Acoustic Event Classification using Low-resolution Multi-label Non-negative Matrix Deconvolution

LODE VUEGEN[*⋆†], PETER KARSMAKERS[*†], BART VANRUMSTE[*†‡§] AND HUGO VAN HAMME[⋆]
(lode.vuegen@kuleuven.be)

[*] *KU Leuven, Dpt. of Electrical Engineering, ESAT-ETC-ADVISE, Kleinhoefstraat 4, B-2440 GEEL, Belgium.*
[⋆] *KU Leuven, Dpt. of Electrical Engineering, ESAT-PSI, Kasteelpark Arenberg 10, B-3001 LEUVEN, Belgium.*
[†] *KU Leuven, Dpt. of Electrical Engineering, ESAT-STADIUS, Kasteelpark Arenberg 10, B-3001 LEUVEN, Belgium.*
[‡] *KU Leuven, Dpt. of Electrical Engineering, ESAT-ETC-eMedia, Andreas Vesaliusstraat 13, B-3000 LEUVEN, Belgium.*
[§] *IMEC, Remisebosweg 1, B-3001 LEUVEN, Belgium.*

Acoustic event classification for monitoring applications is becoming feasible thanks to the increasing number of connected devices with a built-in microphone. The sound event classes are defined by annotating training data, which is a laborious process. Attempts have been made to reduce the workload on annotating the vast amounts of training data, and are referred to as semi-supervised learning and active learning. In this paper, we propose a non-negative matrix deconvolution (NMD) based approach, capable of modelling acoustic events from data labelled on a low-resolution and multi-label level and thereby reducing the annotation workload. We further show that the proposed extension of NMD is successfully applied for the classification of acoustic events, even in noisy conditions and with overlapping events.

0 INTRODUCTION

Acoustic event detection and classification (AED/C) have recently become key challenges in the field of acoustic signal processing and are already widely examined for numerous different applications [1]. Over the past few years it has been investigated for surveillance and healthcare related applications [2, 3, 4], tracking and classifying sound sources of military interest [5], automated wildlife observations [6] and diagnoses of industrial machinery [7]. The promise and success of all these applications relies on robust sensing of the environment. Nowadays, collecting acoustic data has become more accessible than ever before due to the ever increasing number of connected devices in our daily lives containing a built-in microphone (e.g. smartphones, smartwatches, tablets, laptops, televisions, remote cameras, etc.). As a result, recordings can be made continuously all day round with minimal effort. However, the development of acoustic event detection and classification algorithms itself requires labelled data to learn the model-specific parameters. Accurate labelling the data takes at least the duration of the recording and is thus often the main cost in the process of developing a robust sound event classifier.

The amount of effort that can be spent on labelling the data is typically expressed by *'labelling budget'* and defines the maximum number of labels that can be assigned. In situations where the labelling budget is low, or in cases of very large datasets, there are multiple established strategies available in the literature to utilise the abundant amount of unlabelled data with minimal labelling effort. The two best known strategies are *'semi-supervised learning'* and *'active learning'*.

The majority of the semi-supervised learning (SSL) algorithms are either based on the so-called self-training approach or on the pre-training methodology. Self-training permits to automatically annotate unlabelled data by using a pre-existing model trained on a smaller set of labelled data. The classified instances together with the predicted class labels are added to the training data and the model is retrained. This procedure is then repeated iteratively until a certain target performance is achieved or until no more unlabelled data is available. This allows to deal with unlabelled data and is already examined for various acoustic pattern recognition tasks such as speaker identification [8] and musical instrument recognition [9]. Pre-training SSL approaches on the other hand are often used for training deep neural nets (DNN). This methodology starts with learning the feature space from unlabelled data followed by a tuning phase on a smaller set of labelled instances as discussed in [10].

Active learning (AL) approaches on the other hand asks for labelling input on data selected by the algorithm from the set of unlabelled data. The annotated labels are then

used to train an initial classifier and the remaining unlabelled samples are classified using the learned classifier. A batch of samples with the lowest classification certainties are presented by the active learning algorithm to the user for verification and are used to update the learned classifier. This procedure is repeated until a stopping criterion is met. Certainty-based sampling has already been widely studied in the field of acoustic pattern recognition [11, 12]. A slightly different approach of active learning is when the classified instances are directly observed by the user and corrected if necessary. This strategy does not require a certainty based metric for sampling and is recently successfully applied for acoustic pattern recognition tasks such as recognising speech commands from dysarthric speech [13].

All the above-mentioned techniques rely on a classifier either for sampling or prediction of the unlabelled data. We aim at learning the model-specific parameters from both the labelled and unlabelled instances directly by utilising the internal structure of the data. An already widely examined technique for structure discovery in the field of acoustic signal processing are the so-called *'non-negative matrix factorisation'* (NMF) based approaches and their extensions. The main objective of NMF is to decompose an all positive data matrix into a product of two (or more) non-negative lower dimensional submatrices called factors. As a result of the restriction to positive values only, the factors tend to model the data matrix as a linear combination of additive components and thereby revealing the underlying latent structure. In this research, we will focus on using a convolutive variant of NMF, called *'non-negative matrix deconvolution'* (NMD), and is capable of identifying components with a temporal structure as well [14]. Although both NMF and NMD tend to a parts-based representation of the data, there is no absolute guarantee for this behaviour. In this work we present an extension to NMD where the parts-based solutions are promoted by incorporating the available class labels in the learning procedure of NMD. Label information must only be available at a *low-resolution*, i.e. an indication of which events have occurred in an annotation segment, without identifying beginnings nor endings of the individual events. This also leads to a *multi-label* approach in training and classification: a sound segment can contain multiple (different) events which may or may not overlap with each other in time. Lastly, we will assume that not all occurring events in an annotation segment are annotated as occurring. In other words, we assume annotations are incomplete and available at a low temporal resolution, which significantly reduces the annotation cost.

Convolutional neural networks (CNN) are currently the most popular method for sound event classification [15, 16, 17]. However, CNN typically require massive amounts of data to estimate their parameters in good order using a discriminative objective. Furthermore, their feature representations are often hard to interpret. NMD in its original form is a generative data modelling approach that uses optimisation to learn interpretable and consistent data structures potentially using a smaller amount of data compared to CNN. The proposed low-resolution multi-label NMD alternative balances between a generative and discriminative learning objective for even enhanced interpretability. Note that both NMD and CNN can complement each other. The convolutive event model learned by NMD might replace a convolutional layer in CNN. Thereby feeding a deep neural network structure with dictionary element activations as input.

The remainder of this paper is organised as followed: in Section 1 convolutive event modelling is thoroughly explained together with the adopted changes to include the low-resolution multi-labelling information into the learning procedure. The experimental specific details are given in Section 2 while the experiments together with the obtained classification results are discussed in Section 3. Finally, the concluding remarks of this research are given in Section 4.

1 CONVOLUTIVE EVENT MODELLING

1.1 Introduction to non-negative matrix deconvolution

Non-negative matrix deconvolution, also known as convolutive non-negative matrix factorisation (CNMF) or convolutive sparse coding (CSC), is an extension of non-negative matrix factorisation and is capable of identifying components with a temporal structure [14, 18]. The main objective of NMD is to decompose an all-positive data matrix $\mathbf{Y} \in \mathbb{R}_{\geq 0}^{B \times F}$, i.e. a B-dimensional magnitude spectrogram of length F in case of acoustic processing, into the convolution between a set of temporal basis matrices $\mathbf{A}_t \in \mathbb{R}_{\geq 0}^{B \times L}$, with $t \in [1, T]$, and an activation matrix $\mathbf{X} \in \mathbb{R}_{\geq 0}^{L \times F}$, and is formulated by

$$\mathbf{Y} \approx \mathbf{\Psi} = \sum_{t=1}^{T} \mathbf{A}_t \overrightarrow{\mathbf{X}}^{(t-1)}. \tag{1}$$

The operator $\overrightarrow{(\cdot)}^{t}$ denotes a matrix shift of t entries to the right. Columns that are shifted out at the right are discarded, while zeros are shifted in from the left. Conversely, the operation $\overleftarrow{(\cdot)}^{t}$ shifts columns to the left, with zero filling on the right and removal of the shifted out columns on the left. It is now clear how to read equation (1) and it can be noticed that NMD turns into regular NMF for the special case $T = 1$. The complete set of basis data is described by combining all temporal basis matrices $\mathbf{A}_t$ into a global three-way tensor $\mathbf{A} \in \mathbb{R}^{B \times L \times T}$. Each l-th slice of $\mathbf{A}$, denoted further as $\mathbf{A}_{(l)}$, then contains the temporal basis data of the l^{th}-component over time t and can be interpreted as one of the additive time-frequency dictionary elements describing the underlying structure in $\mathbf{Y}$. In other words, the reconstructed data $\mathbf{\Psi}$ is obtained by summing the L convolutions obtained between $\mathbf{A}_{(l)}$ and the l-th row in $\mathbf{X}$. Hence, each row in $\mathbf{X}$ can thus be seen as an activation vector for the corresponding time-frequency dictionary element $\mathbf{A}_{(l)}$.

The objective of the decomposition is to estimate $\mathbf{A}$ and $\mathbf{X}$ such that the error between $\mathbf{Y}$ and $\mathbf{\Psi}$ is minimised. Dif-

ferent error measures are proposed in the literature [19] but we have chosen to use the generalised Kullback-Leibler divergence (beta divergence for $\beta = 1$) which is given by

$$D_{KL}(\mathbf{Y}\|\mathbf{\Psi}) = \sum_{(b,f)\in(B,F)} (\mathbf{Y}_{bf} \log \frac{\mathbf{Y}_{bf}}{\mathbf{\Psi}_{bf}} - \mathbf{Y}_{bf} + \mathbf{\Psi}_{bf}). \tag{2}$$

Although NMF and NMD favour a sparse and parts-based representation of the observation data $\mathbf{Y}$ there is no guarantee for this behaviour [20]. Therefore, an additional weighted L_1-norm penalty is typically applied to the activations to control the sparseness. Hence, the total cost function is expressed as

$$\min_{\mathbf{A},\mathbf{X}} \Big(D_{KL}(\mathbf{Y}||\mathbf{\Psi}) + \lambda \|\mathbf{X}\|_1 \Big), \tag{3}$$

with $\lambda\|\mathbf{X}\|_1$ denoting the weighted sparsity penalty term. Increasing λ will tend to sparser solutions. It is worth to note that this approach of sparsity penalisation is not scale-invariant and can thus be trivially minimised by scaling $\mathbf{A}$ up and $\mathbf{X}$ down without altering $D_{KL}(\mathbf{Y}||\mathbf{\Psi})$. A solution to overcome this scaling misbehaviour is by minimising (3) under the constraint that $\mathbf{A}_{(l)}, \forall l \in L$, is normalised to unit L_2-norm as proposed in [21, 22]. The corresponding iterative multiplicative updates are

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \frac{\sum_{t=1}^{T} \mathbf{A}_t{}' \overleftarrow{\left(\frac{\mathbf{Y}}{\mathbf{\Psi}}\right)}^{(t-1)}}{\sum_{t=1}^{T} \mathbf{A}_t{}' \overleftarrow{\mathbf{1}_u}^{(t-1)} + \lambda}, \tag{4}$$

for the activations and

$$\mathbf{A}_t \leftarrow \mathbf{A}_t \otimes \frac{\left(\frac{\mathbf{Y}}{\mathbf{\Psi}}\right) \overrightarrow{\mathbf{X}}^{(t-1)'} + \mathbf{N}_t}{\mathbf{1}_u \overrightarrow{\mathbf{X}}^{(t-1)'} + \mathbf{D}_t}, \forall t \in [1,T], \tag{5}$$

for the temporal basis data with

$$\mathbf{N}_t = \tilde{\mathbf{A}}_t \otimes \left(\mathbf{1}_v \left[\tilde{\mathbf{A}}_t \otimes \left(\mathbf{1}_u \overrightarrow{\mathbf{X}}^{(t-1)'} \right) \right] \right), \tag{6}$$

and

$$\mathbf{D}_t = \tilde{\mathbf{A}}_t \otimes \left(\mathbf{1}_v \left[\tilde{\mathbf{A}}_t \otimes \left[\left(\frac{\mathbf{Y}}{\mathbf{\Psi}}\right) \overrightarrow{\mathbf{X}}^{(t-1)'} \right] \right] \right), \tag{7}$$

where $\tilde{\mathbf{A}} = \left[\frac{\mathbf{A}_{(1)}}{\|\mathbf{A}_{(1)}\|_2}, \frac{\mathbf{A}_{(2)}}{\|\mathbf{A}_{(2)}\|_2}, \ldots, \frac{\mathbf{A}_{(L)}}{\|\mathbf{A}_{(L)}\|_2} \right]$ denotes the slice-wise normalisation of $\mathbf{A}$ over time t. $\mathbf{1}_u$ and $\mathbf{1}_v$ are both all-one matrices of dimensions $B \times F$ and $B \times B$ respectively.

