SIMILARITY INDUCED GROUP SPARSITY FOR NON-NEGATIVE MATRIX FACTORISATION

Antti Hurmalainen[⋆] *Rahim Saeidi*[†] *Tuomas Virtanen*[⋆]

[⋆] Department of Signal Processing, Tampere University of Technology, Tampere, Finland
[†] Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland

ABSTRACT

Non-negative matrix factorisations are used in several branches of signal processing and data analysis for separation and classification. Sparsity constraints are commonly set on the model to promote discovery of a small number of dominant patterns. In group sparse models, atoms considered to belong to a consistent group are permitted to activate together, while activations across groups are suppressed, reducing the number of simultaneously active sources or other structures. Whereas most group sparse models require explicit division of atoms into separate groups without addressing their mutual relations, we propose a constraint that permits dynamic relationships between atoms or groups, based on any defined distance measure. The resulting solutions promote approximation with components considered similar to each other. Evaluation results are shown for speech enhancement and noise robust speech and speaker recognition.

Index Terms— non-negative matrix factorization, group sparsity, sparse representations, speech recognition, speaker recognition

1. INTRODUCTION

Many natural signals and data sets can be represented as additive combinations of underlying sources and their more primitive components. *Non-negative matrix factorisation* (NMF) is a common algorithm for *compositional modelling* with a *basis* of atoms representing such components [1, 2]. Their relative *activation weights* reveal the estimated presence of each individual component, hence performing separation, classification, and approximation of the content of observations. Factorisations are referred to as *unsupervised*, *semi-supervised*, or *supervised*, depending on whether all, some, or none of the atoms are adapted in the process beside activations.

Supervised factorisation with an overcomplete basis may lead to an infinite number of equivalent solutions. *Sparsity constraints* are often introduced to favour modelling with a small number of best matching components. Discovering such key elements is known as *sparse classification* (SC), and it has been used for diverse purposes such as face recognition [3], music genre classification [4], and general audio analysis [5]. In our previous work, NMF algorithms have been applied to noisy speech for automatic speech recognition (ASR) [6] and speaker recognition/identification [7] via SC.

Common sparsity measures like an L_1 norm penalty term on activations promote reduction of the number of active elements. Although different penalty weights can be set on each atom, in general this does not address the internal structure of activations, that is, co-occurrence of atoms. In many scenarios, however, the components are naturally structured, hence appropriate modelling would be preferable over treating the atoms as independent entries.

This work is supported by Academy of Finland (project # 256961 / 284671).

Group sparsity functions have been proposed for general sparse models in [8, 9, 10, 11, 12]. Recent studies on e.g. *group LASSO* regularisation also permit overlapping and variable-size groups. Models for NMF in particular were demonstrated in [13] and [14] for modelling speech with multiple speaker-dependent bases. In these experiments, a group comprises various phonetic patterns of a single speaker. Under an assumption that only one, unknown target speaker is present in a noisy observation, the model will converge to a solution with only the most likely candidates active, instead of combining features arbitrarily from all speakers. Because less active non-target speech features and other speech-like interferences are suppressed, improvements have been observed in denoising [13, 14], SC-based ASR [15], and speaker recognition [15].

A typical approach to group sparsity is to use two different measures for activations; an inner sparsity function within a group, and another over groups. By appropriate selection of the functions, activations from the same group have a lower cost than the same coefficients distributed across groups. In [15], these measures were L_2 and L_1 norms, respectively. In [13], several combinations were discussed with $\log$ and L_1 functions chosen for experiments to reduce the number of active speaker bases in denoising. They were also used in [16] for separation of music and in [17] for audio events.

There are two problems with this general approach, though. First, regardless of the choice of the inner function, its minimisation will also penalise any activations within the selected group(s), which may not be desired. Second, these nested functions in their basic form do not consider *which* groups will activate, alone or together. While these models promote general group sparsity, for many purposes it would be beneficial to have more control on the relationships of different atoms or groups. For this purpose, we propose a quadratic penalty function, where a distance measure can be defined between elements to promote diverse structures of underlying models. The new group sparse model does not alter the structure of individual activations within groups, hence their penalisation may be defined independently from the group cost function.

