GROUP NONNEGATIVE MATRIX FACTORISATION WITH SPEAKER AND SESSION VARIABILITY COMPENSATION FOR SPEAKER IDENTIFICATION

Romain Serizel, Slim Essid, Gaël Richard

LTCI, CNRS, Télécom ParisTech, Université Paris - Saclay, 75013, Paris, France

ABSTRACT

This paper presents a feature learning approach for speaker identification that is based on nonnegative matrix factorisation. Recent studies have shown that with such models, the dictionary atoms can represent well the speaker identity. The approaches proposed so far focused only on speaker variability and not on session variability. However, this later point is a crucial aspect in the success of the I-vector approach that is now the state-of-the-art in speaker identification.

This paper proposes a method that relies on group nonnegative matrix factorisation and that is inspired by the I-vector training procedure. By doing so the proposed approach intends to capture both the speaker variability and the session variability. Results on a small corpus prove that the proposed approach can be competitive with I-vectors.

***Index Terms*—** Nonnegative matrix factorisation, spectrogram factorisation, feature learning, speaker variability, 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 and if he/she is known, to find his/her identity. Applications of speaker identification are numerous, among which are speaker dependent automatic speech recognition and subject identification based on biometric information. The sentence pronounced by the subject can be unknown and the recordings can be of variable quality. The speaker identification then becomes a highly challenging problem.

Since their emergence almost five years ago, the I-vectors [1] have become the state-of-the-art approach for speaker identification [2]. A typical speaker identification system is composed of I-vector extraction, normalisation [3, 4] and classification with probabilistic linear discriminant analysis (PLDA) [5]. Research on the tandem I-vector/PLDA has focused a lot of attention during the past years and speaker identification systems have reached a high level of performance on databases such as those from the National Institute of Standards and Technology (NIST) [2, 6].

On the other hand, recent studies have shown that approaches such as nonnegative matrix factorisation (NMF) [7] can be successfully applied to spectrogram factorisation [8, 9, 10] or to multimodal co-factorisation [11] to retrieve speaker identity. These results tend to indicate that the activations of NMF dictionary atoms can represent well the speaker identity [10]. Besides, exploiting group sparsity on the activations has then proven to improve further the performance of NMF-based approaches [9]. NMF therefore offers a credible alternative to I-vectors that takes advantage of the intrinsic sparsity of speech [9, 12]. However, to our best knowledge, none of the approaches proposed until now take the recording sessions variability into account. Yet this is a crucial point in the success of I-vectors.

This paper proposes an approach to speaker identification that relies on group-NMF and that is inspired by the I-vector training procedure. Given data measured with several subjects, the key idea in group-NMF is to track inter-subject and intra-subject variations by constraining a set of common bases across subjects in the decomposition dictionaries. This has originally been applied to the analysis of electroencephalograms [13]. The approach presented here extends this idea and proposes to capture inter-speaker and inter-session variabilities by constraining a set of speaker-dependent bases across sessions and a set of session-dependent bases across speakers. This approach is inspired by the joint factor analysis [14] and I-vectors as it takes both speaker variability and session variability into account. In this sense, it differs from previous approaches based on NMF [8, 9, 12] that take only speaker variability into account. Besides, in these previous works similarity constraints were imposed on activations while in the approach proposed here the constraints are on the dictionaries.

The paper is organised as follows. The problem, the notations and the general NMF approach for speaker identification are introduced, in Section 2. The proposed approach is described in Section 3. Experiment results are presented in Section 4. Finally, conclusions and directions for future work are exposed in Section 5.

2. PROBLEM STATEMENT

2.1. Notations

Consider the (nonnegative) time-frequency representation of an audio signal $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ (this could be for example a mel-frequency spectrogram), where F is the number of frequency

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

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 $\mathrm{Card}(.)$, such that $\mathrm{Card}(\mathcal{C}) = C$ 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.

2.2. NMF with Kullback-Leibler divergence

The goal of NMF [7] 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. Given a separable divergence D, NMF model estimation can be formulated as the following optimisation problem:

$$\min_{\mathbf{WH}} D(\mathbf{V}|\mathbf{WH}) \quad \text{s.t.} \quad \mathbf{W} \geq 0,\ \mathbf{H} \geq 0.$$

When considering audio signals, D is often chosen to be the Kullback-Leibler divergence (denoted D_{KL} here) [15] or the Itakura-Saito divergence [16]. In most cases NMF problem is solved using a two-block coordinate descent approach. Each of the factors $\mathbf{W}$ and $\mathbf{H}$ is optimised alternatively. The subproblem in one factor can then be considered as a non-negative least square problem (NNLS) [17]. One of the approaches to solve these NNLS problems leads to the multiplicative update rules for the matrices $\mathbf{W}$ and $\mathbf{H}$, which can be expressed as follows for the D_{KL} [18, 19]:

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T\left[(\mathbf{WH})^{-1} \odot \mathbf{V}\right]}{\mathbf{W}^T\mathbf{1}} \tag{2}$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\left[(\mathbf{WH})^{-1} \odot \mathbf{V}\right]\mathbf{H}^T}{\mathbf{1}\mathbf{H}^T}; \tag{3}$$

where $\odot$ is the element-wise product (Hadamard product) and division and power are element-wise. $\mathbf{1}$ is a matrix of dimension $F \times N$ with all its coefficient equal to 1.