1.2 Event modelling using low-resolution multi-label NMD

In this work, we introduce an extended version of NMD, called low-resolution multi-label non-negative matrix deconvolution (LRM-NMD) where both the observation data and the available labelling information are used during training. The labelling information is included into the learning procedure by partitioning the observation data $\mathbf{Y}^{[o]}$ into so-called annotation segments, i.e. $\mathbf{Y}^{[o]} = [\mathbf{Y}_1^{[o]}, \mathbf{Y}_2^{[o]}, \ldots, \mathbf{Y}_J^{[o]}]$, all containing a sequence of multiple (different) and potentially overlapping sound events. The sound events in the segments are labelled on a low-resolution multi-label level by assigning a label vector $\mathbf{y}_j^{[s]} \in \{0,1\}^C$ to each $\mathbf{Y}_j^{[o]}$ indicating the sound classes, or possibly a subset of the sound classes, that take place in the corresponding segment. The time duration of the annotation segments may vary, i.e. they don't have to be of equal length, and defines the resolution of the labelling. More specifically, longer annotation segments corresponds to a lower resolution compared to shorter annotation segments. Formally, the low-resolution multi-label vectors are defined by

$$\mathbf{y}_{(c,j)}^{[s]} = \begin{cases} 1 & \text{if seg. } j \text{ contains data from sound class } c, \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where $\mathbf{y}_{(c,j)}^{[s]}$ denotes the c^{th}-element in $\mathbf{y}_j^{[s]}$. The main objective of LRM-NMD is still to decompose the observations $\mathbf{Y}_j^{[o]}, \forall j \in J$, by using (1) but with respect to

$$\mathbf{y}_j^{[s]} \approx \boldsymbol{\psi}_j^{[s]} = \mathbf{A}^{[s]} \mathbf{X}_j \mathbf{1}_j, \tag{9}$$

with $\mathbf{1}_j$ an all-one column vector of length F_j and $\mathbf{A}^{[s]}$ functioning as a labelling matrix for the temporal basis data $\mathbf{A}^{[o]}$ and is defined by

$$\mathbf{A}_{(c,l)}^{[s]} = \begin{cases} 1 & \text{if } \mathbf{A}_{(l)}^{[o]} \text{ belongs to sound class } c, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

This additional factorisation favours decompositions such that there is a resemblance with the labelling information as well. More specifically, LRM-NMD encourages that the events in annotation segment $\mathbf{Y}_j^{[o]}$, labelled by $\mathbf{y}_j^{[s]}$, are decomposed with their subset of temporal basis data given the labelling matrix $\mathbf{A}^{[s]}$ and thereby further improving the parts-based representation in $\mathbf{A}^{[o]}$. A graphical example of LRM-NMD is given in Figure 1 where events of three different sound classes in $\mathbf{Y}^{[o]}$ are decomposed into three temporal basis matrices, each four frames long, and their activations.

The cost function of LRM-NMD is expressed by

$$\min_{\mathbf{A}^{[o]},\mathbf{A}^{[s]},\mathbf{X}} \left[\sum_{j=1}^{J} \Big(D_{KL}(\mathbf{Y}_j^{[o]} \| \mathbf{\Psi}_j^{[o]}) + \lambda \|\mathbf{X}_j\|_1 + \eta D_{KL}(\mathbf{y}_j^{[s]} \| \boldsymbol{\psi}_j^{[s]}) \Big) \right], \tag{11}$$

with η as a semantic balancing parameter between the observation data and the labelling information. Larger values of η yield more focus on the labelling information but with risk of overfitting. Conversely, LRM-NMD reduces to conventional NMD when $\eta = 0$. Like for NMD, iterative update formulas for $\mathbf{X}_j$, $\forall j \in J$, $\mathbf{A}^{[o]}$ and $\mathbf{A}^{[s]}$ can be derived to minimise (11) by following the same procedure as pro-

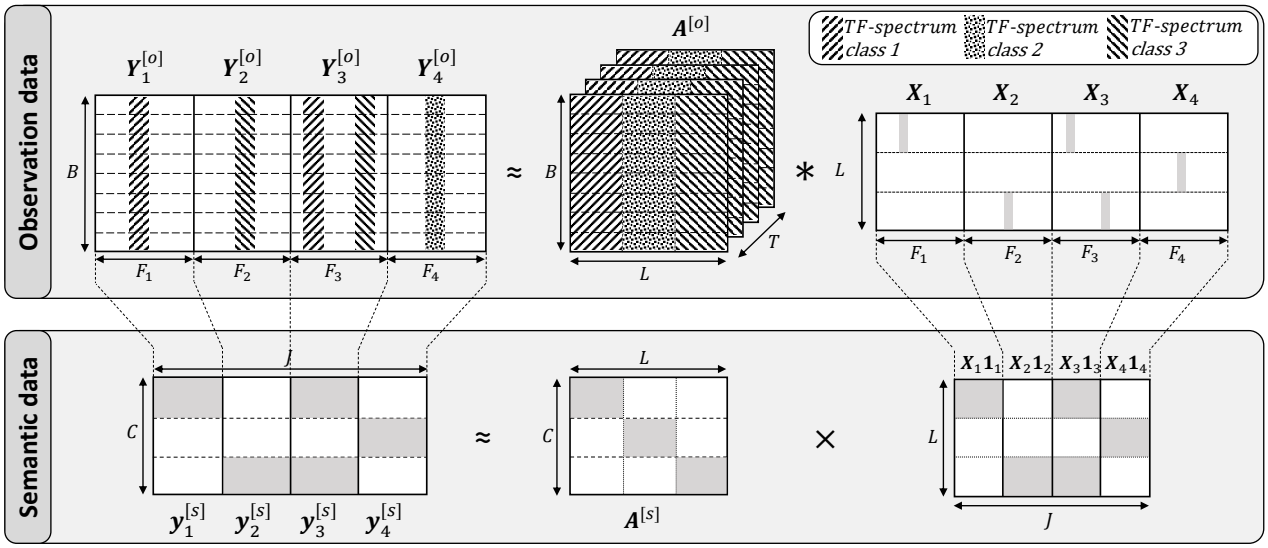

Fig. 1. A graphical example of LRM-NMD where events of three different sound classes ($C = 3$) in $\mathbf{Y}^{[o]}$ are decomposed into a set of temporal basis matrices ($L = 3$) each four frames long ($T = 4$) and their activations. The operators $*$ and $\times$ denote the convolutive factorisation according to (1) and a standard matrix multiplication respectively. The upper part visualises the convolutive modelling of events w.r.t. the low-resolution multi-label vectors given in the lower part. The learned temporal basis data $\mathbf{A}^{[o]}$ can be interpreted as a set of dictionary elements describing the underlying structure in the observation data $\mathbf{Y}^{[o]}$ while $\mathbf{A}^{[s]}$ acts as a labelling matrix of $\mathbf{A}^{[o]}$ indicating to which sound class each learned dictionary element corresponds.

posed in [18]. The corresponding updates are

$$\mathbf{X}_j \leftarrow \mathbf{X}_j \otimes \frac{\left(\sum_{t=1}^{T} \mathbf{A}_t^{[o]\prime} \big(\frac{\mathbf{Y}_j^{[o]}}{\overleftarrow{\mathbf{\Psi}_j^{[o]}}^{(t-1)}}\big)\right) + \eta\left(\mathbf{A}^{[s]\prime}\big(\frac{\mathbf{y}_j^{[s]}}{\boldsymbol{\psi}_j^{[s]}}\big)\mathbf{1}'_j\right)}{\left(\sum_{t=1}^{T} \mathbf{A}_t^{[o]\prime} \overleftarrow{\mathbf{1}_{u_j}}^{(t-1)} + \lambda\right) + \eta\left(\mathbf{A}^{[s]\prime}(\mathbf{1}_w)\mathbf{1}'_j\right)}, \tag{12}$$

$$\mathbf{A}_t^{[o]} \leftarrow \mathbf{A}_t^{[o]} \otimes \frac{\sum_{j=1}^{J}\left(\big(\frac{\mathbf{Y}_j^{[o]}}{\mathbf{\Psi}_j^{[o]}}\big) \overrightarrow{\mathbf{X}_j}^{(t-1)\prime} + \mathbf{N}_{t_j}\right)}{\sum_{j=1}^{J}\left(\mathbf{1}_{u_j} \overrightarrow{\mathbf{X}_j}^{(t-1)\prime} + \mathbf{D}_{t_j}\right)}, \forall t \in [1,T], \tag{13}$$

and

$$\mathbf{A}^{[s]} \leftarrow \mathbf{A}^{[s]} \otimes \frac{\sum_{j=1}^{J}\left(\big(\frac{\mathbf{y}_j^{[s]}}{\boldsymbol{\psi}_j^{[s]}}\big)\big(\mathbf{X}_j\mathbf{1}_j\big)'\right)}{\sum_{j=1}^{J}\left(\mathbf{1}_w\big(\mathbf{X}_j\mathbf{1}_j\big)'\right)}, \tag{14}$$

with $\mathbf{1}_w$ an all-one column vector of length C. Furthermore, it is worth noting that the additional term $D_{KL}(\mathbf{y}_j^{[s]} \| \boldsymbol{\psi}_j^{[s]})$ in the cost function of LRM-NMD does not change the update of $\mathbf{A}^{[o]}$ but that we only have rewritten the numerator and denominator into a sum over the annotation segments. The corresponding $\mathbf{N}_{t_j}$ and $\mathbf{D}_{t_j}$ are computed from $\mathbf{Y}_j^{[o]}$, $\mathbf{\Psi}_j^{[o]}$ and $\mathbf{X}_j$ using (6) and (7) respectively. The complete training procedure of LRM-NMD is given in Algorithm 1.

1.3 Finding activations from the basis data

Once a representative set of basis data is learned from the training data, the task of LRM-NMD becomes decomposing unseen observation data $\mathbf{Y}^{[o]}$ such that the estimated activations $\mathbf{X}$ minimise (3) under the fixed learned basis data $\mathbf{A}^{[o]}$. Like in training, the activations are obtained by applying iterative updates to $\mathbf{X}$ but with keeping $\mathbf{A}^{[o]}$ fixed. It is worth mentioning that the updates are done by using (4), instead of (12), since the class labels are hidden and cannot be used.

The obtained activations are well suited for the purpose of structure discovery in $\mathbf{Y}^{[o]}$ due to the parts-based representation of $\mathbf{A}^{[o]}$. Generally, the degree of activation in $\mathbf{X}$ indicates where events in $\mathbf{Y}^{[o]}$ occur while the labelling matrix $\mathbf{A}^{[s]}$ reveals the sound classes of the detected events. The complete procedure of activation estimation is given in Algorithm 2.

2 EXPERIMENTAL SPECIFIC DETAILS

2.1 From raw data to features

In this work the LRM-NMD method is evaluated in a context of monitoring persons at home based on acoustics. For this purpose the publicly available NAR-dataset was selected. This dataset contains a set of real-life isolated domestic audio events, collected with a humanoid robot Nao, and is recorded especially for acoustic classification benchmarking in domestic environments [23, 24].