The proposed model with its mathematical properties is introduced in Section 2. An experimental framework demonstrating the method on recognition of noisy speech by an unknown speaker from a closed set is described in Section 3. Results and conclusions are given in Sections 4 and 5, respectively.

2. DISTANCE-BASED GROUP SPARSITY MODEL

2.1. Sparse representation

Let us define the baseline sparse NMF model within a B-dimensional feature space so that an observation vector $\mathbf{y}$ is estimated with L basis vectors $\mathbf{a}$ stored in a $B \times L$ matrix $\mathbf{A}$. The estimate of $\mathbf{y}$ is $\boldsymbol{\psi}$, given as

$$\boldsymbol{\psi} = \mathbf{A}\mathbf{x}, \tag{1}$$

where $\mathbf{x}$ is a length L *activation vector*. All arrays in the model are non-negative. The core task is to find the sparse representation $\mathbf{x}$, which minimises a cost function of a general form

$$f_{\mathrm{tot}} = d(\mathbf{y}, \boldsymbol{\psi}) + f_\lambda(\mathbf{x}) \tag{2}$$

for some *estimate distance function* d between the observation and its estimate, and a *sparsity cost function* f_λ. Typical choices for d include Euclidean distance and generalised Kullback-Leibler divergence. Basic sparsity for individual atoms is often achieved by using the L_1 norm of $\mathbf{x}$ as the cost function. The aforementioned L_2/L_1 group sparsity is denoted by $f_{\mathrm{g}} = \sum_{s=1}^{S} \left\|\mathbf{x}_s\right\|_2$, where $\mathbf{x}_s$ is a subvector of $\mathbf{x}$ containing the indices belonging to group s of S.

2.2. Group cost for atom distances

Let us define an $L \times L$ *group distance matrix* $\mathbf{M}$, where each element $m_{i,j}$ is a distance measure between atoms i and j. The cost function to be minimised is

$$f_{\mathrm{m}}(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x}, \tag{3}$$

which measures the sum of all $x_i m_{i,j} x_j$ coefficients for $i, j \in L$. We notice that the formulation itself is equivalent to Tikhonov regularisation, whose special case employing a weighted identity matrix is often used for *ridge regression* in least squares problems. Indeed, nonzero diagonal elements may be used for the same purpose here as well. However, in the following analysis we concentrate on the non-diagonal cross terms, and actually assume the diagonal to consist of zeros according to a general interpretation of distances, which should be zero to the element itself and non-negative between any two elements. The common definition of a distance also implies symmetricity, and even for any measures violating it (such as KL-divergence), the original matrix $\mathbf{M}$ may be replaced with its symmetric counterpart $\hat{\mathbf{M}} = (\mathbf{M} + \mathbf{M}^{\mathrm{T}})/2$ without affecting any mathematical or practical aspects of this work.

2.3. Properties and analysis

We can immediately observe the following properties of the elements of $\mathbf{M}$:

1. Co-occurrence of x_i and x_j will not be penalised if and only if both cross-terms $m_{i,j}$ and $m_{j,i}$ are zeros.
2. An infinite entry $m_{i,j} = \infty$ (or a very large value) prevents x_i and x_j from activating together.
3. All other m values produce a quadratic penalty weight for co-occurrence of the corrersponding $\mathbf{x}$ entries.

Consequently, the model will favour activation patterns, where the cross-terms between all activated atoms are low, standing for a small distance and thus high similarity.

Function f_{m} is quadratic. Its gradient, used for iterative NMF update rules, is $(\mathbf{M}+\mathbf{M}^{\mathrm{T}})\mathbf{x}$ and its Hessian matrix $\mathbf{M}+\mathbf{M}^{\mathrm{T}}$. Both are non-negative. These become $2\mathbf{M}\mathbf{x}$ and $2\mathbf{M}$, respectively, when symmetry is assumed or enforced. Because we defined the diagonal elements as zeros or small compared to the cross-terms, the 2×2 principal minors of $\mathbf{M}$ are non-positive, and the Hessian indefinite. Thereby the cost function is strongly nonconvex by design. For minimisation of general quadratic functions in open sets this would be a major issue. However, in non-negative modelling, it only causes the minima to appear on the axes. Also, when the measure is combined with the original modelling task and its residual cost function, the convexity and behaviour of the overall function are not smooth.

