

Supervised Group Nonnegative Matrix Factorisation With Similarity Constraints And Applications To Speaker Identification

Romain Serizel, Victor Bisot, Slim Essid, Gael Richard

▶ **To cite this version:**

Romain Serizel, Victor Bisot, Slim Essid, Gael Richard. Supervised Group Nonnegative Matrix Factorisation With Similarity Constraints And Applications To Speaker Identification. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Mar 2017, New Orleans, United States. hal-01484744

HAL Id: hal-01484744

https://inria.hal.science/hal-01484744

Submitted on 7 Mar 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

SUPERVISED GROUP NONNEGATIVE MATRIX FACTORISATION WITH SIMILARITY CONSTRAINTS AND APPLICATIONS TO SPEAKER IDENTIFICATION

Romain Serizel[⋆§†], *Victor Bisot*[‡], *Slim Essid*[‡], *Gaël Richard*[‡]

[⋆] Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54506, France
[§]Inria, Villers-lès-Nancy, F-54600, France
[†]CNRS, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54506, France
[‡]LTCI, CNRS, Télécom ParisTech, Université Paris - Saclay, F-75013, Paris, France

ABSTRACT

This paper presents supervised feature learning approaches for speaker identification that rely on nonnegative matrix factorisation. Recent studies have shown that group nonnegative matrix factorisation and task-driven supervised dictionary learning can help performing effective feature learning for audio classification problems.

This paper proposes to integrate a recent method that relies on group nonnegative matrix factorisation into a task-driven supervised framework for speaker identification. The goal is to capture both the speaker variability and the session variability while exploiting the discriminative learning aspect of the task-driven approach. Results on a subset of the ESTER corpus prove that the proposed approach can be competitive with I-vectors.

***Index Terms*—** Nonnegative matrix factorisation, feature learning, dictionary learning, online learning, speaker identification

1. INTRODUCTION

The main target of speaker identification is to assert whether or not the speaker of a test segment is known and if he/she is known, to determine his/her identity. Since their emergence about five years ago, the I-vectors [1] have become the state-of-the-art approach for speaker identification [2] and a typical speaker identification system is composed of I-vector extraction, normalisation [3, 4] and classification with probabilistic linear discriminant analysis (PLDA) [5].

On the other hand, recent studies have shown that approaches such as nonnegative matrix factorisation (NMF) [6] can be successfully exploited to perform spectrogram factorisation [7, 8, 9] or multimodal co-factorisation [10] to retrieve speaker identity. Capitalising on this, we have recently proposed an approach based on group-NMF (GNMF) [11] and inspired by the I-vector training procedure that allowed them to take into account inter-speaker and inter-session variability by constraining a set of speaker-dependent bases across sessions and a set of session-dependent bases across speakers [12]. This approach was shown to be competitive with a state-of-the-art I-vector system.

The GNMF approach allows one to exploit, to some extent, annotations about recording sessions and speakers [12]. However, this approach does not enable the possibility to enforce the "discriminativity" of the learned dictionaries, which can be of tremendous importance when the final target is a classification problem (as it is the case here). A supervised matrix factorization approach proposed recently and known as Task-driven Dictionary Learning (TDL) [13] allows for learning a dictionary jointly with a classifier, therefore enforcing the discriminative quality of the dictionary. This approach was later extended to nonnegative dictionaries and adapted to audio classification problem demonstrating significant performance improvement compared to unsupervised approaches [14].

In this paper, we propose a new formulation of the GNMF method. Using the Euclidean distance as the divergence for the GNMF problem, the dictionary learning based on GNMF is integrated in a supervised framework inspired by TDL [13]. The choice of the Euclidean distance renders the latter process more efficient and and allows to improve previous results be up to 9 % in some cases. In a first step, the nonnegative extension of the TDL [14, 15] is applied to the dictionaries obtained with standard NMF or GNMF, in order to fine-tune the dictionaries. We then consider a task-driven formulation for the GNMF method with speaker and session variability constraints [12]. This approach allows our system to discriminatively learn nonnegative dictionaries that also capture both the speaker variability and the session variability, and improve the performance further.

The paper is organised as follows. The problem, the notations and the general NMF approach are introduced in Section 2. The GNMF approach with Euclidean distance is described in Section 3. Then, the task-driven NMF approaches are described in Section 4. Experiment results are presented in Section 5. Finally, conclusions are exposed in Section 6.