In total 42 different sound classes were recorded and can be categorised into *'kitchen related events'*, *'office related events'*, *'non-verbal events'* and *'verbal events'*. The verbal events are not used in this research since we are not interested in detecting and classifying speech commands such

Algorithm 1 LRM-NMD training procedure

Inputs: $\mathbf{Y}_j^{[o]}$, $\mathbf{y}_j^{[s]}$, λ and η with $j \in [1, J]$
Initialise: $\mathbf{A}^{[o]}$, $\mathbf{A}^{[s]}$ and $\mathbf{X}_j$
Normalise: $\mathbf{A}_{(l)}^{[o]} \leftarrow \mathbf{A}_{(l)}^{[o]} / \|\mathbf{A}_{(l)}^{[o]}\|_2, \forall l \in L$
repeat
  for $j = 1 : J$ **do**
    $\boldsymbol{\Psi}_j^{[o]} = \sum_{t=1}^{T} \mathbf{A}_t^{[o]} \overrightarrow{\mathbf{X}_j^{(t-1)}}$
    $\mathbf{X}_j \leftarrow \mathbf{X}_j \otimes \frac{\left(\sum_{t=1}^{T} \mathbf{A}_t^{[o]\prime} \left(\frac{\mathbf{y}_j^{[o]}}{\boldsymbol{\Psi}_j^{[o]}} \right) \right) + \eta \left(\mathbf{A}^{[s]\prime} \left(\frac{\mathbf{y}_j^{[s]}}{\boldsymbol{\psi}_j^{[s]}} \right) \mathbf{1}'_j \right)}{\left(\sum_{t=1}^{T} \mathbf{A}_t^{[o]\prime} \overleftarrow{\mathbf{1}_{u_j}^{(t-1)}} + \lambda \right) + \eta \left(\mathbf{A}^{[s]\prime} (\mathbf{1}_w) \mathbf{1}'_j \right)}$
    $\boldsymbol{\Psi}_j^{[o]} = \sum_{t=1}^{T} \mathbf{A}_t^{[o]} \overrightarrow{\mathbf{X}_j^{(t-1)}}$
  end for
  $\mathbf{A}_t^{[o]} \leftarrow \mathbf{A}_t^{[o]} \otimes \frac{\sum_{j=1}^{J} \left(\left(\frac{\mathbf{Y}_j^{[o]}}{\boldsymbol{\Psi}_j^{[o]}} \right) \overrightarrow{\mathbf{X}_j^{(t-1)\prime}} + \mathbf{N}_{t_j} \right)}{\sum_{j=1}^{J} \left(\mathbf{1}_{u_j} \overrightarrow{\mathbf{X}_j^{(t-1)\prime}} + \mathbf{D}_{t_j} \right)}, \forall t \in [1, T]$
  $\mathbf{A}_{(l)}^{[o]} \leftarrow \mathbf{A}_{(l)}^{[o]} / \|\mathbf{A}_{(l)}^{[o]}\|_2, \forall l \in L$
  for $j = 1 : J$ **do**
    $\boldsymbol{\Psi}_j^{[o]} = \sum_{t=1}^{T} \mathbf{A}_t^{[o]} \overrightarrow{\mathbf{X}_j^{(t-1)}}$
    $\boldsymbol{\psi}_j^{[s]} = \mathbf{A}^{[s]} \mathbf{X}_j \mathbf{1}_j$
  end for
  $\mathbf{A}^{[s]} \leftarrow \mathbf{A}^{[s]} \otimes \frac{\sum_{j=1}^{J} \left(\left(\frac{\mathbf{y}_j^{[s]}}{\boldsymbol{\psi}_j^{[s]}} \right) \left(\mathbf{X}_j \mathbf{1}_j \right)' \right)}{\sum_{j=1}^{J} \left(\mathbf{1}_w \left(\mathbf{X}_j \mathbf{1}_j \right)' \right)}$
  for $j = 1 : J$ **do**
    $\boldsymbol{\psi}_j^{[s]} = \mathbf{A}^{[s]} \mathbf{X}_j \mathbf{1}_j$
  end for
until convergence **or** when max. iterations is reached
return $\mathbf{A}^{[o]}$, $\mathbf{A}^{[s]}$ and $\mathbf{X}_j, \forall j \in J$

Algorithm 2 LRM-NMD evaluation procedure

Inputs: $\mathbf{Y}^{[o]}$, $\mathbf{A}^{[o]}$, $\mathbf{A}^{[s]}$ and λ
Initialise: $\mathbf{X}$
repeat
  $\boldsymbol{\Psi}^{[o]} = \sum_{t=1}^{T} \mathbf{A}_t^{[o]} \overrightarrow{\mathbf{X}^{(t-1)}}$
  $\mathbf{X} \leftarrow \mathbf{X} \otimes \frac{\sum_{t=1}^{T} \mathbf{A}_t^{[o]\prime} \left(\frac{\mathbf{Y}^{[o]}}{\boldsymbol{\Psi}^{[o]}} \right)}{\sum_{t=1}^{T} \mathbf{A}_t^{[o]\prime} \overleftarrow{\mathbf{1}_u^{(t-1)}} + \lambda}$
  $\boldsymbol{\Psi}^{[o]} = \sum_{t=1}^{T} \mathbf{A}_t^{[o]} \overrightarrow{\mathbf{X}^{(t-1)}}$
until convergence **or** when max. iterations is reached
return $\mathbf{X}$

as *'left'*, *'right'*, *'start'*, *'stop'*, etc. Therefore, the dataset is reduced to 20 sound classes each containing 20 or 21 recordings of an isolated event. A detailed overview of the used data together with the average duration per sound class is given in Table 1.

In addition, to examine the influence of background noise we have artificially mixed the clean events with natural environmental sounds as well, i.e. recordings of *'rain'*, *'hail'* and *'blowing wind'*, which are a realistic type of background noise in case of home monitoring. The examined SNR levels are *'clean'*, 20, 10, 5, 3 and 0 dB.

A good choice of non-negative and approximately additive acoustic features are the so-called MEL-magnitude spectrograms [25]. The MEL-magnitude spectrograms spanning 40 bands are computed using a Hamming window with a frame length of 25 ms and a frame shift of 10 ms as proposed in [26]. The used filter bank is constructed such that the begin frequency of the first MEL-filter and the end frequency of the last MEL-filter correspond to the frequency range of the microphone, i.e. 300 Hz and 18 kHz. Furthermore, all events are L_2-normalised over the duration of the recording to prevent the higher energetic events from dominating those with lower energy during training.

The features for the training and test sets are obtained by randomly sampling events from each sound class in the NAR-dataset with a ratio of 60% training data and 40% test data. This implies that each class is represented with twelve training examples and eight or nine test examples resulting into class balanced training and test sets. In total four independent folds are generated to examine the statistical relevance of the results as well.

2.2 Basis initialisation

Like NMF, Algorithm 1 can only be proven to find a local optimum of its cost function, and not a global one, and is therefore sensitive as well to the initialisation of its factors [18, 19]. Careful initialisation of $\mathbf{A}^{[o]}$, $\mathbf{A}^{[s]}$ and $\mathbf{X}$ can improve the speed and accuracy of the factorisation, as it can produce faster convergence to an improved local minimum. Although random initialisation with small positive numbers is by far the most commonly used method in the literature, we have chosen an exemplar-based initialisation which has been found to be effective as well in the case of acoustic event modelling [25, 27]. Hence, the temporal basis data $\mathbf{A}^{[o]}$ is initialised with one example per sound class randomly sampled from the training data and the labelling matrix $\mathbf{A}^{[s]}$ is initialised according to (10). The silence frames before and after the events are trimmed and the shorter events are padded with small positive random values (between 0 and 10^{-4}) such that they are all of equal length. One exception is *'running tapwater'*, where the length is limited to the maximum length of the other chosen events, i.e. 40 frames in this work, due to stationarity in its feature frames. The overall dimensions of $\mathbf{A}^{[o]}$ are thus $B = 40$, $L = 20$ and $T = 40$. The activations $\mathbf{X}$ on the other hand are initialised uniformly with the value $1/L$.

Table 1. Overview NAR-dataset

Cat.	Sound class	# Events	Avg. duration in (s)
Kitchen	Alarm microwave	21	0.52 ± 0.06
	Alarm refrigerator	21	0.55 ± 0.03
	Closing microwave	21	0.37 ± 0.07
	Opening microwave	21	0.28 ± 0.05
	Drawer	21	1.04 ± 0.11
	Taking cutlery	21	0.35 ± 0.06
	Toaster	21	0.45 ± 0.10
	Running tapwater	21	1.78 ± 0.43
Non-verbal	Coughing	21	0.84 ± 0.09
	Fingersnap	20	0.23 ± 0.06
	Handclap	20	0.36 ± 0.08
	Tongue click	20	0.18 ± 0.06
Office	Doorknock	20	0.32 ± 0.06
	Locking/unlocking door	20	0.29 ± 0.08
	Opening door	20	0.29 ± 0.12
	Closing door	20	0.44 ± 0.05
	Moving chair	21	0.97 ± 0.10
	Tearing paper	20	0.46 ± 0.08
	Opening zipper 1	20	0.34 ± 0.04
	Opening zipper 2	20	0.40 ± 0.06

Note: The verbal events of the NAR-dataset are not listed in this table. See [24] for a complete overview.

2.3 Background modelling

The implementation of a background noise model is done on a straightforward manner by adding isolated noise examples to the basis data $\mathbf{A}^{[o]}$, without linking them to a specific sound class given the labelling matrix $\mathbf{A}^{[s]}$, and keeping them fixed during training. Here, only one example for each type of noise is added to $\mathbf{A}^{[o]}$ resulting into four additional dictionary elements representing the noise data, i.e. *'Nao fan noise'*, *'rain'*, *'hail'* and *'blowing wind'* respectively. As a result of the background modelling, the total number of temporal basis vectors is increased to $L = 24$. In addition, the sparsity penalty weights corresponding to the background noise basis data are set to zero and the noise activations are restricted to be constant over periods of one second to mimic the stationarity and slow time-varying properties of the background noise. This form of parameter tying can be expressed in the optimisation problem (11) and is tantamount to overwriting the rows in the numerator and denominator of (12) corresponding to the noise activations with their mean value computed from non-overlapping one second windows.

2.4 Event classification

The classification task is implemented by feeding features of isolated non-overlapping events to the LRM-NMD evaluation procedure given by algorithm (2). The obtained activations are summed over time and the predicted class label $\hat{C}$ is defined by the temporal basis matrix $\mathbf{A}_{(l)}$ with the highest accumulated activation in which the background noise related activations are disregarded. The used metric to evaluate the classification performance is classification accuracy which defines the percentage of correctly classified events.

3 EXPERIMENTS AND RESULTS

3.1 Experimental setup

The experiments in this work will mainly focus on investigating the relation between the number of labelled data w.r.t. the classification performance of LRM-NMD. This relation is examined in two different operating modes, i.e. *'single-label learning'* and *'low-resolution multi-label learning'*, and are discussed in more detail in section 3.3 and section 3.4 respectively. The main difference between both operating modes is that in single-label learning each $\mathbf{Y}_j^{[o]}$, $\forall j \in J$, is an isolated sound event, while in low-resolution multi-label learning all $\mathbf{Y}_j^{[o]}$, $\forall j \in J$, are segments spanning multiple sound events. The latter operating mode is of more interest in case of a real-life setting, e.g. home-monitoring, since it reduces the workload on labelling the training data significantly. The noise robustness of both operating modes will be examined and in case of low-resolution multi-label learning we will examine the influence of overlapping events in the annotating segments as well. Furthermore, the behaviour of the semantic balancing parameter (η) and the sparsity penalty (λ) is also investigated in detail.

