements (presence of sound reflections and diffraction), the distance between acoustic sources.
Since we are interested at performing a recognition of sound sources, we explore here different ways to obtain contributions from a mixture based on a classification approach.
The first way is based on a classifier called Gaussian Mixtures model, the second is linked to another classifier called Fisher Linear Discriminant and the last one is derived from a Non-negative Matrix Factorization technique. We evaluate performances on artificially mixed sounds and on a realistic scenario.

2. Classification of environmental sounds

Sound classification derives directly from machine learning techniques. The purpose is to create a system that learns from the data, how to classify them in different classes. Therefore, there is a training phase where the system is set up with representative data. When the training data are labelled, it is called supervised learning. We use this approach in this paper. The process consists generally of two steps. We first extract some features, in order to reduce the dimensionality of the input time signals. Then we apply a classifier which determines the class of a given signal. Different types of features and classifiers exist. We discuss some of them below and estimate their performances.

Different features and classifiers have been successively applied to different classes of environmental sounds [1, 2]. As popular examples of features, we can mention the Mel-Frequency Cepstral Coefficients (MFCC), the Linear Prediction Cepstral Coefficients (LPCC) and the Perceptual Linear Predictive (PLP). Different classifiers have been used for these sounds. We describe briefly below the Gaussian Mixture Model (GMM) classifier and the Fisher Linear Discriminant (FLD) classifier, since they are applied below to analyze mixtures of sounds. We also introduce the Non-negative matrix factorization (NMF) technique, as a tool to decompose mixtures into source contributions.

2.1. Classification using a Gaussian mixture model

GMM represents a parametric model of continuous features as a linear combination of Gaussian component densities. The linear combination can be expressed as [3]

$$p(\boldsymbol{x}) = \sum_{i=1}^{N} w_i g(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \quad (1)$$

where N is the number of Gaussian component densities, w_i is the weight of the mixture, and $g(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is a Gaussian function having mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. The model can be derived using an expectation maximization (EM) algorithm [3] based on the maximization of the likelihood function of the model.

2.2. Classification using a Fisher Linear Discriminant

FLD [4] projects M-dimensional vectors to an optimal line of separation so that inter-class scatter is maximized while the intra-class scatter is minimized. This can be achieved by maximizing the following cost function.

$$\boldsymbol{J}(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{\Sigma}_B \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{\Sigma}_w \boldsymbol{w}} \quad (2)$$

where $\boldsymbol{w}$ is a weighting vector that is used for the projection of feature vectors, $\boldsymbol{\Sigma}_B$ is the inter-class covariance and $\boldsymbol{\Sigma}_w$ is the intra-class covariance. Once the weighting vector is derived, the separation line between classes can be defined as the mid-point from the means of the classes.

2.3. Non-negative matrix factorization

NMF is derived from the idea that a signal could be decomposed into components that add up constructively in term of magnitude. This is a good approximation for sound signals in many cases, when using the magnitude of the power spectrum $\boldsymbol{X}(t,f)$. Applying this idea to short-term power spectra, we can decompose the signals as [5]:

$$\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H}. \quad (3)$$

$\boldsymbol{W}$ is a matrix that contains basis vectors. They define the frequency content of each component. The number of columns specifies the number of components. $\boldsymbol{H}$ is an activation matrix indicating the timing of activity for each component. To achieve this decomposition, different criteria are possible. One of the most popular consists in using the Kullback-Leibler divergence to minimize the distance between the two terms [6]. We can add different constraints in the minimization problem to improve the decomposition, such as for sparsity or temporal continuity criteria [6].

3. Mixture of sounds

3.1. Different approaches

In the following sections, we are interested in using sound classification framework to estimate contributions from the different sounds included in the recorded signals. Our first approach is related to a Gaussian Mixture Model classifier. It is somewhat a natural path, since it is a probabilistic model that provides a probability of belonging to different classes. Our second technique consists in using a Fisher Linear Discriminant classifier. It is a different approach for the estimation of contributions in a mixture, since this classifier establishes borders between classes. The last investigation in this report is derived from Non-negative Matrix Factorization processing. It has be shown to be a successful tool to decompose an audio signal into components [5]. We apply it here for sound classification purpose. These techniques applied to mixtures are explained in the following sections.

3.2. Contribution of a target source using Gaussian Mixtures Model

GMM has been applied to classify the source of environmental noises [7], but the method has been restricted to a single source classification. GMM assigns the percentage of belonging to a class in such a way that the sum of percentages becomes always 100%. This may be particularly problematic for outliers, which does not belong to any of the trained classes.