2. PROBLEM STATEMENT

2.1. Notations

Consider the (nonnegative) time-frequency representation of an audio signal $\mathbf{V} \in \mathbb{R}_+^{F \times N}$, where F is the number of frequency components and N the number of frames. $\mathbf{V}$ is composed of data collected during S recording sessions with speech segments originating from C speakers. In each session several speakers can be present and a particular speaker can be present in several sessions. Let $\mathcal{C}$ denote the set of speakers and $\mathcal{S}$ the set of sessions. The number of elements in an ensemble is denoted by $\mathrm{Card}(.)$. $\mathrm{Card}(\mathcal{C}) = C$ and $\mathrm{Card}(\mathcal{S}) = S$. Let $\mathcal{C}^s$ denote the subset of speakers that appear in the session s ($\mathcal{C}^s \subset \mathcal{C}$) and $\mathcal{S}^c$ the subset of sessions in which the speaker c is active ($\mathcal{S}^c \subset \mathcal{S}$). In the remainder of this paper, superscripts ^{c} and ^{s} will denote the current speaker and session, respectively.

This work was partly funded by the European Union under the FP7-LASIE project (grant 607480).

2.2. NMF with Euclidean distance

The goal of NMF [6] is to find a factorisation for $\mathbf{V}$ of the form:

$$\mathbf{V} \approx \mathbf{WH} \tag{1}$$

where $\mathbf{W} \in \mathbb{R}_+^{F \times K}$, $\mathbf{H} \in \mathbb{R}_+^{K \times N}$ and K is the number of components in the decomposition. NMF model estimation can be formulated as the following optimisation problem:

$$\min_{\mathbf{WH}} \|\mathbf{V} - \mathbf{WH}\|_2^2 \quad \text{s.t.} \quad \mathbf{W} \geq 0,\ \mathbf{H} \geq 0. \tag{2}$$

with $\|.\|_2$ the Euclidean distance. One algorithm that is built to solve this problem exploits multiplicative update rules which are obtained using the heuristic consisting in expressing the gradient of the cost function (2) as the difference between a positive contribution and a negative contribution [16]:

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T\mathbf{V}}{\mathbf{W}^T\mathbf{WH}} \quad \text{and} \quad \mathbf{W} \leftarrow \mathbf{W} \odot \frac{\mathbf{VH}^T}{\mathbf{WHH}^T}; \tag{3}$$

where $\odot$ is the element-wise product (Hadamard product) and division is element-wise.

3. GNMF WITH SPEAKER AND SESSION SIMILARITY

We have recently proposed an approach that derives from GNMF [11] and intends to take speaker and session variability into account [12]. The approach was proposed for NMF with the generalised Kullback-Leibler divergence [17]. Here we introduce its counterpart for the Euclidean distance.

We first decompose $\mathbf{V}$ into portions $\mathbf{V}^{(cs)}$ of length $N^{(cs)}$ that are recorded in a session s in which only the speaker c is active. The global cost function (J_{global}) minimized in (2) can then be seen as the sum of all local divergences:

$$J_{\text{global}} = \sum_{c=1}^{C} \sum_{s \in \mathcal{S}_c} \|\mathbf{V}^{(cs)} - \mathbf{W}^{(cs)}\mathbf{H}^{(cs)}\|_2^2. \tag{4}$$

3.1. Class and session similarity constraints

We further decompose the dictionaries $\mathbf{W}^{(cs)}$ as follows:

$$\mathbf{W}^{(cs)} = [\underbrace{\mathbf{W}_{\text{SPK}}^{(cs)}}_{\leftarrow K_{\text{SPK}} \rightarrow} \mid \underbrace{\mathbf{W}_{\text{SES}}^{(cs)}}_{\leftarrow K_{\text{SES}} \rightarrow} \mid \underbrace{\mathbf{W}_{\text{RES}}^{(cs)}}_{\leftarrow K_{\text{RES}} \rightarrow}]$$

with $K_{\text{SPK}} + K_{\text{SES}} + K_{\text{RES}} = K$ and where K_{SPK}, K_{SES} and K_{RES} are the number of components in the speaker-dependent bases, the session-dependent bases and the residual bases, respectively.