3.2 Baselines

This section briefly discusses the used baselines, together with their implementation specific details, which have been used to benchmark LRM-NMD. The used baselines are chosen w.r.t. the methods described in [23, 24] and are:

1) A **Gaussian mixture model (GMM)** is a weighted sum of M-multivariate Gaussian distributions modelling the feature space $\mathbf{Y}^{[o]}$ such that the a-posteriori probability is maximised [28, 29]. During classification, the objective is to find the class model $\boldsymbol{\theta}_c$ with the highest a-posteriori probability given the features $\mathbf{Y}_{te}^{[o]}$ of an unlabelled test event using

$$\hat{C} = \underset{\forall c \in C}{\operatorname{argmax}}\, p(\boldsymbol{\theta}_c | \mathbf{Y}_{te}^{[o]}) = \underset{\forall c \in C}{\operatorname{argmax}}\, p(\mathbf{Y}_{te}^{[o]} | \boldsymbol{\theta}_c), \quad (15)$$

where the second equation is because of Bayes' rule with the assumption of equal class priors, i.e. $p(\boldsymbol{\theta}_c) = 1/C$, and the class independency of $p(\mathbf{Y}_{te}^{[o]})$.

As opposed to NMD, GMMs are typically used in combination with MEL-frequency cepstral coefficients (MFCCs) due to their ability to represent the amplitude spectrogram in a compact and

decorrelated form [29]. The MFCCs are computed from the MEL-features by applying a discrete cosine transform (DCT) on the log-transformed MEL-magnitudes. Here, 13^{th} order MFCCs are computed, without the deltas (rate of change) and delta-deltas (acceleration), and are normalised such that the complete training set has zero mean and unit standard deviation in each cepstral dimension. The number of mixture components are tuned on a separate development set obtained by further partitioning the test data into two equal sized subsets, alternately serving for development and test respectively. The global classification accuracy of the fold is computed by averaging the sub results. Furthermore, all GMMs are modelled with diagonal covariance matrices due to the limited amount of training data.

2) **Support vector machine (SVM)** is a binary classifier formally characterised by a separating hyperplane. The operation of the SVM algorithm is based on estimating the hyperplane such that the margin between the two classes is maximised. This method of construction yields that the decision function is typically specified by the points closest to the hyperplane, referred to as the support vectors. An unlabelled test vector $\mathbf{y}_{te}^{[o]}$ is classified by means of its position to the separating hyperplane using

$$\hat{C} = sign\left(\sum_{j=1}^{J} \alpha_j y_j K(\mathbf{y}_j^{[o]}, \mathbf{y}_{te}^{[o]}) + \omega \right), \qquad (16)$$

where $[\alpha_1, \alpha_2, \ldots, \alpha_J]$ and ω are the SVM model parameters, $\mathbf{y}_j^{[o]}$, $\forall j \in J$, are the training vectors with class label $y_j \in \{-1, +1\}$, and $K(\mathbf{y}, \mathbf{y}^*) = \phi(\mathbf{y})'\phi(\mathbf{y}^*)$ is the Kernel-function which can be seen as a function that describes the similarity between two feature vectors [30]. The used kernel in this work is the well-known radial basis function (RBF) and is defined by $K(\mathbf{y}, \mathbf{y}^*) = exp(-\|\mathbf{y} - \mathbf{y}^*\|_2^2 / 2\sigma^2)$ with σ as hyper-parameter defining the kernel bandwidth.

Several solutions are presented in the literature to expand this binary classification problem into a multiclass classification problem. Here, $1 - vs - 1$ is used as coding scheme, resulting into the estimation of $(1/2)C(C-1)$ SVMs discriminating one class from another. The overall classification result is computed by a majority vote over the sub results [31].

In contrast to GMM, SVM typically uses the normalised mean and standard deviation of each MFCC dimension as acoustic features instead of using them individually. The latter is done to reduce the computational complexity of SVM since it is linearly dependent on the number of training examples. This implies that each $\mathbf{Y}_j^{[o]}$, $\forall j \in J$ is transformed into a single feature vector $\mathbf{y}_j^{[o]}$ containing the mean and standard deviation of each MFCC dimension. The SVM hyper-parameters, i.e. kernel bandwidth and regularisation parameter, are tuned as with GMM by a grid search such that the classification accuracy on the development set is maximised.

3) **Exemplar NMD (E-NMD)** constructs the temporal basis data matrix $\mathbf{A}^{[o]}$ directly from the J labelled examples in the training data $\mathbf{Y}^{[o]}$ instead of learning them as with LRM-NMD. Hence, the training procedure can be omitted but with as main disadvantage that the complexity of the evaluation procedure in E-NMD is linear with the number of exemplars ($L = J$).

   In this work, an unlabelled event is classified by summing the obtained activation vectors of the exemplars corresponding to the same class and the class label $\hat{C}$ is defined by the group of exemplars with the highest accumulated activation.

4) **Supervised NMD (S-NMD)** estimates one dictionary element ($L = 1$) for each class from the isolated class data. This implies that the learning procedure must be run C times. Subsequently, all learned basis matrices are combined into a global temporal basis data matrix $\mathbf{A}^{[o]}$ representing all classes ($L = C$) .

   Furthermore, it is worth to note that the evaluation and classification procedure of S-NMD is exactly the same as with LRM-NMD.

3.3 Single-label learning mode

Single-label learning mode learns the temporal basis data matrix $\mathbf{A}^{[o]}$ from the isolated training events in $\mathbf{Y}_j^{[o]}$, $\forall j \in J$, and the relation between the amount of training data, semantic balancing parameter and sparsity penalty w.r.t. the classification performance will be examined. This experiment is in fact a fully supervised setting and is performed to compare LRM-NMD with other supervised learning algorithms such as GMM, SVM and S-NMD. Therefore, the number of training events per sound class, further denoted by n_{tr}, is artificially increased in the sequence of $n_{tr} \in \{1, 2, 3, 4, 5, 10, all\}$ and the examined semantic balancing parameter values and sparsity penalisation weights are $\eta \in \{0, 0.1, 1, 10, 100, 200, 500, 1000\}$ and $\lambda \in \{0, 5, 10\}$ respectively.

All classification results are listed in Table 2 in Appendix A.1 and indicate that the additional weighted sparsity penalty on the activations improves the classification performance of LRM-NMD. However, too much penalty tends to lower classification scores as a result of a sub-optimal factorisation due to the sparseness in $\mathbf{X}$. In this work, the best scores are obtained when $\lambda = 5$, which are also shown in Figure 2, and indicate that both the amount of training data and the semantic balancing parameter have a significant impact on the classification performance of LRM-NMD.

First of all, including the labelling information $\mathbf{y}_j^{[s]}$, $\forall j \in J$, into the learning procedure of LRM-NMD, by setting $\eta > 0$, boosts the classification performance for all choices of n_{tr} as a result of the improved parts-based representation in $\mathbf{A}^{[o]}$. The highest classification accuracies for most n_{tr} settings are achieved for $\eta = 100$, which are situated between 81% and 94%, and remains more or less constant

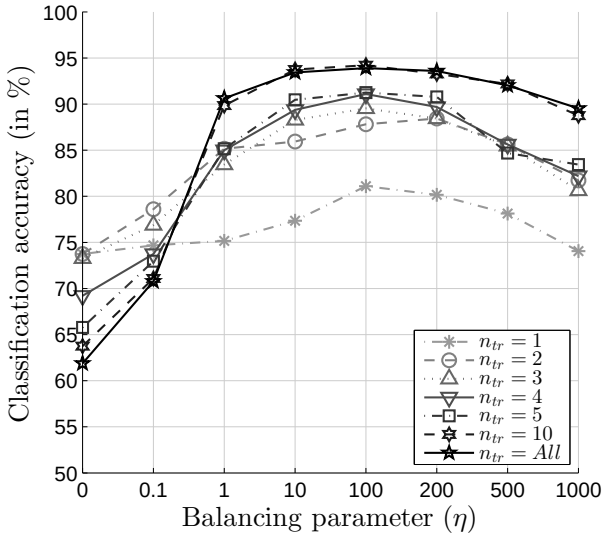

Fig. 2. Influence of the semantic balancing parameter (η) on the classification performance of LRM-NMD in single-label learning mode for different number of training events per sound class (n_{tr}) when $\lambda = 5$.

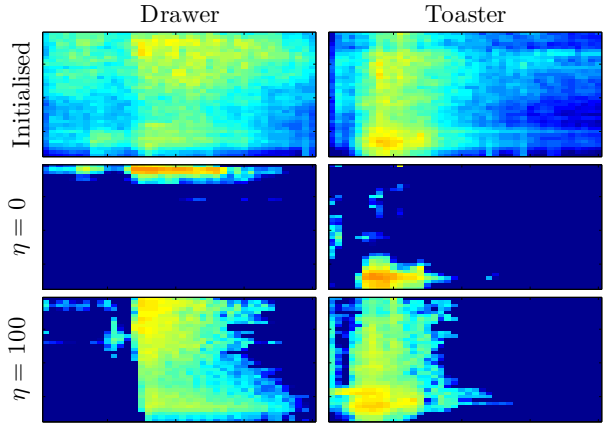

Fig. 3. A graphical representation of the temporal basis data for the sound classes *'drawer'* and *'toaster'* for $\lambda = 5$ and $n_{tr} = all$. The top row are the initialised basis data matrices while the next two rows are the obtained dictionary elements after learning when the semantic balancing parameter is set to $\eta = 0$ and $\eta = 100$ respectively. As one can see, including the labelling information, by setting $\eta > 0$, tends to model complete events and thereby improving the parts-based representation of $\mathbf{A}^{[o]}$.

in the range of $\eta \in [10, 200]$. This implies that the choice of η is not critical in this operating mode. However, larger values of η yield decreasing classification scores as a result of overfitting due to too much emphasis on correct factorisation of the labelling information instead of the acoustic event data. Secondly, increasing the number of training examples per sound class boosts the classification scores as well because of an improved generalisation of the learned temporal basis data $\mathbf{A}^{[o]}$. However, it should be noticed this improvement is only realised when the labelling information is used during training, by setting the semantic balancing parameter sufficiently large, i.e. $\eta \geq 1$. For smaller values of η, increasing the amount of training data yields more complex factorisations due to increased intra-class variabilities with as a result that some sound classes are modelled by two or more dictionary elements while other sound classes are not represented as all. This happens because the number of dictionary elements usable for modelling the acoustic event data in the basis data matrix is equal to the number of sound classes since $\mathbf{A}^{[o]}$ is initialised with one example per sound class.

The aspect of the improved parts-based representation in $\mathbf{A}^{[o]}$ is visualised in Figure 3 where the initialised temporal basis data of *'drawer'* and *'toaster'* are shown with the learned temporal basis data for a semantic balancing parameter of $\eta = 0$ and $\eta = 100$ when all training data is used and a sparsity penalty of $\lambda = 5$. By comparing the initialised data, i.e. first row, with the learned temporal basis data for $\eta = 0$ and $\eta = 100$, i.e. second and third row respectively, it can be clearly seen that the semantic balancing parameter affects the parts-based representation of $\mathbf{A}^{[o]}$ since $\eta = 0$ only models the spectro-temporal regions with the highest energy while the setting $\eta = 100$ tends to model complete events.

By comparing the classification scores of LRM-NMD when $\eta = 100$ and $\lambda = 5$ with the baseline results in function of the amount of training data, which are shown in Figure 4, it can be clearly seen that LRM-NMD outperforms the baselines, with exception of S-NMD, when the available number of training events per sound class is rather limited, i.e. $n_{tr} = 5$ or less. However, the main advantage of LRM-NMD over S-NMD is that it can be used in a low-resolution multi-label learning mode, i.e. larger segments are labelled by a single multi-label vector indicating sound classes of the events, or possibly a subset of events, that take place during that period, and thereby reducing the workload on labelling the training data significantly. The sparsity penalty for S-NMD and E-NMD is set to $\lambda = 5$ as well. The highest classification scores when all training data is used are obtained by GMM followed by SVM, S-NMD and LRM-NMD respectively. Furthermore, learning the temporal basis data $\mathbf{A}^{[o]}$ from the training data is preferred over an exemplar based approach since the E-NMD results are always exceeded by both LRM-NMD and S-NMD for all n_{tr} settings.