Hansen (2012) [7] defined a threshold by relating the Mahalanobis distance between an observation vector and the each of component Gaussian densities trained during the process of learning. The derived threshold takes into account the distribution of each component Gaussian density function in contrast to a direct distance measure from the mean vector of the component Gaussian distribution to the new data point. The current investigation will show how such a threshold can be used to estimate the contribution from a target environmental source within a mixture of environmental noises.

3.3. Contribution of multiple sound sources based on Fisher Linear Discriminant

The distance from the separation line to a new projected observation point determines which of the trained classes the new data belongs to. When having a mixture of environmental noise sources, there may exist more than one positive (belonging to the class) distances. The contribution of individual noise sources to the mixture may be calculated by the ratio between the individual positive distance and the sum of positive distances [8]. In this method, any arbitrary mixture of sources can be considered assuming individual sources can be isolated for training the FLD algorithm.

3.4. Direct estimation of sound contributions using Non-negative Matrix Factorization technique

In the context of classification, we would like to obtain the basis vectors for training data belonging to the same class. We can then build different sets of basis vectors for each of the classes. We can arrange them in a matrix form. For our dataset, we obtain $\boldsymbol{W_c}$ for the 'car' class, $\boldsymbol{W_a}$ for the 'aircraft' class and $\boldsymbol{W_t}$ for the 'train' class. In the second phase of the classification process, we decompose the input signal according to (3), but this time $\boldsymbol{W}$ is fixed and set to $\boldsymbol{W_c}$, $\boldsymbol{W_a}$ or $\boldsymbol{W_t}$. The output for each of these cases provides the contribution of 'car', 'aircraft' and 'train' sources in the recording. In the results section, we apply the technique described in [9, 10].

4. Application to real environmental sounds

4.1. Dataset

We are particularly interested in transportation sounds for this paper. Three main classes are included in this study: cars, aircrafts and trains in traffic conditions. The recordings were obtained in various traffic conditions using a free-field microphone. The database contains 469 seconds of aircraft noise, 332 seconds of car noise and 158 seconds of train noise. We consider as well other types of sources, such an artificial white noise added to the predefined classes.
To test the different techniques, we generate mixture of recordings at selected mixing rates. In the following sections, we mix signals from different classes: car-train or car-aircraft for examples. In the case of GMM, we mix car traffic noise with white noise.

4.2. Results

4.2.1. Contribution of a target source using GMM

The performance of a classification algorithm is heavily influenced by the extracted features from environmental noise measurements. The recognition rate of different feature-classifier combinations is summarized in Table 1. For both

GMM and FLD algorithm, MFCC results in the best recognition rate. For this reason, the following investigation uses only MFCC as feature.

feature / classifier	MFCC	LPCC	PLP
FLD	90.7	72.1	90.6
GMM	88.7	56.1	85.0

Table 1 Recognition rate in percentage for different feature-classifier combination

To demonstrate a possible application using the GMM threshold, a highway noise of 800 s is edited to contain white noise having various sound pressure levels in the half of the total duration. The upper graph in Figure 1 shows how the white noise is added to the highway noise. The level of white noise is controlled so that a broad range of signal-to-noise ratio (SNR) is covered. The resulting mixture is divided into 100 milliseconds frames, and the GMM threshold is applied to each of these frames to find out whether the frame belongs to the high way noise. If it is classified as an outlier, then the level of the frame is set by interpolating the level of neighboring frames.

When SNR is poor, the influence of added white noise is significant (see total L_{Aeq} in Figure 1).