In order to capture speaker and session variability we define two constraints [12]. The first constraint is related to the distance between the speaker bases:

$$J_{\text{SPK}} = \frac{1}{2} \sum_{c=1}^{C} \sum_{s \in \mathcal{S}_c} \sum_{\substack{s_1 \in \mathcal{S}_c \\ s_1 \neq s}} \|\mathbf{W}_{\text{SPK}}^{(cs)} - \mathbf{W}_{\text{SPK}}^{(cs_1)}\|^2 < \alpha_1 \tag{5}$$

with α_1, the similarity level on speaker-dependent bases.

The second constraint is related to the distance between the session bases:

$$J_{\text{SES}} = \frac{1}{2} \sum_{s=1}^{S} \sum_{c \in \mathcal{C}_s} \sum_{\substack{c_1 \in \mathcal{C}_s \\ c_1 \neq c}} \|\mathbf{W}_{\text{SES}}^{(cs)} - \mathbf{W}_{\text{SES}}^{(c_1 s)}\|^2 < \alpha_2 \tag{6}$$

where α_2 is the similarity level for session-dependent bases.

Minimizing the global divergence (4) subject to constraints (5) and (6) is related to the following problem:

$$\min_{\mathbf{W},\mathbf{H}} J_{\text{global}} + \mu_1 J_{\text{SPK}} + \mu_2 J_{\text{SES}} \quad \text{s.t.} \quad \mathbf{W} \geq 0,\ \mathbf{H} \geq 0 \tag{7}$$

which in turn leads to the multiplicative update rules for the dictionaries $\mathbf{W}_{\text{SPK}}^{(cs)}$ and $\mathbf{W}_{\text{SES}}^{(cs)}$ that are given in equations (12) and (13), respectively. The update rules for $\mathbf{W}_{\text{RES}}^{(cs)}$ and for the activations ($\mathbf{H}^{(cs)}$) are left unchanged (see [12]).

4. TASK-DRIVEN NMF BASED DICTIONARY LEARNING

TDL [13] has recently been applied with nonnegativity constraints to perform speech enhancement [15] or to acoustic scene classification, where temporally integrated projections are classified with multinomial logistic regression [14]. In this paper we extend the latter approach to the GNMF case.

4.1. Task-driven NMF

The general idea of nonnegative TDL or task-driven NMF (TNMF) is to unite the dictionary learning with NMF and the training of the classifier in a joint optimization problem [15, 14]. Influenced by the classifier, the basis vectors are encouraged to explain the discriminative information in the data while keeping a low reconstruction cost. The TNMF model first considers the optimal projections $\mathbf{h}^\star(\mathbf{v}, \mathbf{W})$ of the data points $\mathbf{v}$ on the dictionary $\mathbf{W}$, which are defined as solutions of the nonnegative elastic-net problem [18], expressed as:

$$\mathbf{h}^\star(\mathbf{v}, \mathbf{W}) = \min_{\mathbf{h} \in \mathbb{R}_+^K} \frac{1}{2}\|\mathbf{v} - \mathbf{Wh}\|_2^2 + \lambda_1\|\mathbf{h}\|_1 + \frac{\lambda_2}{2}\|\mathbf{h}\|_2^2; \tag{8}$$

where λ_1 and λ_2 are nonnegative regularization parameters. Given each data segment $\mathbf{V}^{(l)}$ of length M frames, associated with a label y in a fixed set of labels $\mathcal{Y}$, we want to classify the mean of the projections of the data points $\mathbf{v}^{(l)}$ belonging to the segment l, such that $\mathbf{V}^{(l)} = [\mathbf{v}_0^{(l)}, ..., \mathbf{v}_{M-1}^{(l)}]$. We define $\hat{\mathbf{h}}^{(l)}$ as the averaged projection of $\mathbf{V}^{(l)}$ on the dictionary, where $\hat{\mathbf{h}}^{(l)} = \frac{1}{M}\sum_{m=0}^{M-1} \mathbf{h}^\star(\mathbf{v}_m^{(l)}, \mathbf{W})$. The corresponding classification loss (here using multinomial logistic regression) is defined as $l_s(y, \mathbf{A}, \hat{\mathbf{h}}^{(l)})$, where $\mathbf{A} \in \mathcal{A}$ are the parameters of the classifier. The TNMF problem is then expressed as a joint minimization of the expected classification loss over $\mathbf{W}$ and $\mathbf{A}$:

$$\min_{\mathbf{W} \in \mathcal{W}, \mathbf{A} \in \mathcal{A}} f(\mathbf{W}, \mathbf{A}) + \frac{\nu}{2}\|\mathbf{A}\|_2^2, \tag{9}$$

with

$$f(\mathbf{W}, \mathbf{A}) = \mathbb{E}_{y, \mathbf{V}^{(l)}}[l_s(y, \mathbf{A}, \hat{\mathbf{h}}^{(l)}(\mathbf{v}^{(l)}, \mathbf{W}))]. \tag{10}$$