The robustness to environmental background noise of LRM-NMD when $\eta = 100$ and $\lambda = 5$ is shown in Figure 5. These results clearly indicate that LRM-NMD can deal with noise levels down to 10 dB SNR without any significant loss in classification performance. Lower SNRs yield decreasing classification scores, however the absolute reduction in classification performance from 10 dB to 0 dB strongly depends on the amount of training data, i.e. less than 10% in case of $n_{tr} = \{10, all\}$ and somewhere around 20% to 25% in case of $n_{tr} = \{1, 2\}$. The parameter settings of $n_{tr} = \{3, 4, 5\}$ achieve still promising classification scores for a SNR of 0 dB and are situated in the range of 75% and 80% respectively.

3.4 Low-resolution multi-label learning mode

Low-resolution multi-label learning mode learns the temporal basis data $\mathbf{A}^{[o]}$ from segments spanning several

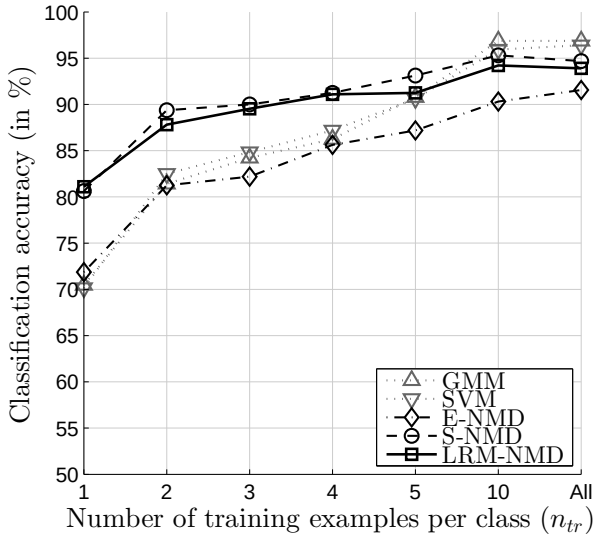

Fig. 4. Classification scores of LRM-NMD in single-label learning mode for $\eta = 100$ and $\lambda = 5$ together with the obtained baseline results in function of different number of training examples per sound class (n_{tr}).

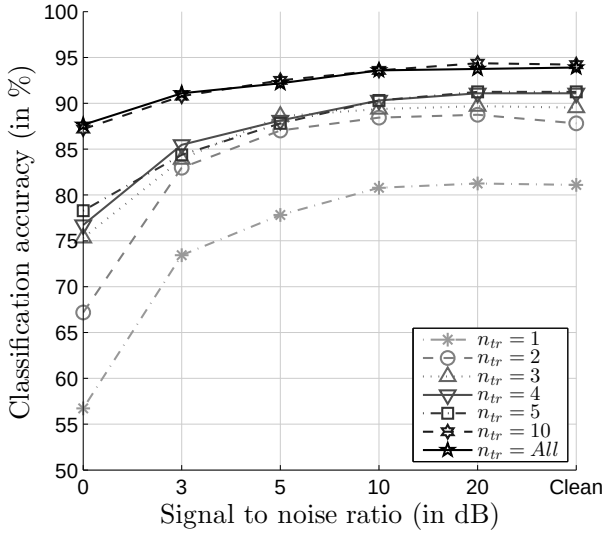

Fig. 5. Background noise robustness of LRM-NMD in single-label learning mode when $\eta = 100$ and $\lambda = 5$ in function the signal to noise ratio (SNR) for different number of training examples per sound class (n_{tr}).

events instead of the isolated events data as with single label learning. Hence, each $\mathbf{Y}_j^{[o]}$, $\forall j \in J$, is a annotating segment labelled by $\mathbf{y}_j^{[s]}$ indicating the sound classes, or a subset of sound classes, of the events that take place. This approach of annotating the data reduces the labelling workload significantly since we do not need to segment the data on the event level by indicating the onset and offset times. In this work, each annotating segment is made up of five events by randomly combining events from the training set. The events in the annotation segments are separated by T frames containing small positive random values (between 0 and 10^{-4}) to mimic short periods of silence between two successive events. In total 48 annotating segments are created per fold since each fold is represented by 240 isolated training examples which are allowed to be used once. The minimum and maximum time duration of the annotation segments varies in the range of 1.86 and 6.41 seconds respectively. In this experiment, the influence of the semantic balancing parameter (η) together with the number of labelled events per segment, further denoted by n_{lbl}, will be examined. For instance, $n_{lbl} = 1$ yield only one labelled event out of five events in each segment. Furthermore, the events in the annotating segments are sampled in such way that they all have a different sound class combination and that the amount of labelled data per sound class is balanced for all choices of $n_{lbl} \in [1,5]$. In addition, the influence of overlapping events in the annotating segments on the classification performance of LRM-NMD will be investigated as well and in the range of $n_{overlap} \in \{0\%, 25\%, 50\%, 75\%, 100\%\}$. Important to note is that the events in the segments are sorted with increasing duration and that $n_{overlap}$ is defined as the percentage of overlapping frames between two successive events computed on basis of the shortest one. Special cases are $n_{overlap} = 100\%$ where all five events in each segment have the same onset time and $n_{overlap} = 0\%$ where the events are still separated with T frames containing small positive random values.

All classification results on the clean and non-overlapping data are listed in Table 3 in Appendix A.2.1 and the best classification scores are again when the sparsity penalty is set to $\lambda = 5$. The corresponding results are shown in Figure 6 and indicate that the optimal choice of semantic balancing parameter in this learning mode depends on the number of labelled events per annotating segment since increasing η tends to earlier classification performance breakoffs for lower n_{lbl} settings. The latter is a result of the penalty induced by the semantic balancing parameter η on modelling the non-labelled event data in the annotating segments $\mathbf{Y}_j^{[o]}$, using their temporal basis matrices in $\mathbf{A}^{[o]}$, which leads to an increased divergence between $\mathbf{y}_j^{[s]}$ and $\boldsymbol{\psi}_j^{[s]}$. In this work, a good choice of semantic balancing parameter is $\eta = 5$ as it achieves high classification scores for all n_{lbl} settings, i.e. around 74% for $n_{lbl} = 1$ and 90% for $n_{lbl} = 5$, and will thus be used during the next experiments.

By comparing the classification scores of LRM-NMD for $\eta = 5$ and $\lambda = 5$ with the results obtained by GMM, SVM and S-NMD in function of n_{lbl}, which are shown in Figure 7, it can be clearly seen that LRM-NMD is capable of dealing with low-resolution multi-label training data. It is worth to note that the sparsity penalty for S-NMD is also set to $\lambda = 5$ and that E-NMD is not evaluated in this learning mode since the isolated exemplars are not available. Furthermore, the sound class model parameters of GMM, SVM and S-NMD are estimated from the annotating segments containing at least one labelled example of the corresponding sound class. This implies that the training data contains acoustic information of other sound classes as well and thereby resulting into less accurate model parameter estimation which explains the decreasing classification accuracies. In addition, it can be seen that GMM is also

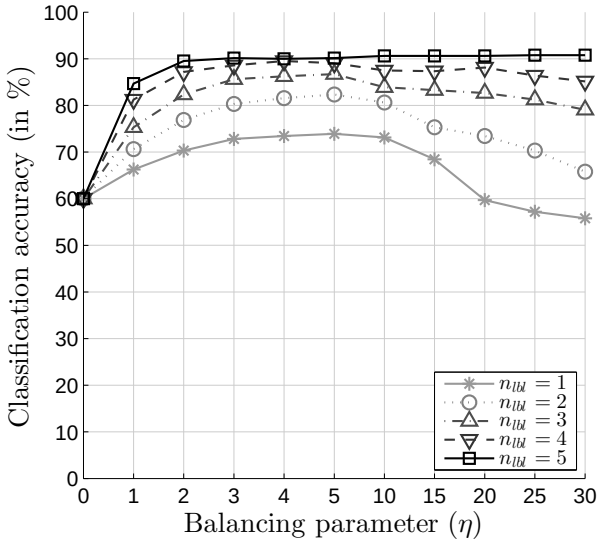

Fig. 6. Influence of the semantic balancing parameter (η) on the classification performance of LRM-NMD in low-resolution multi-label learning mode in function of different number of labelled events per annotating segment (n_{lbl}) when $\lambda = 5$.

able to deal with low-resolution multi-label training data, as a result of the ability to model different regions in the feature space and thereby distinguishing the different event classes, but that the classification performance is strongly related to the number of labelled examples.

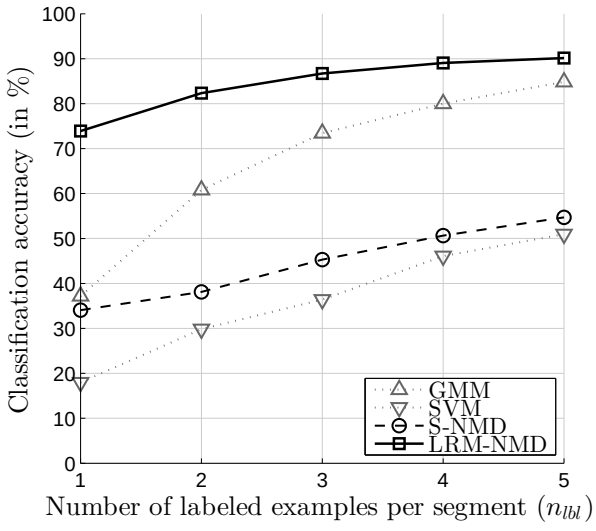

Fig. 7. Classification scores of LRM-NMD in low-resolution multi-label learning mode for $\eta = 5$ and $\lambda = 5$ together with the obtained baseline results in function of different number of labelled events per annotating segment (n_{lbl}).

Figure 8 are the classification results of LRM-NMD in noisy conditions and indicate that, as with single-label learning, LRM-NMD in low-resolution multi-label learning mode can deal with noise levels up to 10 dB without any significant loss in classification performance. Lowering the SNR further decreases classification scores. The absolute reduction in classification performance from 10 dB to 0 dB is situated somewhere between 30% and 35% for all n_{lbl} settings. This implies that the number of labelled events per segment has a limited influence on the noise robustness of LRM-NMD since the amount of training data remains the same for all n_{lbl} settings in this operating mode. Furthermore, it can be observed that the highest drop in classification performance is when the SNR is lowered from 3 dB till 0 dB and that an SNR of 3 dB achieves still classification scores in the range of 70% to 85% as long as $n_{lbl} \geq 2$.

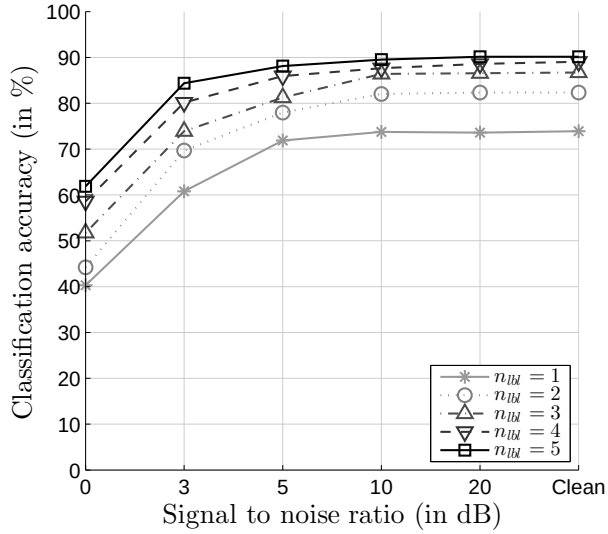

Fig. 8. Background noise robustness of LRM-NMD in low-resolution multi-label learning mode when $\eta = 5$ and $\lambda = 5$ in function of the signal to noise ratio (SNR) for different number of labelled events per annotating segment (n_{lbl}).