Let us illustrate the matter with a simple example, where the function to be minimised is $f_{\mathrm{tot}} = d(\mathbf{y}, \boldsymbol{\psi}) + \lambda f_{\mathrm{m}}(\mathbf{x})$ for arrays

$$\mathbf{y} = \begin{bmatrix} 4 \\ 4 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}, \quad \mathbf{M} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

If the sparsity weight λ is zero, $\mathbf{x}$ clearly has a solution $\begin{bmatrix} 1 & 1 \end{bmatrix}^{\mathrm{T}}$ for any common distance function d. However, assuming Euclidean distance d_{Euc} as the residual measure, the overall function has a uniform Hessian

$$\mathbf{H} = \begin{bmatrix} 20 & 12 \\ 12 & 20 \end{bmatrix} + \lambda \begin{bmatrix} 0 & 2 \\ 2 & 0 \end{bmatrix}. \tag{4}$$

If $\lambda < 4$, the global minimum is at $\mathbf{x} = 16/(\lambda + 16) \begin{bmatrix} 1 & 1 \end{bmatrix}^{\mathrm{T}}$. If $\lambda > 4$, the function becomes a saddle with its minima in the non-negative activation space at $\mathbf{x} = \begin{bmatrix} 1.6 & 0 \end{bmatrix}^{\mathrm{T}}$ and $\begin{bmatrix} 0 & 1.6 \end{bmatrix}^{\mathrm{T}}$. In other words, cross-activations have been eliminated at the cost of estimate $\boldsymbol{\psi}$ coefficients becoming 1.6 and 4.8, either way. If the distance function is changed to generalised KL-divergence d_{KL}, the same fundamental behaviour remains, but non-uniform, $\mathbf{x}$-dependent convexity of the overall function allows a small range of λ values, where both the local minimum on the diagonal and the two global minima on the axes are present simultaneously.

For actual data and more diverse $\mathbf{M}$ matrices, the solutions become varied. For example, a large $m_{i,j}$ value may split the activation space into two halves, while both halves still have their own minima, unhindered by the f_{m} cost if these cross-terms are low. Small λ values alter the selection between almost equivalent atoms without dominating the result. Conversely, large λ values result in strong nonconvexity, which calls for careful initialisation to guide the solution to the correct minimum. In our experiments, running initial NMF iterations without the group cost term and then introducing it gradually produced stable descent to accurate group sparse solutions.

2.4. Extension to groups of activations

The model described in Equation (3) applies to single activation vectors with matrix $\mathbf{M}$ defined between individual atoms. This already permits group structures in the sense that e.g. multiple atoms may be assigned to the same group by giving them zero cross-distances. However, in our speech processing work, the common application is to find optimal multi-column activation matrices $\mathbf{X}$, where

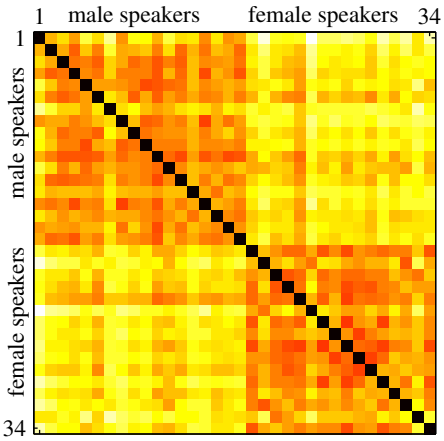
1. large sets of atoms belong to few groups, and
2. structures should apply coherently to multiple observation indices, that is, columns of $\mathbf{X}$.