Here, $\mathcal{W}$ is defined as the set of nonnegative dictionaries containing unit l_2-norm basis vectors and ν is a regularization parameter on the classifier parameters, meant to prevent over-fitting. The problem in equation (10) is optimized with mini-batch stochastic gradient descent as described in the paper of Bisot *et al.* [14].

4.2. Task-driven GNMF

In task-driven GNMF (TGNMF) we propose to perform jointly the dictionary learning based on GNMF [12] and the training of a multinomial logistic regression. The dictionary $\mathbf{W}$ is then the concatenation of all the sub-dictionaries $\mathbf{W}^{(cs)}$ and the optimal projections $\mathbf{h}^\star(\mathbf{v}, \mathbf{W})$ are the solutions of (8).

$$\mathbf{W}_{\mathrm{SPK}}^{(cs)} \leftarrow \mathbf{W}_{\mathrm{SPK}}^{(cs)} \odot \frac{\mathbf{V}^{(cs)}{\mathbf{H}_{\mathrm{SPK}}^{(cs)}}^T + \frac{\mu_1}{2} \sum\limits_{\substack{s_1 \in \mathcal{S}_c \\ s_1 \neq s}} \mathbf{W}_{\mathrm{SPK}}^{(cs_1)}}{\mathbf{W}^{(cs)}\mathbf{H}^{(cs)}{\mathbf{H}_{\mathrm{SPK}}^{(cs)}}^T + \frac{\mu_1}{2}\left(\mathrm{Card}(\mathcal{S}_c) - 1\right)\mathbf{W}_{\mathrm{SPK}}^{(cs)}} \tag{12}$$

$$\mathbf{W}_{\mathrm{SES}}^{(cs)} \leftarrow \mathbf{W}_{\mathrm{SES}}^{(cs)} \odot \frac{\mathbf{V}^{(cs)}{\mathbf{H}_{\mathrm{SES}}^{(cs)}}^T + \frac{\mu_2}{2} \sum\limits_{\substack{c_1 \in \mathcal{C}_s \\ c_1 \neq c}} \mathbf{W}_{\mathrm{SES}}^{(c_1 s)}}{\mathbf{W}^{(cs)}\mathbf{H}^{(cs)}{\mathbf{H}_{\mathrm{SES}}^{(cs)}}^T + \frac{\mu_2}{2}\left(\mathrm{Card}(\mathcal{C}_s) - 1\right)\mathbf{W}_{\mathrm{SES}}^{(cs)}} \tag{13}$$

Including the similarity constraints (5) and (6), the TGNMF is thus expressed as the minimization of the following problem:

$$\min_{\mathbf{W}\in\mathcal{W},\mathbf{A}\in\mathcal{A}} f(\mathbf{W},\mathbf{A}) + \frac{\nu}{2}\|\mathbf{A}\|_2^2 + \mu_1 J_{\mathrm{SPK}} + \mu_2 J_{\mathrm{SES}}, \tag{11}$$

with $f(\mathbf{W},\mathbf{A})$ as defined above. The problem is again optimized with mini-batch stochastic gradient descent. However, as opposed to the previous algorithm, for each data point $\mathbf{v}$ belonging to a particular $\mathbf{V}^{(cs)}$, only the corresponding sub-dictionaries ($\mathbf{W}^{(cs)}$) are updated, whereas the other dictionaries are left unchanged in order to match the GNMF adaptation scheme [12].

5. EXPERIMENTS

5.1. Corpus

The approach presented here is evaluated on a subset of ESTER, a corpus for automatic speech recognition composed of data recorded from broadcast radio [19]. The subset of ESTER is composed of non-overlapping speech and decomposes as follows: 6 hours and 11 minutes of training data and 3 hours 40 minutes of test data both distributed among 95 speakers. The amount of training data per speaker ranges from 10 seconds to 6 minutes [12]. One target of this article is to act as a proof of concept for supervised dictionary learning methods applied to speaker identification. This corpus is small enough to allow for testing several configuration and reasonably large to perform statistically significant experiments. It is therefore suited for the task targeted in this article.