The influence of overlapping events in the annotating segments on the classification performance of LRM-NMD is shown in Figure 9 where the obtained classification results on the clean data are visualised in function of $n_{overlap}$. As expected, increasing $n_{overlap}$ yields decreasing classification scores, however the drop in classification performance from $n_{overlap} = 0\%$ to $n_{overlap} = 50\%$ is rather limited for all n_{lbl} settings and is situated around 5%. Increasing the amount of overlap further results in higher decreasing classification scores. The classification results on the noisy data in function of $n_{overlap}$ are all listed in Table 4 in Appendix A.2.2 and indicate the same trends as with the clean data.

4 CONCLUSION

This paper focusses on modelling sound events from acoustic data labelled on a low-resolution and multi-label level using non-negative matrix deconvolution. The low-resolution multi-labelling information simply indicates the sound classes of the events that take place over a longer period of time in the acoustic data without identifying beginning nor endings of the individual events. This approach of labelling the data requires less annotation work compared to protocols where the data must be labelled on an event level. In order to cope with this type of labelling, we have adopted the existing NMD algorithm into a low-resolution

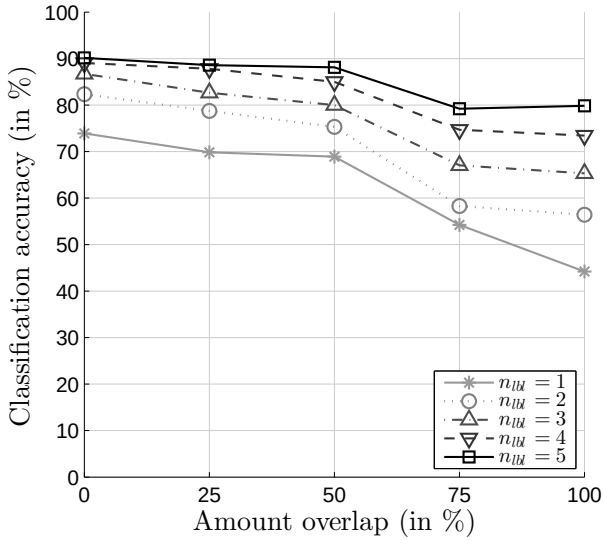

Fig. 9. Classification scores of LRM-NMD in low-resolution multi-label learning mode for $\eta = 5$ and $\lambda = 5$ in function of the amount of overlap within the training events ($n_{overlap}$) on clean data for different number of labelled events per annotating segment (n_{lbl}).

multi-label variant named LRM-NMD by incorporating the low-resolution multi-labelling information into the learning procedure of NMD.

Firstly, we have examined LRM-NMD in a single-label learning context, i.e. all events are labelled individually, to benchmark our proposed algorithm with other already widely examined sound event classifiers such as GMM, SVM, S-NMD and E-NMD [23, 24, 32, 33]. The obtained classification results show that LRM-NMD, together with S-NMD, outperforms GMM, SVM and E-NMD in situations when the amount of training examples per sound class is rather limited, i.e. $n_{tr} < 5$, and are situated in the range of 81% and 92% when $n_{tr} = 1$ and $n_{tr} = 4$ respectively. However, the main advantage of LRM-NMD over S-NMD is that it can be used in a low-resolution multi-label setting which is of more interest in case of real-life settings. In addition, LRM-NMD is also able to deal directly with overlapping event data without any further modifications to the algorithm due to its additive property, which is an interesting property in real-life settings as well, while this is not the case for GMM and SVM. It is also shown that learning the dictionary elements from the acoustic data is preferred over an exemplar based approach since E-NMD is always outperformed by both LRM-NMD and S-NMD. Furthermore, the robustness of LRM-NMD to stationary environmental background noise has been shown till SNRs of 10 dB without any significant loss in classification performance. Lowering the SNRs further starts decreasing classification scores, but the absolute reduction in classification performance is strongly related to the number of training examples per sound class.

Secondly, we have shown that LRM-NMD is capable of modelling sound events directly from acoustic data labelled on a low-resolution and multi-label level. In this work, the annotation segments are all constructed out of five events and the relation between the number of labelled events per segment (n_{lbl}) w.r.t. the classification performance of LRM-NMD is examined. The corresponding classification accuracies are situated in the range of 75% and 90% when $n_{lbl} = 1$ and $n_{lbl} = 5$ respectively and outperform the baselines. Although GMM is able to achieve a classification accuracy slightly below LRM-NMD when all events in the annotation segments are labelled, it rapidly decreases to an accuracy of less than 40% when $n_{lbl} = 1$ and thereby making it less usable compared to LRM-NMD. The robustness to background noise has been shown again with SNRs of down to 10 dB without any significant decrease in classification performance. We have also shown that LRM-NMD can deal with overlapping events in the annotating segments as well, i.e. a drop in absolute classification performance of somewhere around 5% when $n_{overlap} = 50\%$. Further increasing $n_{overlap}$ yields larger declines in classification scores as a result of a higher complexity in the acoustic observation data.

Further research will mainly focus on evaluating LRM-NMD on a real-life dataset recorded over a longer period of time for the purpose of monitoring elderly at home. Therefore, the relation between individual sound events and activities of daily living (ADL) must be investigated. A possible solution is by using the distributions of the sound classes occurring in each activity. For instance, the sound of running tap water together with the sound of a tooth brush is more likely related to the activity brushing teeth than for example cooking. Other interesting research topics that might further improve the performance of LRM-NMD are automatic relevance determination to select the correct number of dictionary elements in the learned basis data matrix $\mathbf{A}^{[o]}$ [34] and incremental or adaptive learning strategies for fine tuning $\mathbf{A}^{[o]}$ to current situation of the elderly being monitored [35].

5 ACKNOWLEDGMENT

This work was performed in the context of following projects: VLAIO doctoral scholarship (contract 121565) and Sound INterfacing through the Swarm - SINS (VLAIO-SBO contract 130006).

6 REFERENCES

[1] D. Wang and G. J. Brown, *Computational auditory scene analysis: principles, algorithms, and applications*. Wiley-IEEE Press, October 2006.

[2] A. Temko, R. Malkin, C. Zieger, D. Macho, and C. Nadeu, "Acoustic event detection and classification in smart-room environments: Evaluation of chil project systems," *Cough*, vol. 65, no. 48, p. 5, 2006.

[3] G. Roma, J. Janer, S. Kersten, M. Schirosa, P. Herrera, and X. Serra, "Ecological acoustics perspective for content-based retrieval of environmental sounds," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 1, p. 960863, 2010.

[4] J. Schroeder, S. Wabnik, P. W. J. van Hengel, and S. Goetze, *Ambient Assisted Living*, vol. 4, ch. Detection

and Classification of Acoustic Events for In-Home Care, pp. 181–195. Berlin, Heidelberg: Springer, January 2011.

[5] B. G. Ferguson and K. W. Lo, "Acoustic cueing for surveillance and security applications," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, vol. 6201, pp. 62011K–62011K–7, May 2006.

[6] S. H. Gage, W. Joo, E. Kasten, and S. Biswas, "Acoustic observations in agricultural landscapess," in *The Ecology of Agricultural Landscapes: Long-Term Research on the Path to Sustainability* (S. K. Hamilton, J. E. Doll, and G. P. Robertso, eds.), pp. 360–377, Oxford University Press, New York, USA, 2015.

[7] J. Schroeder, M. Brandes, D. Hollosi, J. Wellmann, M. Wittorf, O. Jung, V. Grtzmacher, and S. Goetze, "Foreign object detection in tires by acoustic event detection," in *Deutsche Jahrestagung fr Akustik (DAGA)*, vol. 41, (Nrnberg), pp. 1266–1269, March 2015.

[8] P. J. Moreno and S. Agarwal, "An experimental study of em-based algorithms for semi-supervised learning in audio classification," in *International Conference on Machine Learning (ICML) Workshop on the Continuum from Labeled to Unlabeled data*, August 2003.

[9] A. Diment, T. Heittola, and T. Virtanen, "Semi-supervised learning for musical instrument recognition," in *21st European Signal Processing Conference (EUSIPCO 2013)*, pp. 1–5, September 2013.

[10] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?," *Journal of Machine Learning Research*, vol. 11, pp. 625–660, March 2010.

[11] D. Hakkani Tur, G. Riccardi, and A. Gorin, "Active learning for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, pp. 3904–3907, May 2002.

[12] G. Riccardi and D. Hakkani Tur, "Active learning: Theory and applications to automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 504–511, July 2005.

[13] J. F. Gemmeke, B. Ons, N. Tessema, H. Van hamme, J. van de Loo, G. De Pauw, W. Daelemans, J. Huyghe, J. Derboven, L. Vuegen, B. Van Den Broeck, P. Karsmakers, and B. Vanrumste, "Self-taught assistive vocal interfaces: an overview of the ALADIN project," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, pp. 2039–2043, 2013.

[14] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *International Conference on Independent Component Analysis and Signal Separation*, pp. 494–499, Springer, 2004.

[15] S. Adavanne and T. Virtanen, "Sound event detection using weakly labeled dataset with stacked convolutional and recurrent neural network," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, pp. 12–16, November 2017.

[16] D. Lee, S. Lee, Y. Han, and K. Lee, "Ensemble of convolutional neural networks for weakly-supervised sound event detection using multiple scale input," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, pp. 74–79, November 2017.

[17] E. Cakir and T. Virtanen, "Convolutional recurrent neural networks for rare sound event detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, pp. 27–31, November 2017.

[18] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[19] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems (NIPS)* (T. K. Leen, T. G. Dietterich, and V. Tresp, eds.), vol. 4, pp. 556–562, MIT Press, 2000.

[20] P. O. Hoyer, "Non-negative sparse coding," in *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing*, pp. 557–565, 2002.

[21] J. Le Roux, F. J. Weninger, and J. R. Hershey, "Sparse nmf - half-baked or well done?," Tech. Rep. TR2015-023, Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA, March 2015.

[22] P. D. O'Grady and B. A. Pearlmutter, "Discovering convolutive speech phones using sparseness and non-negativity constraints," in *Proceedings of the Seventh International Conference on Independent Component Analysis*, pp. 520–527, 2007.

[23] M. Janvier, X. Alameda-Pineda, L. Girin, and R. Horaud, "Sound-event recognition with a companion humanoid," in *Humanoids 2012 - IEEE International Conference on Humanoid Robotics*, (Osaka, Japan), pp. 104–111, IEEE, November 2012.

[24] M. Janvier, X. Alameda-Pineda, L. Girin, and R. Horaud, "Sound representation and classification benchmark for domestic robots," in *2014 IEEE International Conference on Robotics and Automation (ICRA 2014)*, (Hong-Kong, China), pp. 104–111, IEEE, May 2014.

[25] J. F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste, and H. Van hamme, "An exemplar-based NMF approach to audio event detection," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–4, October 2013.

[26] B. Gold, N. Morgan, and D. Ellis, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. New York, NY, USA: John Wiley & Sons, Inc., 2nd ed., 1999.

[27] A. N. Langville, C. D. Meyer, and R. Albright, "Initializations for the nonnegative matrix factorization," 2006.

[28] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[29] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72–83, January 1995.

[30] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, COLT '92, pp. 144–152, ACM Press, 1992.

[31] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–27, April 2011.

[32] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, pp. 1733–1746, October 2015.

[33] T. Virtanen, A. Mesaros, T. Heittola, M. Plumbley, P. Foster, E. Benetos, and M. Lagrange, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*. Tampere University of Technology. Department of Signal Processing, 2016.

[34] V. Renkens and H. Van hamme, "Automatic relevance determination for nonnegative dictionary learning in the gamma-poisson model," *Signal Processing*, vol. 132, pp. 121–133, 2017.

[35] J. Driesen and H. Van hamme, "Modelling vocabulary acquisition, adaptation and generalization in infants using adaptive bayesian plsa," *Neurocomput.*, vol. 74, pp. 1874–1882, May 2011.