Both objectives can be met by defining the distances for *sets of activations*, which may span multiple rows and columns of $\mathbf{X}$ or even more complex structures. Nevertheless, to simplify the formulation, we concentrate on rectangular blocks of $\mathbf{X}$ ($L \times W$), whose W columns correspond to *window indices* of audio spectrograms. The set activation weights z_s for S sets are computed as

$$z_s = \sum_{l \in G_s} \sum_{w=1}^{W} \mathbf{X}_{l,w} \tag{5}$$

with G_s denoting atoms indices belonging to set s. Group distance $\mathbf{M}$ with order $S \times S$ is then defined between whole groups of atoms, and the cost computed as $f_{\mathrm{m}} = \mathbf{z}^{\mathrm{T}}\mathbf{M}\mathbf{z}$ for sets of activations instead of individual entries. The group cost is introduced to common NMF update rules [1, 2] by adding the strictly non-negative gradient of f_{m}, $2\mathbf{M}\mathbf{x}$ (or $2\mathbf{M}\mathbf{z}$) to the denominator of activation update, in the latter case applied to all activations belonging to each z_s. For example, the update rule for atom-level group sparsity with KL-divergence will be

Fig. 1. Distance matrix for CHiME speakers, sorted by gender. Bright/yellow entries correspond to high distance values.



$$\mathbf{x} \leftarrow \mathbf{x} \otimes \frac{\mathbf{A}^{\mathrm{T}}(\mathbf{y}/(\mathbf{A}\mathbf{x}))}{\mathbf{A}^{\mathrm{T}}\mathbf{1} + 2\mathbf{M}\mathbf{x}} \tag{6}$$

with $\otimes$-multiplication and all divisions taken elementwise.

3. EXPERIMENTS ON NOISY SPEECH

3.1. Noisy speech corpus

The proposed method is tested in enhancement and sparse classification of noisy speech from the 1st CHiME Challenge [18]. The task is based on GRID corpus [19], where one of 34 speakers utters a six-word command sentence following a linear grammar with a vocabulary of 51 words. In the CHiME corpus, utterances are mixed with living room noise at six SNRs from +9 to -6 dB, each using a different instance of noise without rescaling. Long noise context is available to both directions for adapting the models. For training of speech models, there are 500 utterances per speaker without additive noise. Development and test sets comprise 600 utterances per SNR.

The challenge task is to conduct automatic speech recognition, where two keywords per utterance ('letter' and 'digit') are scored. In the original challenge, speaker identity is known. Nevertheless, to demonstrate the group sparse methods, we run the factorisation and recognition tasks blindly with concatenated speaker models, expecting the model to find the correct identity while suppressing others. The single, correct speaker model thus acts as an oracle baseline.

3.2. Experimental set-up

Most of the set-up follows our earlier NMF experiments on the same corpus [6, 7, 15]. Features were extracted as 40 monaural mel magnitudes with 25 ms frame length and 10 ms shift. A speech basis of 250 templates, each a 40×25 spectrogram segment, was extracted for all 34 speakers by computing an average spectro-temporal profile of each back-end state and its context from 300 training utterances per speaker. These were concatenated to a joint speech basis of $34 \cdot 250$ atoms. For noisy sets, 250 noise atoms of the same size were extracted from the neighbourhood of each utterance individually by sampling the noise context. Thereby 8500 atoms were used in factorisation of clean speech, and 8750 in noisy conditions.

Each single-speaker basis forms a group in our system. Because matching atom indices contain the same linguistic content in every speaker's basis, it is relatively straightforward to compute a distance between each pair of bases. After experimenting with e.g. KL and Euclidean distances between atoms, we settled for computing the angle between vectorised atom spectrograms $\mathbf{a}$ as

$$\angle(\mathbf{a}_l^{(i)}, \mathbf{a}_l^{(j)}) = \cos^{-1} \frac{\mathbf{a}_l^{(i)} \cdot \mathbf{a}_l^{(j)}}{\|\mathbf{a}_l^{(i)}\| \|\mathbf{a}_l^{(j)}\|} \tag{7}$$

pairwise for each atom index l in the bases of speakers i and j. These were summed over atoms to form the $m_{i,j}$ entries. Finally, the 34×34 distance matrix $\mathbf{M}$ was normalised to a mean value of 1 over non-diagonal entries. This measure was found to produce plausible estimates of speaker similarity. For example, the average distance between speakers of the same gender was approximately 20% lower than across genders. The matrix is shown in Figure 1, sorted by gender. Apart from general trends, diverse levels of pairwise similarity between speakers can be observed in individual entries.