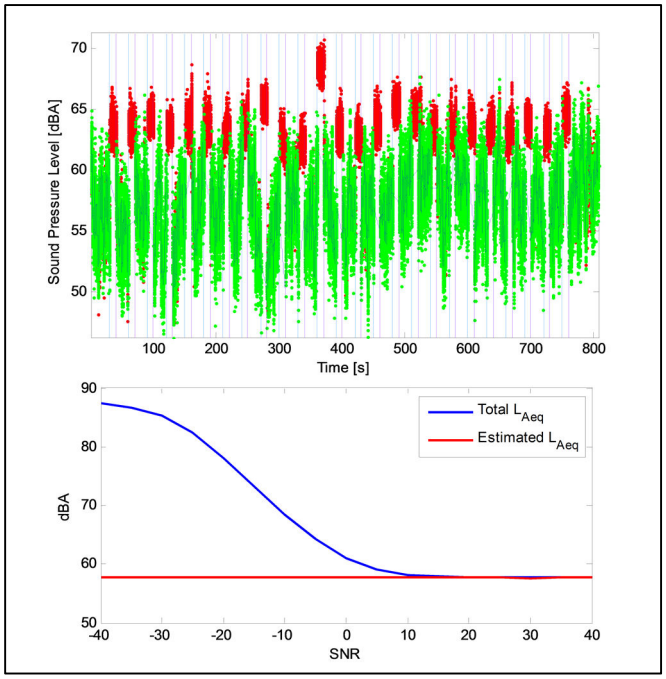


Figure 1 Accurate LAeq estimation of highway noise in the presence of white background noise. The SNR lower than 0 dB means that the level of white noise is higher than the level of highway noise.

However, the GMM threshold method removes the most of the frames containing the white noise sample, and the estimated L_{Aeq} is almost unchanged independent of the level of white noise added. While the threshold successfully removes the influence of interference noises, the method may not be used to quantify the contribution of all individual sources in terms of their sound pressure level since it does not estimate the level of outliers in a mixture.

4.2.2. Contribution of individual sources in a mixture using FLD

In contrast to the GMM threshold method, the percentage of belonging to the class in a mixture using FLD can be useful to quantify the contribution from individual noise sources. However, the percentage of belonging to the class cannot be directly used to quantify the Sound Pressure Level (SPL) contribution, and for that, we need to relate the SPL contribution to the percentage of belonging for artificially mixed signals. Notice that this relationship holds only for the specific conditions where the relationship is known.

In order to get this empirical relationship among noise sources, a pair of noise sources, e.g. train noise and airplane noise, is mixed with different combination of sound pressure level. This is achieved by fixing the SPL of one source while the SPL of the other source is adjusted accordingly. The percentage of belonging to the class is displayed along with the corresponding target noise to total ratio in terms of SPL (see Figure 2). There is a monotonic relationship between the two quantities, meaning that the target SPL may be calculated reliably by the percentage of belonging from FLD. The relationship is also derived for train vs. car and car vs. airplane.

To illustrate how the percentage of belonging behaves for more realistic scenarios, a time recording was performed at a place close to both a train station and a highway. Figure 3 shows the resulting percentage of belonging as well as the direct classification result when a train passed by while the highway noise was present. The upper graph displays the traditional FLD result, in which only one dominant class is identified for each frame. On the other hand, the lower graph shows the percentage of belonging, and it follows quite well with the temporal influence of train noise on the continuous highway noise. As indicated earlier, the empirical relationship only holds for the

specific sources, with which the classifiers are trained. In practical applications, isolating individual sources for the training process may be challenging to achieve.

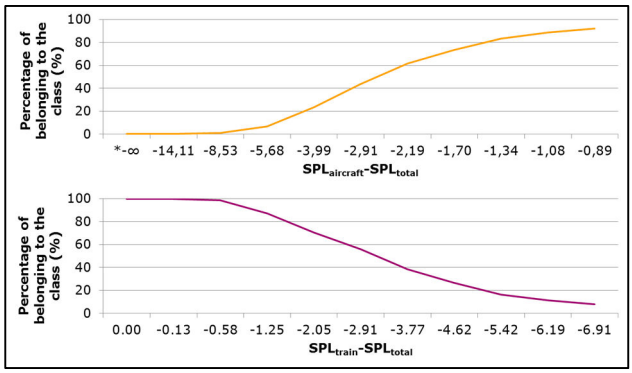


Figure 2 The relationship between the calculated percentage belonging to the class and the contribution of a specific noise type to the overall sound pressure level.