5.2. I-vector baseline

A baseline I-vector-based system is trained with the LIUM speaker diarisation toolkit [20]. The acoustic features are computed with YAAFE [21]. They are 20 mel frequency cepstral coefficients (MFCC) [22], including the energy coefficient. They are computed on 32 ms frames with 16 ms overlap. The MFCC are augmented with their first and second derivatives to form a 60-dimensional feature vector. A universal background model (UBM) with 256 Gaussian components per acoustic feature is trained on the full training set and the dimension of the total variability space is set to 100. The parameter values are in the range of the values commonly found in the literature for datasets of similar size. Eigen factor radial normalisation (EFR) is applied on I-vectors before classification [4].

5.3. NMF-based feature learning

NMF-based systems are trained on GPGPU with an in-house software[1] exploiting the Theano toolbox [23]. The acoustic features are 132 constant-Q transform coefficients (CQT) [24] computed on 16 ms frames with YAAFE [21]. To cope with the well-known problem of non-uniqueness of the NMF solution, NMF and GNMF are initialised randomly 6 times and trained independently for 100 iterations. In each case, the factorisation with the lowest cost function value at the end of the training is selected to extract features. After preliminary tests, the number of components for the NMF has been set to $K = 100$. The number of components for each data portion of the GNMF is set to $K = 8$ ($K_{\mathrm{SPK}} = 4,\ K_{\mathrm{SES}} = 2, K_{\mathrm{RES}} = 2$). Only speaker-related bases and session-related bases are kept to project the data at runtime. There are 236 unique (speaker, session) couples, so the dimension of the feature vectors extracted with the GNMF is $K = 1416$. The weights μ_1 and μ_2 are scaled such that, respectively, for $\mu_1 = 1$ the contributions from (4) and (5) to (7) are equivalent, and for $\mu_2 = 1$ the contributions from (4) and (6) to (7) are equivalent. The features extracted with NMF are scaled to unit variance before classification. In the remained of this paper, GNMF applied without similarity constraints ($\mu_1 = 0$ and $\mu_2 = 0$) is denoted GNMF$_0$. Similarly, GNMF with similarity constraints ($\mu_1 = 0.4$ and $\mu_2 = 0.15$) is denoted GNMF$_c$.

5.4. Multinomial logistic regression

Normalised I-vectors and feature vectors extracted with NMF and GNMF are classified with a multinomial logistic regression performed with the scikit-learn toolkit [25]. The logistic regression is preferred to PLDA as the latter is known to perform quite poorly when the number of samples becomes small compared to the feature dimensionality, which is the case here. This approach has proven successful in our previous work [12].

5.5. Task-driven approaches

TNMF and TGNMF are applied to fine-tune the dictionaries obtained with the unsupervised NMF and GNMF described above[2]. The projections on the dictionary (corresponding to equation (8)) are computed using the *lasso* function from the *spams* toolbox [26]. The classifier is updated using one iteration of the scikit-learn [25] implementation of the multinomial logistic regression with the L-BFGS solver. The model is trained over $I = 5$ full passes over the data (epochs). When the initial dictionary is obtained with standard NMF ($K = 100$), the initial gradient update step is 0.0005 and the parameters for the elastic net problem are $\lambda_1 = 0.001$ and $\lambda_2 = 0.001$. When the initial dictionary is obtained with GNMF ($K = 1416$), the initial gradient update step is 0.0001 and the parameter for the elastic net problem are $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$. The decaying of the gradient steps over iterations follows the same heuristic as suggested

[1] Source code is available at `https://github.com/rserizel/groupNMF`

[2] Source code is available at `https://github.com/rserizel/TGNMF`

in [13]. The hyper parameters are obtained after performing a grid search over several reasonable values. After 5 epochs, the dictionaries are kept fixed and the classifier alone is trained for at most 50 epochs. In the remained of this paper, TGNMF applied without similarity constraints ($\mu_1 = 0$ and $\mu_2 = 0$) is denoted TGNMF$_0$. Similarly, TGNMF with similarity constraints ($\mu_1 = 0.0001$ and $\mu_2 = 0.0001$) is denoted TGNMF$_c$.