Activation matrices were computed with convolutive NMF described in [6, 15] using a combination of KL-divergence, 1-norm of activations weighted by matrix $\mathbf{\Lambda}_1$ (with weights 0.1 and 0.085 for speech and noise, respectively), and the new group sparsity cost together as the target function,

$$f_{\mathrm{tot}} = d_{\mathrm{KL}}(\mathbf{Y}, \mathbf{\Psi}) + \|\mathbf{\Lambda}_1 \otimes \mathbf{X}\|_1 + \lambda_{\mathrm{m}} f_{\mathrm{m}}(\mathbf{z}). \tag{8}$$

Apart from changing the group sparsity function, the set-up was identical to [15]. Group sparsity factor λ_{m} was optimised to 0.0025 on the development set, and multiplied by the mean of atom 1-norms like other sparsity weights in related work. To address the possibility of multiple minima in the cost function, the group cost was introduced linearly from zero to λ_{m} over iterations 101–200 so that an approximate non-sparse solution would be reached first before converging into a reduced number of speaker bases. All in all, 400 iterations were used for sufficient convergence of the final cost.

Table 1. Average results over the noisy CHiME test set, measured by the five metrics given in Section 3.3. Rows correspond to different speaker selection methods; joint bases without group sparsity ('no GS'), previous L_2/L_1 group sparsity [15], proposed distance-based model, and oracle factorisation with the correct speaker's basis only.

method	act.%	SDR/dB	FE-ASR	SC-ASR	spk.rec
no GS	35.6	6.52	83.5	76.5	95.0
L_2/L_1	45.8	7.40	83.8	80.9	95.7
proposed	79.1	8.01	83.6	80.1	94.6
oracle	100	8.62	85.2	81.8	100

3.3. Result metrics

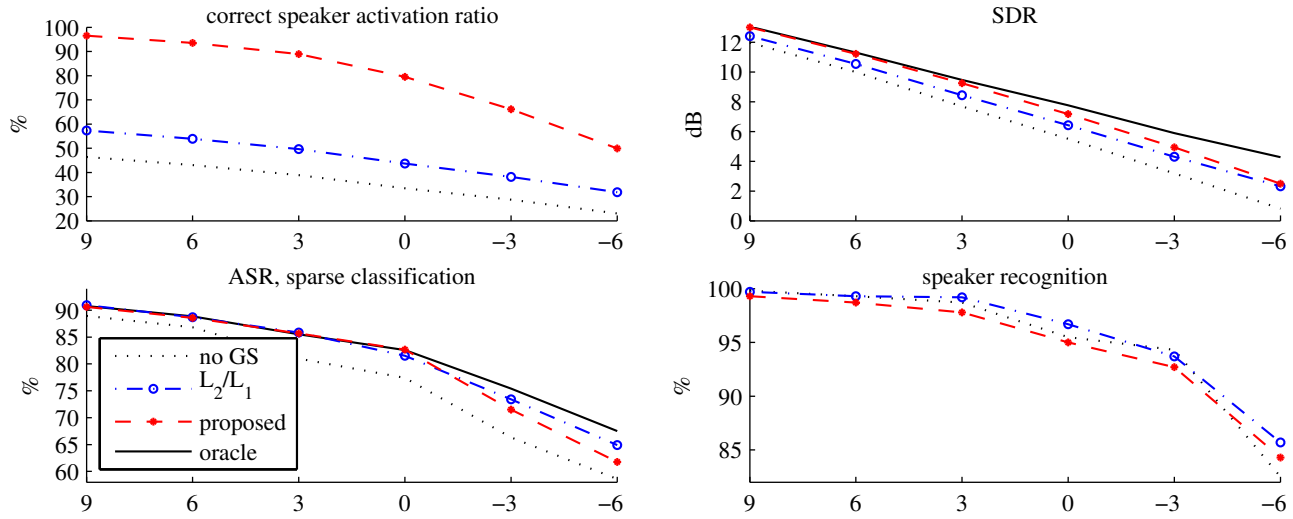
The following metrics were used to evaluate the results.

Correct speaker activity percentage (act.%): For each utterance, a normalised histogram of speaker activity (the $\mathbf{z}$ vector) was calculated. The amount of correct speaker's activity was averaged over all 600 utterances per condition.