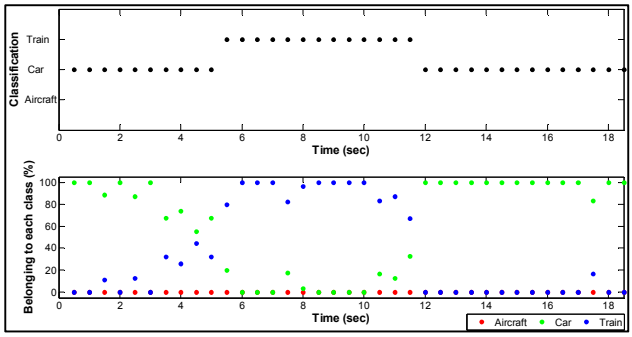
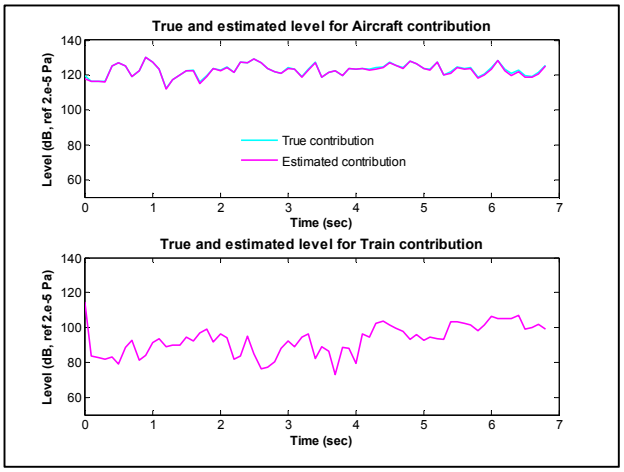


Figure 3 Classification of noise type (above) and percentage of belonging to each class (below) for an example, in which a train was passing close to a high way.

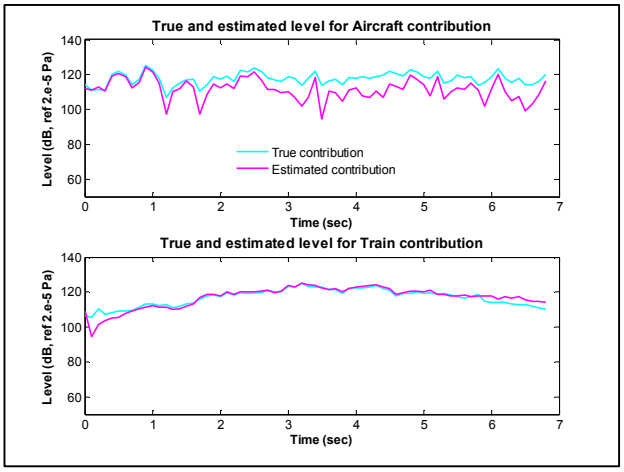
4.2.3. Sound Contribution from NMF

The NMF framework provides a direct way to extract sound contributions and from there, an estimate of the sound pressure level for the different sources. Using each of the basis vectors matrix linked to one class of sound, we can calculate its contribution in the recording. To evaluate the performances of this approach, we proceed the same way as for FLD in previous section: we create different mixtures. Knowing the true level for each of the sources in the mixture, we can compare them with those provided by the algorithm. In the following examples, we look at a mixture of an aircraft and a train at different mixing ratio. The audio samples used to create the mixtures are not part of the training set.

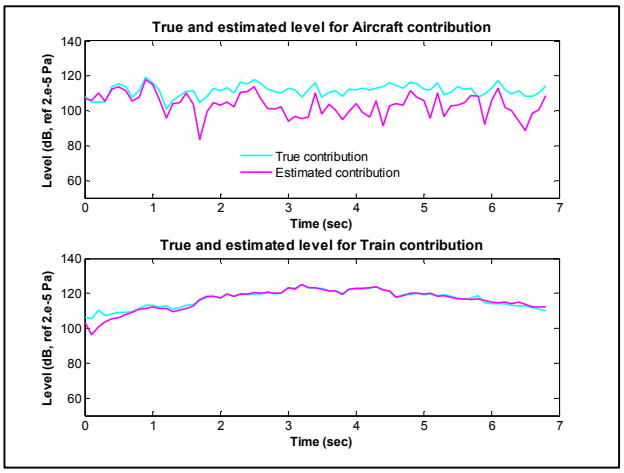
In Figure 4, we compare the true and estimated contribution levels for the aircraft noise in one hand and for the train noise in the other hand.



(a)



(b)



(c)

Figure 4 Comparison of true and estimated level after NMF. The signal is a mixture of aircraft and train sound at different mixing rate. (a): No train contribution, (b): the same overall SPL for both contributions, (c): Overall SPL(Train) = Overall SPL(Aircraft)+3dB.

The performances are visibly depending on the mixing ratio. When only aircraft noise is present in

the signal, the estimated contribution of train noise is approximately 30 dB below the contribution of aircraft. In this case, the aircraft noise contribution is accurately estimated.