5.6. Performance evaluation

In order to mitigate the effect of the imbalance between speakers in the test set, the classification performance is measured with weighted F1-score [27] where the F1-score is computed for each class separately and weighted by the number of utterances in the class. Variations in identification performance are validated using the McNemar test [28] with significance level .05. In the remainder of the paper, unless stated otherwise explicitly, when a performance change is mentioned it is statistically significant.

F1-score performance obtained with the different approaches described above is presented in Table 1. Each column corresponds to a different initialisation method (NMF, GNMF$_0$ and GNMF$_c$). The first row (labeled **unsupervised**) presents the reference performance for each initialisation method, where the feature learning model and the classifier are learned independently. For the sake of simplicity, in the remainder of the paper these methods are referred to as unsupervised, as opposed to supervised methods (TNMF and TGNMF), even though some level of supervision is necessary for GNMF. The second row (labeled **TNMF**) presents the performance obtained when applying TNMF in a similar way as in Bisot et al. [14], initialised with the dictionaries obtained with NMF and GNMF. The last rows present the performance obtained when applying TGNMF$_0$ and TGNMF$_c$, initialised with the dictionaries obtained with GNMF.

5.7. Discussion

Two main tendencies can be observed from the results in Table 1. First, on small dictionaries (NMF with $K = 100$), TNMF allows for a large improvement compared to unsupervised methods and good performance. Secondly, TGNMF can sometimes provide large improvement reducing the performance difference between systems using initialisations with GNMF$_0$ and GNMG$_c$.

Unsupervised reference methods
The performance obtained with unsupervised methods tends to confirm previous findings where NMF (75.6%) is behind other systems and where the GNMF (81.7% with GNMF$_c$ or 80.7% GNMF$_0$) is better than the baseline I-vector system (76.1%). These systems also improve the performance compared to previous experiments from the authors with GNMF with generalised Kullback-Leibler divergence [17] applied on Mel-spectrums coefficients [12].

TNMF
Applying TNMF in a similar way as in Bisot et al. [14], initialised on the dictionaries learned with standard NMF allows for a large performance improvement (from 75.6% to 79.9%), whereas TNMF initialised with concatenated dictionaries obtained with GNMF leads to improvements that are not statistically significant. This could be due to the fact that the dictionaries are then too large and that one of the advantages of TNMF is that it is the most efficient when considering dictionaries smaller than those used with unsupervised methods.

Features	I-vector	NMF	GNMF$_0$	GNMF$_c$
Unsupervised	76.1%	75.6%	80.7%	81.7%
TNMF	–	79.9%	81.1%	81.9%
TGNMF$_0$	–	–	81.7%	82.1%
TGNMF$_c$	–	–	82.0%	**82.2%**

Table 1. Weighted F1-scores for speaker classification ($K = 100$ for NMF and $K = 1446$ for GNMF). Each column corresponds to a different initialisation method and each row corresponds to the method applied after the initialisation (for the first row no processing is done after the initialisation). The subscripts $_0$ and $_c$ correspond to method without and with constraints, respectively (see also 5.3, 5.5 and 5.6 for more detailed explanations).

TGNMF$_0$
GNMF$_0$ allows for focusing on learning some sub-dictionaries related to portions of the data originating from a specific speaker or session. This already proved effective on the unsupervised methods. This observation is confirmed when applying TGNMF$_0$ on dictionaries obtained with GNMF$_0$. TGNMF$_0$ then allows for a F1-score increase from 80.7% to 81.7%. The system obtains similar performance as the best reference system (GNMF$_c$), without exploiting the similarity constraints. The gain is less important when applying TGNMF$_0$ initialised with GNMF$_c$ where the annotations were already exploited to some extent.

TGNMF$_c$
Imposing similarity constraints during TGNMF helps improving the performance further, up to 82.2% when initialised with dictionaries obtained with GNMF$_c$. This is our best performance to date on this corpus. However, this is not significantly better than performance obtained with other TGNMF systems. This tends to indicate that both methods (GNMF and TGNMF) are to some extent redundant in the way to exploit the information from the annotations to structure the dictionaries and that we maybe reached a saturation point for these methods applied to speaker identification on rather small corpora such as the subset of ESTER. Future works should then include validation of these methods on larger corpora.