APPENDIX

A.1 Single-label learning mode classification results

Table 2. Classification scores of LRM-NMD in single-label learning mode on the clean data in function of the number of training examples per sound class (n_{tr}), semantic balancing parameter (η) and sparsity penalisation (λ).

λ	η	**Number of training examples per sound class** (n_{tr})						
		1	**2**	**3**	**4**	**5**	**10**	**All**
0	**0**	$49.1 \pm 2.4\%$	$45.5 \pm 1.6\%$	$41.4 \pm 5.1\%$	$39.4 \pm 3.2\%$	$37.3 \pm 3.4\%$	$37.3 \pm 2.5\%$	$35.6 \pm 2.0\%$
	0.1	$50.8 \pm 3.4\%$	$48.6 \pm 3.2\%$	$47.0 \pm 3.2\%$	$43.9 \pm 4.3\%$	$42.2 \pm 4.3\%$	$40.9 \pm 3.3\%$	$40.2 \pm 2.8\%$
	1	$57.0 \pm 4.9\%$	$59.7 \pm 3.6\%$	$60.0 \pm 4.7\%$	$59.2 \pm 5.3\%$	$59.7 \pm 6.6\%$	$63.9 \pm 3.1\%$	$63.9 \pm 0.8\%$
	10	$57.8 \pm 5.6\%$	$61.3 \pm 2.4\%$	$63.1 \pm 3.3\%$	$63.8 \pm 7.0\%$	$65.8 \pm 7.4\%$	$72.7 \pm 2.4\%$	$71.7 \pm 2.5\%$
	100	$58.8 \pm 5.5\%$	$62.3 \pm 2.9\%$	$64.4 \pm 2.8\%$	$63.8 \pm 4.4\%$	$65.3 \pm 4.6\%$	$72.0 \pm 1.6\%$	$72.3 \pm 1.7\%$
	200	$58.4 \pm 6.5\%$	$63.4 \pm 3.2\%$	$63.9 \pm 3.9\%$	$63.4 \pm 3.6\%$	$65.5 \pm 6.1\%$	$72.5 \pm 2.0\%$	$72.2 \pm 1.9\%$
	500	$58.4 \pm 7.4\%$	$63.4 \pm 3.5\%$	$64.8 \pm 4.8\%$	$64.7 \pm 4.5\%$	$66.1 \pm 4.2\%$	$71.7 \pm 1.9\%$	$73.1 \pm 1.5\%$
	1000	$59.7 \pm 8.3\%$	$64.5 \pm 2.5\%$	$65.8 \pm 3.9\%$	$65.5 \pm 3.2\%$	$66.9 \pm 3.3\%$	$72.2 \pm 1.7\%$	$72.7 \pm 1.5\%$
5	**0**	$73.8 \pm 6.2\%$	$73.8 \pm 3.1\%$	$73.3 \pm 3.7\%$	$69.2 \pm 4.9\%$	$65.8 \pm 5.7\%$	$63.8 \pm 2.1\%$	$61.9 \pm 2.2\%$
	0.1	$74.7 \pm 6.9\%$	$78.6 \pm 5.9\%$	$76.9 \pm 1.8\%$	$73.8 \pm 3.8\%$	$73.0 \pm 4.6\%$	$71.1 \pm 3.1\%$	$70.8 \pm 2.4\%$
	1	$75.2 \pm 6.4\%$	$85.2 \pm 4.2\%$	$83.4 \pm 5.5\%$	$80.5 \pm 5.1\%$	$85.2 \pm 3.8\%$	$89.8 \pm 2.0\%$	$90.6 \pm 2.2\%$
	10	$77.3 \pm 6.9\%$	$85.9 \pm 4.2\%$	$88.3 \pm 1.7\%$	$89.4 \pm 2.2\%$	$90.5 \pm 3.3\%$	$93.8 \pm 0.9\%$	$93.4 \pm 0.6\%$
	100	$\underline{81.1 \pm 4.2\%}$	$87.8 \pm 3.4\%$	$\underline{89.5 \pm 3.5\%}$	$\underline{91.1 \pm 2.2\%}$	$\underline{91.3 \pm 2.2\%}$	$\underline{94.2 \pm 1.1\%}$	$\underline{93.9 \pm 1.2\%}$
	200	$80.2 \pm 3.1\%$	$\underline{88.4 \pm 1.9\%}$	$88.4 \pm 3.6\%$	$89.7 \pm 2.3\%$	$90.8 \pm 2.4\%$	$93.3 \pm 2.1\%$	$93.6 \pm 1.6\%$
	500	$78.1 \pm 3.6\%$	$85.6 \pm 2.7\%$	$85.5 \pm 4.0\%$	$85.6 \pm 2.6\%$	$84.7 \pm 4.5\%$	$92.2 \pm 3.6\%$	$92.0 \pm 2.7\%$
	1000	$74.1 \pm 4.8\%$	$81.7 \pm 2.8\%$	$80.6 \pm 3.3\%$	$82.2 \pm 3.6\%$	$83.4 \pm 5.2\%$	$88.8 \pm 4.5\%$	$89.5 \pm 3.6\%$
10	**0**	$71.4 \pm 5.3\%$	$69.5 \pm 3.2\%$	$67.7 \pm 3.8\%$	$63.9 \pm 6.1\%$	$61.1 \pm 4.8\%$	$58.9 \pm 4.3\%$	$55.5 \pm 2.1\%$
	0.1	$72.5 \pm 6.7\%$	$74.4 \pm 5.1\%$	$70.3 \pm 2.9\%$	$67.5 \pm 5.2\%$	$67.3 \pm 5.6\%$	$64.5 \pm 5.5\%$	$63.8 \pm 5.5\%$
	1	$75.9 \pm 7.3\%$	$83.9 \pm 3.5\%$	$83.8 \pm 3.1\%$	$83.1 \pm 2.2\%$	$81.6 \pm 4.5\%$	$87.8 \pm 2.8\%$	$88.3 \pm 3.1\%$
	10	$75.9 \pm 6.9\%$	$85.6 \pm 4.4\%$	$88.4 \pm 2.5\%$	$87.8 \pm 2.8\%$	$88.1 \pm 3.2\%$	$92.8 \pm 1.9\%$	$93.6 \pm 1.4\%$
	100	$77.7 \pm 4.5\%$	$85.0 \pm 3.5\%$	$85.5 \pm 3.9\%$	$84.2 \pm 1.4\%$	$84.5 \pm 2.2\%$	$90.6 \pm 0.7\%$	$90.0 \pm 1.4\%$
	200	$73.0 \pm 5.3\%$	$84.1 \pm 3.2\%$	$83.3 \pm 3.9\%$	$82.0 \pm 2.2\%$	$84.8 \pm 1.8\%$	$89.7 \pm 1.1\%$	$89.4 \pm 1.4\%$
	500	$69.8 \pm 6.4\%$	$78.0 \pm 2.8\%$	$75.6 \pm 3.9\%$	$73.9 \pm 2.4\%$	$76.9 \pm 4.0\%$	$86.4 \pm 3.4\%$	$86.6 \pm 1.9\%$
	1000	$66.1 \pm 5.3\%$	$72.0 \pm 3.2\%$	$68.8 \pm 2.8\%$	$67.7 \pm 2.7\%$	$69.7 \pm 6.1\%$	$80.5 \pm 3.2\%$	$80.6 \pm 3.2\%$

Note: The best classification scores for each n_{tr} setting are underlined.

A.2 Low-resolution multi-label learning mode classification results

A.2.1 Non-overlapping events in the annotating segments

Table 3. Classification scores of LRM-NMD in low-resolution multi-label learning mode on the clean data in function of the number of labelled events per annotating segment (n_{lbl}), semantic balancing parameter (η) and sparsity penalisation (λ).

λ	η	**Number of labelled events per annotating segment** (n_{lbl})				
		1	**2**	**3**	**4**	**5**
0	**0**	$34.8 \pm 3.3\%$	$34.8 \pm 3.3\%$	$34.8 \pm 3.3\%$	$34.8 \pm 3.3\%$	$34.8 \pm 3.3\%$
	1	$37.7 \pm 2.7\%$	$42.3 \pm 3.1\%$	$46.3 \pm 2.7\%$	$49.4 \pm 4.5\%$	$51.4 \pm 4.1\%$
	2	$41.1 \pm 1.5\%$	$47.5 \pm 2.3\%$	$53.6 \pm 2.8\%$	$55.0 \pm 3.7\%$	$59.5 \pm 2.7\%$
	3	$42.3 \pm 1.5\%$	$48.4 \pm 1.1\%$	$56.1 \pm 3.5\%$	$59.2 \pm 2.6\%$	$64.7 \pm 3.3\%$
	4	$42.8 \pm 2.3\%$	$50.3 \pm 2.1\%$	$57.8 \pm 4.3\%$	$61.3 \pm 2.1\%$	$66.7 \pm 3.3\%$
	5	$42.7 \pm 2.1\%$	$53.1 \pm 3.1\%$	$58.9 \pm 3.1\%$	$62.7 \pm 2.8\%$	$69.7 \pm 3.4\%$
	10	$45.2 \pm 1.9\%$	$51.7 \pm 6.7\%$	$59.4 \pm 2.0\%$	$65.5 \pm 1.9\%$	$75.2 \pm 3.7\%$
	15	$45.2 \pm 4.9\%$	$47.8 \pm 9.0\%$	$59.5 \pm 4.4\%$	$68.0 \pm 2.0\%$	$77.3 \pm 3.9\%$
	20	$41.7 \pm 3.3\%$	$44.1 \pm 9.5\%$	$59.2 \pm 7.5\%$	$69.7 \pm 2.8\%$	$78.0 \pm 4.8\%$
	25	$38.0 \pm 4.9\%$	$44.1 \pm 8.1\%$	$58.6 \pm 7.3\%$	$70.0 \pm 3.3\%$	$78.3 \pm 4.6\%$
	30	$37.2 \pm 6.1\%$	$44.1 \pm 8.4\%$	$58.4 \pm 7.6\%$	$69.2 \pm 3.1\%$	$78.3 \pm 5.1\%$
5	**0**	$60.0 \pm 3.7\%$	$60.0 \pm 3.7\%$	$60.0 \pm 3.7\%$	$60.0 \pm 3.7\%$	$60.0 \pm 3.7\%$
	1	$66.3 \pm 3.6\%$	$70.6 \pm 4.2\%$	$75.3 \pm 4.9\%$	$81.3 \pm 5.1\%$	$84.7 \pm 3.7\%$
	2	$70.3 \pm 3.4\%$	$76.9 \pm 7.0\%$	$82.3 \pm 3.6\%$	$87.2 \pm 3.1\%$	$89.5 \pm 3.1\%$
	3	$72.8 \pm 4.2\%$	$80.3 \pm 5.4\%$	$85.6 \pm 2.1\%$	$88.6 \pm 2.1\%$	$90.2 \pm 2.9\%$
	4	$73.4 \pm 3.7\%$	$81.6 \pm 4.4\%$	$86.3 \pm 1.5\%$	$\underline{89.5 \pm 1.7\%}$	$90.0 \pm 3.7\%$
	5	$\underline{73.9 \pm 5.3\%}$	$\underline{82.3 \pm 5.6\%}$	$\underline{86.7 \pm 2.1\%}$	$89.1 \pm 1.2\%$	$90.2 \pm 3.9\%$
	10	$73.1 \pm 6.0\%$	$80.6 \pm 1.8\%$	$83.9 \pm 2.5\%$	$87.5 \pm 1.6\%$	$90.6 \pm 3.8\%$
	15	$68.4 \pm 4.9\%$	$75.3 \pm 3.9\%$	$83.3 \pm 1.4\%$	$87.3 \pm 1.1\%$	$90.6 \pm 3.4\%$
	20	$59.7 \pm 4.9\%$	$73.4 \pm 5.4\%$	$82.7 \pm 3.5\%$	$88.1 \pm 1.4\%$	$90.6 \pm 2.9\%$
	25	$57.2 \pm 3.7\%$	$70.3 \pm 7.8\%$	$81.3 \pm 3.1\%$	$86.4 \pm 3.1\%$	$\underline{90.8 \pm 2.7\%}$
	30	$55.8 \pm 5.1\%$	$65.8 \pm 9.6\%$	$79.1 \pm 4.0\%$	$85.2 \pm 4.8\%$	$\underline{90.8 \pm 2.7\%}$
10	**0**	$56.1 \pm 2.2\%$	$56.1 \pm 2.2\%$	$56.1 \pm 2.2\%$	$56.1 \pm 2.2\%$	$56.1 \pm 2.2\%$
	1	$60.3 \pm 3.0\%$	$63.6 \pm 1.3\%$	$68.3 \pm 3.1\%$	$72.5 \pm 5.7\%$	$75.8 \pm 3.9\%$
	2	$63.1 \pm 4.3\%$	$69.2 \pm 4.3\%$	$74.5 \pm 3.9\%$	$81.6 \pm 2.1\%$	$85.3 \pm 5.4\%$
	3	$63.6 \pm 4.6\%$	$72.2 \pm 4.9\%$	$79.4 \pm 3.1\%$	$85.8 \pm 2.9\%$	$86.4 \pm 4.6\%$
	4	$63.4 \pm 4.0\%$	$73.3 \pm 4.0\%$	$81.3 \pm 3.8\%$	$86.3 \pm 2.0\%$	$87.2 \pm 5.1\%$
	5	$64.2 \pm 4.3\%$	$75.5 \pm 4.6\%$	$80.5 \pm 3.1\%$	$86.9 \pm 0.9\%$	$87.5 \pm 4.8\%$
	10	$65.2 \pm 4.3\%$	$77.2 \pm 3.3\%$	$80.5 \pm 3.5\%$	$86.7 \pm 1.3\%$	$88.1 \pm 4.4\%$
	15	$61.3 \pm 6.6\%$	$72.7 \pm 5.8\%$	$80.8 \pm 2.9\%$	$86.6 \pm 1.3\%$	$88.4 \pm 4.3\%$
	20	$53.0 \pm 7.2\%$	$72.0 \pm 4.2\%$	$78.1 \pm 2.7\%$	$83.4 \pm 4.3\%$	$88.4 \pm 4.3\%$
	25	$46.4 \pm 8.3\%$	$68.4 \pm 4.3\%$	$76.3 \pm 5.5\%$	$82.7 \pm 4.0\%$	$88.4 \pm 4.3\%$
	30	$42.3 \pm 10.3\%$	$63.4 \pm 2.8\%$	$74.4 \pm 5.7\%$	$82.4 \pm 4.2\%$	$88.4 \pm 4.3\%$