Signal-to-distortions ratio (SDR): Speech features were enhanced by computing a spectro-temporal filter $\mathbf{\Psi}^{\mathrm{s}}/\mathbf{\Psi}^{\mathrm{tot}}$ binwise between utterance estimates from speech-only and all atoms. Both were converted from mel to DFT magnitudes by the pseudoinverse of the mel matrix as in [6, 15]. Enhancement quality of synthesised waves was measured by SDR using BSSeval toolkit (without highpass filtering) [20] and averaged over utterances.

ASR by feature enhancement (FE): Keyword accuracy was determined by recognising the enhanced utterances with the same multi-condition trained CHiME GMM back-end as in earlier work [6, 15].

Fig. 2. Results for separation and recognition experiments, plotted over the noise levels of 1st CHiME test set.



ASR by sparse classification (SC): For SC, state likelihoods were determined directly from activation weights by assigning $Q \times T$ mapping matrices to atoms, where $Q = 250$ is the number of back-end states and $T = 25$ the number of frames in an atom [6]. These were learnt from factorisations of 200 training matrices per speaker by using the deconvolutive algorithm described in [21]. State likelihood matrices were decoded with CHiME baseline HMMs.

Speaker recognition: As in [7, 15], the average activation vector over each utterance was used to map a variable-length utterance to a fixed-length high-dimensional vector. A sparse linear discriminant analysis (LDA) model is trained using 200 factored clean utterances per speaker, and the cosine similarity metric is employed after LDA.

All factorisation and recognition parameters were optimised with development data. No optimisation took place during test set scoring.

4. RESULTS AND DISCUSSION

Results for the CHiME test set are shown as averages in Table 1, and plotted over SNR in Figure 2. In both representations, the proposed method is compared to joint-basis factorisation with no group sparsity, the previous L_2/L_1 group cost, and oracle factorisation using only the correct speaker's basis (where applicable). We notice that both group sparsity models clearly surpass joint factorisation without such constraint, but lose to oracle factorisation as expected. Different goals and modelling methods are discussed briefly as follows.

Measuring the relative speaker activities reveals that the proposed model is very efficient in finding the correct speaker and emphasising the corresponding activations. Although the sparsity weight λ obviously plays a role in the shaping, the new function was clearly more discriminative for all λ values producing plausible recognition results. This is explained by the new function's tendency to remove nonmatching groups from the model completely.

Because a majority of activations was condensed on the correct speaker, SDRs of the enhanced signals improved uniformly. However, this did not produce significant changes in enhancement-based ASR. All joint-basis methods were roughly tied. A possible explanation is that even incorrect speaker models are able to separate speech from other noises. All factorisation methods were still successful, as the average SDR for unenhanced signals was -0.78 dB and the corresponding keyword recognition rate 74.7%.

In sparse classification of speech, the results were not completely consistent. Despite better basis selection, the average recognition rate decreased compared to the L_2/L_1 model due to worse performance at low SNRs. One explanation is that the atom-level shrinkage of L_2/L_1 sparsity may actually change the activation pattern favourably for noisy conditions. Nevertheless, both costs clearly surpass baseline joint factorisation.

In speaker recognition, the results did not improve over previous models. The primary reason is that clean training factorisations concentrated very heavily on single bases, producing classifiers that do not generalise well to mismatched noisy activation patterns.

Overall, the results show the method's potential, although the exact outcome depends on the task. One major factor in this scenario is that especially at low SNRs, CHiME utterances contain competing voices, whose energy may exceed the target speech. In such cases, the new model may converge to a non-target speaker, largely losing the target speech, whereas models without similarity constraints can keep both. Inspecting the behaviour of single utterances revealed that this indeed happened, causing a few strongly negative outliers among generally improving separation. The task of picking a lower energy speaker blindly is fundamentally difficult, though. A better noise model or spatial methods might help in tackling these scenarios. Another factor is the training of classifiers from clean data, whose activation pattern in this model is more mismatched than earlier, thus multi-condition training should be used for SC instead.