6. CONCLUSIONS

This paper presented supervised feature learning approaches for speaker identification including an approach that integrates GNMF into a TDL supervised framework. The goal was to capture both the speaker variability and the session variability while exploiting the discriminative learning qualities of the task-driven approach.

Evaluations on a subset of the ESTER corpus have shown that TNMF can allow for large improvements compared to unsupervised methods and good performance on small dictionaries obtained with NMF. When considering larger dictionaries obtained with GNMF, TGNMF allowed for focusing on training some sub-dictionaries related to portions of the data and taking into account speaker and session variability. Therefore TGNMF provided large improvement when initialised with dictionaries obtained with GNMF without similarity constraints and significant improvement when initialised on dictionaries obtained with GNMF with similarity constraints providing our best performance to date on the subset of ESTER.

7. REFERENCES

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[2] C. S. Greenberg, D. Bansé, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, T. Kinnunen, A. F. Martin, A. McCree, M. Przybocki, and D. A. Reynolds, "The NIST 2014 Speaker Recognition I-Vector Machine Learning Challenge," in *Proc. of Odyssey: The Speaker and Language Recognition Workshop*, 2014, pp. 224–230.

[3] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of I-vector length normalization in speaker recognition systems," in *Proc. of Interspeech*, 2011, pp. 249–252.

[4] P.-M. Bousquet, D. Matrouf, and J.-F. Bonastre, "Intersession Compensation and Scoring Methods in the i-vectors Space for Speaker Recognition.," in *Proc. of Interspeech*, 2011, pp. 485–488.

[5] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. of ICCV*, 2007, pp. 1–8.

[6] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization.," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[7] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Noise Robust Speaker Recognition with Convolutive Sparse Coding," in *Proc. of Interspeech*, 2015.

[8] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Similarity induced group sparsity for non-negative matrix factorisation," in *Proc. of ICASSP*, 2015, pp. 4425–4429.

[9] R. Saeidi, A. Hurmalainen, T. Virtanen, and A. Van Leeuwen, "Exemplar-based Sparse Representation and Sparse Discrimination for Noise Robust Speaker Identification," in *Proc. of Odyssey 2012: The Speaker and Language Recognition Workshop*, 2012.

[10] N. Seichepine, S. Essid, C. Fevotte, and O. Cappe, "Soft nonnegative matrix co-factorization," *IEEE Transactions on Signal Processing*, vol. 62, no. 22, pp. 5940–5949, 2014.

[11] H. Lee and S. Choi, "Group nonnegative matrix factorization for EEG classification," in *Proc. of AISTATS*, 2009, pp. 320–327.

[12] R. Serizel, S. Essid, and G. Richard, "Group nonnegative matrix factorisation with speaker and session variability compensation for speaker identification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5470–5474.

[13] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 791–804, 2012.

[14] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Feature Learning with Matrix Factorization Applied to Acoustic Scene Classification," HAL-archives ouvertes: working paper or preprint (hal-01362864), Sept. 2016.

[15] P. Sprechmann, A. M. Bronstein, and G. Sapiro, "Supervised non-euclidean sparse NMF via bilevel optimization with applications to speech enhancement," in *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on*, May 2014, pp. 11–15.

[16] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. of NIPS*, 2000, pp. 556–562.

[17] S. Kullback and R. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[18] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[19] G. Gravier, J.-F. Bonastre, E. Geoffrois, S. Galliano, K. Mc Tait, and K. Choukri, "ESTER, une campagne d'évaluation des systemes d'indexation automatique d'émissions radiophoniques en français.," in *Proc. of Journées d'Etude sur la Parole*, 2004.

[20] M. Rouvier, G. Dupuy, P. Gay, and E. Khoury, "An open-source state-of-the-art toolbox for broadcast news diarization," in *Proc. of Interspeech*, 2013.

[21] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "YAAFE, an easy to use and efficient audio feature extraction software," in *Proc. of ISMIR*, 2010, pp. 441–446.

[22] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[23] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio, "Theano: new features and speed improvements," Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

[24] J. Brown, "Calculation of a constant q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., "Scikit-learn: Machine learning in python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[26] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *The Journal of Machine Learning Research*, vol. 11, pp. 19–60, 2010.

[27] C. J. Van Rijsbergen, *Information Retrieval*, Butterworth-Heinemann, 1979.

[28] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.