When the part of the train sound contribution increases in the mixture, we notice bigger differences between estimated and true aircraft noise contributions. It reaches about 6 dB error in average when overall aircraft SPL is equal to overall train SPL and about 8 dB when overall train SPL is 3 dB higher than overall aircraft SPL (we can see locally much higher errors, up to 15 dB). However, the estimated aircraft noise level is always close to the true one at the beginning of the recording: the train noise is not much present in this section of the signal. Concerning the train noise, the estimation of its level is relatively accurate for all the tested configurations.

In order to test the method on a realistic scenario, we apply this technique to the sample used for testing the FLD: a train passing by close to a high way. Figure 5 shows that the crossings of levels between the two contributions happen at the same times as indicated by FLD above. We notice also that estimated contribution from the car noise is relatively constant (with a loss of about 5 dB when the train is coming and leaving).

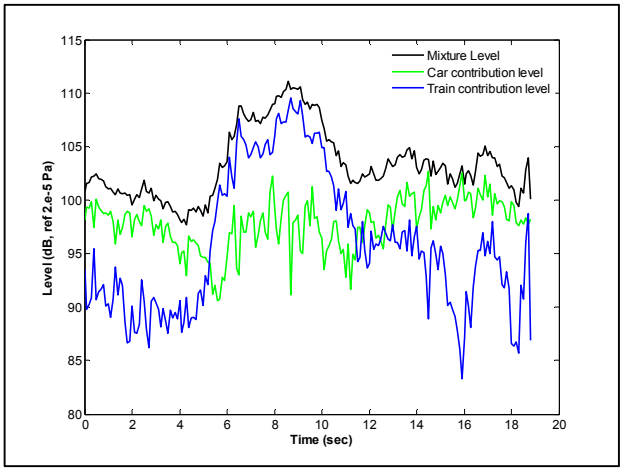


Figure 5 Comparison of mixture, car contribution and train contribution levels. The decomposition is based on NMF. The signal is the same as illustrated in figure 3.

5. Conclusions

We have presented three different techniques to analyze mixtures of environmental sounds, in the context of classification. The first one is based on a GMM classifier, the second is derived from a FLD classifier, and the last is based on one NMF algorithm. In the present study, we consider three classes of environmental sounds: train, aircraft and car noise in traffic condition. GMM is applied here on a case where a predefined class is mixed with a different competing sound. It quantifies the SPL of a target source successfully. The results from FLD approach show that we can estimate contributions in term of percentage of belonging to each class. From there, we can derive the individual SPL contributions. The thirst approach, NMF, alone does not provide a classification, but it directly decomposes recordings into signals that can be identified to a sound class. In this case, the derivation of the SPL for each sound contribution is straightforward. In a future work, we will expand the number of classes, and investigate how individual contributions can be calculated in the presence of non-trained sources.

References

[1] Sachin Chachada, C.-C. Jay Kuo: Environmental Sound Recognition: A Survey. Proc. Signal and Information Processing Association Annual Summit and Conference (APSIPA), Asia-PacificAcustica 4 (1954) 594-600, 2013.

[2] M. Cowling and R. Sitte: Comparison of techniques for environmental sound recognition. Pattern Recognition Letters, Vol 24, pp 2895—2907, 2003.

[3] C. M. Bishop: Pattern recognition and machine learning, springer New York, 2006.

[4] R. A. Fisher: The Use of Multiple Measurements in Taxonomic Problems, Annals of Eugenics, Vol 7, pp 179-188, 1936.

[5] P. Smaragdis and J. C. Brown: Non-negative matrix factorization for polyphonic music transcription. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp 177–180, 2003.

[6] P. Smaragdis, C. Févotte, G. Mysore, N. Mohammadiha and M. Hoffman: Static and dynamic source separation using nonnegative factorizations: A unified view, IEEE Signal Processing Magazine, May 2014.

[7] T. H. Hansen, Classification of Environmental Sounds. Pattern Recognition. Report 2 for bachelor internship, Technical University of Denmark, 2012.

[8] Ester Creixell, Karim Haddad, Wookeun Song, Shashank Chauhan and Xavier Valero: A method for recognition of coexisting environmental sound sources based on the Fisher's linear discriminant classifier, Internoise, 2013.

[9] T. Virtanen, J. F. Gemmeke, and B. Raj. Active-Set Newton Algorithm for Overcomplete Non-Negative Representations of Audio, IEEE Transactions on Audio Speech and Language Processing, volume 21 issue 11, 2013.

[10] T. Virtanen, B. Raj, J. F. Gemmeke, and H. Van hamme, Active-set Newton algorithm for non-negative sparse coding of audio, in proc.International Conference on Acoustics, Speech and Signal Processing, 2014.