5. CONCLUSIONS

A group sparsity model was proposed for non-negative matrix factorisation, based on defining distance measures between groups. The model favours solutions, where a small number of groups is active with a further preference for mutual similarity. The measure may be defined by any means suitable for the task, and it permits customisable pairwise penalties ranging from zero to complete exclusion. The function does not affect the distribution of activations within groups, unlike commonly used group costs. Initial experiments on factorisation of speech with joint speaker bases suggest that better convergence to the correct speaker within a closed set is achieved. However, as the function is nonconvex by design, proper initialisation is necessary. As future work, we expect to apply the model to other separation and classification tasks, and to refine its details.

6. REFERENCES

[1] D. D. Lee and H. S. Seung, "Algorithms for Non-negative Matrix Factorization," in *Advances in Neural Information Processing Systems 13*, 2001, pp. 556–562.

[2] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations*, Wiley, New York, NY, USA, 2009.

[3] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust Face Recognition via Sparse Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[4] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Music Genre Classification via Sparse Representations of Auditory Temporal Modulations," in *Proc. of the 17th EUSIPCO*, Glasgow, Scotland, UK, 2009, pp. 1–5.

[5] T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis, "Compositional models for audio processing," *IEEE Signal Processing Magazine*, vol. 32, no. 2, 2015.

[6] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Modelling non-stationary noise with spectral factorisation in automatic speech recognition," *Computer Speech & Language*, vol. 27, no. 3, pp. 763–779, 2013.

[7] R. Saeidi, A. Hurmalainen, T. Virtanen, and D. A. van Leeuwen, "Exemplar-based Sparse Representation and Sparse Discrimination for Noise Robust Speaker Identification," in *Odyssey speaker and language recognition workshop*, Singapore, 2012.

[8] S. Bengio, F. C. N. Pereira, Y. Singer, and D. Strelow, "Group Sparse Coding," in *Proc. of NIPS*, 2009, pp. 82–89.

[9] M. Kowalski, K. Siedenburg, and M. Dörfler, "Social Sparsity! Neighborhood Systems Enrich Structured Shrinkage Operators," *IEEE Transactions on Signal Processing*, vol. 61, pp. 2498–2511, 2013.

[10] G. Obozinski, L. Jacob, and J.-P. Vert, "Group Lasso with Overlaps: the Latent Group Lasso approach," Tech. Rep., INRIA, 2011, arXiv:1110.0413.

[11] Q. Tan and S. Narayanan, "Novel Variations of Group Sparse Regularization Techniques with Applications to Noise Robust Automatic Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 1337–1346, 2012.

[12] S. Gishkori and G. Leus, "Compressed Sensing for Block-Sparse Smooth Signals," in *Proc. of the 39th ICASSP*, Florence, Italy, 2014, pp. 4166–4170.

[13] D. L. Sun and G. J. Mysore, "Universal Speech Models for Speaker Independent Single Channel Source Separation," in *Proc. of the 38th ICASSP*, Vancouver, BC, Canada, 2013, pp. 141–145.

[14] M. Kim and P. Smaragdis, "Mixtures of Local Dictionaries for Unsupervised Speech Enhancement," *IEEE Signal Processing Letters*, vol. 22, no. 3, pp. 294–297, 2015.

[15] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Group Sparsity for Speaker Identity Discrimination in Factorisation-based Speech Recognition," in *Proc. of the 13th INTERSPEECH*, Portland, OR, USA, 2012, pp. 2138–2141.

[16] A. Lefèvre, F. Bach, and C. Févotte, "Itakura-Saito Nonnegative Matrix Factorization with Group Sparsity," in *Proc. of the 36th ICASSP*, Prague, Czech Republic, 2011, pp. 21–24.

[17] D. El Badawy, N. Q. K. Duong, and A. Ozerov, "On-the-fly Audio Source Separation," in *Proc. of 24th MLSP*, Reims, France, 2014.

[18] J. Barker, E. Vincent, N. Ma, C. Christensen, and P. Green, "The PASCAL CHiME Speech Separation and Recognition Challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, 2013.

[19] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An Audio-visual Corpus for Speech Perception and Automatic Speech Recognition," *Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[20] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[21] A. Hurmalainen and T. Virtanen, "Learning State Labels for Sparse Classification of Speech with Matrix Deconvolution," in *Proc. of ASRU*, Olomouc, Czech Republic, 2013, pp. 168–173.