Note: The best classification scores for each n_{lbl} setting are underlined.

A.2.2 Overlapping events in the annotating segments

Table 4. Classification scores of LRM-NMD in low-resolution multi-label learning mode for $\eta = 5$ and $\lambda = 5$ in function of the number of labelled events per annotating segment (n_{lbl}), signal to noise ratio (SNR) and degree of overlap in the annotating segments ($n_{overlap}$).

$\mathbf{n_{overlap}}$ (in %)	**SNR** (in dB)	**Number of labelled events per annotating segment** (n_{lbl})				
		1	**2**	**3**	**4**	**5**
0	**Clean**	$73.9 \pm 5.3\%$	$82.3 \pm 5.6\%$	$86.7 \pm 2.1\%$	$89.1 \pm 1.2\%$	$90.2 \pm 3.9\%$
	20	$73.6 \pm 5.1\%$	$82.3 \pm 5.3\%$	$86.6 \pm 2.1\%$	$88.6 \pm 1.1\%$	$90.2 \pm 3.4\%$
	10	$73.8 \pm 2.6\%$	$82.0 \pm 4.9\%$	$86.4 \pm 1.9\%$	$87.7 \pm 0.9\%$	$89.5 \pm 3.3\%$
	5	$71.9 \pm 3.6\%$	$78.0 \pm 4.4\%$	$81.3 \pm 3.6\%$	$85.9 \pm 1.1\%$	$88.1 \pm 3.3\%$
	3	$60.8 \pm 2.4\%$	$69.7 \pm 3.9\%$	$73.9 \pm 4.4\%$	$80.2 \pm 2.2\%$	$84.4 \pm 3.1\%$
	0	$40.3 \pm 7.0\%$	$44.2 \pm 5.0\%$	$51.7 \pm 4.7\%$	$58.6 \pm 3.1\%$	$61.9 \pm 3.7\%$
25	**Clean**	$69.8 \pm 5.0\%$	$78.8 \pm 4.1\%$	$82.7 \pm 3.7\%$	$87.8 \pm 1.5\%$	$88.6 \pm 3.4\%$
	20	$69.7 \pm 6.1\%$	$78.8 \pm 3.9\%$	$82.5 \pm 3.8\%$	$88.1 \pm 1.1\%$	$88.4 \pm 3.4\%$
	10	$70.3 \pm 4.7\%$	$78.6 \pm 4.9\%$	$82.8 \pm 3.7\%$	$85.8 \pm 2.4\%$	$88.3 \pm 4.7\%$
	5	$67.7 \pm 3.7\%$	$78.0 \pm 3.6\%$	$79.8 \pm 2.3\%$	$83.3 \pm 3.0\%$	$86.1 \pm 4.6\%$
	3	$58.4 \pm 3.3\%$	$67.2 \pm 2.9\%$	$69.7 \pm 1.7\%$	$77.7 \pm 2.3\%$	$81.7 \pm 4.3\%$
	0	$38.1 \pm 8.0\%$	$41.1 \pm 5.6\%$	$46.1 \pm 3.0\%$	$52.7 \pm 7.3\%$	$57.5 \pm 1.4\%$
50	**Clean**	$68.9 \pm 4.3\%$	$75.3 \pm 6.7\%$	$80.0 \pm 3.5\%$	$85.0 \pm 3.7\%$	$88.1 \pm 4.9\%$
	20	$69.4 \pm 4.4\%$	$75.3 \pm 6.4\%$	$80.0 \pm 3.7\%$	$85.6 \pm 3.3\%$	$88.1 \pm 5.3\%$
	10	$68.3 \pm 3.7\%$	$75.3 \pm 6.2\%$	$80.6 \pm 3.9\%$	$85.3 \pm 3.9\%$	$87.0 \pm 4.9\%$
	5	$62.7 \pm 4.1\%$	$71.4 \pm 5.3\%$	$76.4 \pm 5.1\%$	$82.5 \pm 4.2\%$	$85.0 \pm 3.9\%$
	3	$51.9 \pm 4.0\%$	$60.6 \pm 4.4\%$	$65.3 \pm 3.1\%$	$72.8 \pm 3.2\%$	$77.8 \pm 3.8\%$
	0	$30.0 \pm 5.7\%$	$36.3 \pm 4.7\%$	$41.3 \pm 5.0\%$	$46.6 \pm 6.2\%$	$52.2 \pm 3.1\%$
75	**Clean**	$54.2 \pm 1.3\%$	$58.3 \pm 5.7\%$	$67.0 \pm 4.8\%$	$74.7 \pm 5.2\%$	$79.2 \pm 6.1\%$
	20	$53.8 \pm 0.9\%$	$58.6 \pm 6.8\%$	$66.9 \pm 4.1\%$	$75.6 \pm 6.3\%$	$79.1 \pm 5.4\%$
	10	$52.3 \pm 2.4\%$	$57.3 \pm 7.4\%$	$65.8 \pm 4.5\%$	$76.1 \pm 7.3\%$	$78.8 \pm 4.2\%$
	5	$47.3 \pm 5.7\%$	$55.5 \pm 3.1\%$	$59.7 \pm 1.1\%$	$69.1 \pm 4.1\%$	$75.8 \pm 3.2\%$
	3	$39.5 \pm 3.7\%$	$45.8 \pm 3.1\%$	$51.9 \pm 0.5\%$	$64.2 \pm 5.5\%$	$68.6 \pm 2.5\%$
	0	$30.6 \pm 2.7\%$	$34.7 \pm 5.0\%$	$40.8 \pm 4.4\%$	$43.9 \pm 0.3\%$	$45.6 \pm 1.4\%$
100	**Clean**	$44.2 \pm 5.7\%$	$56.4 \pm 1.2\%$	$65.3 \pm 5.8\%$	$73.4 \pm 4.7\%$	$79.8 \pm 4.0\%$
	20	$42.8 \pm 7.8\%$	$55.5 \pm 2.4\%$	$65.0 \pm 5.5\%$	$73.0 \pm 3.9\%$	$78.8 \pm 4.7\%$
	10	$40.5 \pm 5.6\%$	$52.2 \pm 3.7\%$	$64.4 \pm 4.8\%$	$73.6 \pm 2.6\%$	$77.3 \pm 3.5\%$
	5	$36.6 \pm 9.8\%$	$48.4 \pm 8.8\%$	$60.9 \pm 8.7\%$	$63.9 \pm 5.6\%$	$72.0 \pm 3.6\%$
	3	$34.2 \pm 7.6\%$	$43.8 \pm 8.3\%$	$54.2 \pm 11.2\%$	$55.6 \pm 3.4\%$	$62.5 \pm 1.4\%$
	0	$26.7 \pm 5.1\%$	$34.4 \pm 4.4\%$	$37.7 \pm 1.1\%$	$40.5 \pm 1.1\%$	$42.8 \pm 2.6\%$

THE AUTHORS

Lode Vuegen

Peter Karsmakers

Bart Vanrumste

Hugo Van hamme

Lode Vuegen obtained his Master's degree in Electronics ("industriel ingenieur") from KH Kempen, Geel, Belgium, in 2012. His thesis research involved a comparative study on sound classification algorithms. Supported by a VLAIO doctoral scholarship he currently is researching novel acoustical classification and descriptive models for human monitoring. He currently is working on a novel semi-supervised methodology based on Non-Negative Matrix Deconvolution that allows to build classifier models with state-of-the-art classification performance but with less requirements regarding annotation. His main research interest include applications of acoustic pattern recognition, audio signal processing, and machine learning.

●

Peter Karsmakers received a Ph.D. degree in electrical engineering from the KU Leuven in 2010, a M.Sc. degree in Artificial Intelligence from the KU Leuven in 2004, a M.Sc. in electronics-ICT ("industrieel ingenieur") from the KH Kempen in 2001. Since 2001 he is teaching at the KU Leuven campus Geel. In 2013 he became a senior researcher within the AdvISe research lab from the KU Leuven. His expertise resides in the application of state-of-the-art signal processing, and machine learning algorithms to real-life problems.

●

Bart Vanrumste received his M.Sc. in electrical engineering and M.Sc. in biomedical engineering both from Ghent University in 1994 and 1998, respectively. In 2001 he received his Ph.D. in engineering from the same institute. He worked as a post-doctoral fellow from 2001 until 2003 at the electrical & computer engineering department of the University Of Canterbury, New Zealand. He is member of the faculty of engineering technology at KU Leuven. He is affiliated with eMedia Lab at GroupT, ESAT-STADIUS division and IMEC. His research interests is decision support in healthcare in general and ambient assisted living in particular. His current research activities focus among other on multimodal sensor integration for monitoring of older persons and patients with chronic illnesses at their homes. He is senior member of IEEE Engineering in Medicine and Biology and member of the International Society for Bioelectromagnetism.

●

Hugo Van hamme received the Ph.D. degree in electrical engineering from Vrije Universiteit Brussel (VUB), Brussel, Belgium, in 1992, the M.Sc. degree from Imperial College, London, U.K., in 1988, and the Master's degree in engineering ("burgerlijk ingenieur") from VUB, in 1987. From 1993 to 2002, he worked for L&H Speech Products and ScanSoft, initially as a Senior Researcher and later as a Research Manager. Since 2002, he has been a Professor in the Department of Electrical Engineering, KU Leuven, Leuven, Belgium. His main research interests include applications of speech technology in education and speech therapy, computational models for speech recognition and language acquisition, and noise robust speech recognition.