2.3. NMF for feature learning in speaker identification

In this paper, NMF is used for feature learning in a speaker identification framework. First, the factorisation is learnt on a training set and activations are used as input features to train a general classifier. The dictionaries $\mathbf{W}$ obtained on the training set are then used to extract features (activations) on a test set. These features are used as input to the general classifier to perform speaker identification.

3. GROUP NMF WITH SPEAKER AND SESSION SIMILARITY

In the approach presented above, the feature learning step is totally unsupervised and does not account for speaker variability or session variability. The approach introduced here intends to take these variabilities into account. It derives from group-NMF [13] and is inspired by exemplar-based approaches [8, 9]. The idea of a decomposition across speaker was originally used by Saeidi *et al.* [10] but session variability was not considered.

3.1. NMF on speaker utterances for speaker identification

In order to better model speaker identity, we now consider the portion of $\mathbf{V}$ recorded in a session s in which only the speaker c is active. This is denoted by $\mathbf{V}^{(cs)}$, its length is $N^{(cs)}$ and it can be decomposed according to (1):

$$\mathbf{V}^{(cs)} \approx \mathbf{W}^{(cs)}\mathbf{H}^{(cs)} \quad \forall\ (c,s) \in \mathcal{C} \times \mathcal{S}_c$$

under nonnegative constraints.

We define a global cost function which is the sum of all local divergences:

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

Each $\mathbf{V}^{(cs)}$ can be decomposed independently with standard multiplicative rules (2, 3). The bases learnt on the training set are then concatenated to form a global basis. The latter basis is then used to produce features on test sets.

3.2. Class and session similarity constraints

In order to take the session and speaker variabilities into account we propose to further decompose the dictionaries $\mathbf{W}$ similarly as what was proposed by Lee *et al.* [13]. The matrix $\mathbf{W}^{(cs)}$ can indeed be arbitrarily decomposed as follows:

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

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.

The first target is to capture speaker variability. This is related to finding vectors for the speaker bases ($\mathbf{W}^{(cs)}_{\text{SPK}}$) for each speaker c that are as close as possible across all the sessions in which the speaker is present, leading to the constraint:

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

$$\mathbf{W}_{\mathrm{SPK}}^{(cs)} \leftarrow \mathbf{W}_{\mathrm{SPK}}^{(cs)} \odot \frac{\left[(\mathbf{W}^{(cs)}\mathbf{H}^{(cs)})^{-1} \odot \mathbf{V}^{(cs)}\right] {\mathbf{H}_{\mathrm{SPK}}^{(cs)}}^T + \frac{\lambda_1}{2} \sum\limits_{\substack{s_1 \in \mathcal{S}_c \\ s_1 \neq s}} \mathbf{W}_{\mathrm{SPK}}^{(cs_1)}}{\mathbf{1}{\mathbf{H}_{\mathrm{SPK}}^{(cs)}}^T + \frac{\lambda_1}{2} \left(\mathrm{Card}(\mathcal{S}_c) - 1\right) \mathbf{W}_{\mathrm{SPK}}^{(cs)}} \quad (8)$$

$$\mathbf{W}_{\mathrm{SES}}^{(cs)} \leftarrow \mathbf{W}_{\mathrm{SES}}^{(cs)} \odot \frac{\left[(\mathbf{W}^{(cs)}\mathbf{H}^{(cs)})^{-1} \odot \mathbf{V}^{(cs)}\right] {\mathbf{H}_{\mathrm{SES}}^{(cs)}}^T + \frac{\lambda_2}{2} \sum\limits_{\substack{c_1 \in \mathcal{C}_s \\ c_1 \neq c}} \mathbf{W}_{\mathrm{SES}}^{(c_1 s)}}{\mathbf{1}{\mathbf{H}_{\mathrm{SES}}^{(cs)}}^T + \frac{\lambda_2}{2} \left(\mathrm{Card}(\mathcal{C}_s) - 1\right) \mathbf{W}_{\mathrm{SES}}^{(cs)}} \quad (9)$$

with $\|.\|^2$ the Euclidean distance and α_1 is the similarity constraint on speaker-dependent bases.

The second target is to capture session variability. This can be accounted for by finding vectors for the sessions bases ($\mathbf{W}_{\mathrm{SES}}^{(cs)}$) for each session s that are as close as possible across all the speakers that speak in the session, leading to the constraint:

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

where α_2 is the similarity constraint on session-dependent bases.

The vectors composing the residual bases $\mathbf{W}_{\mathrm{RES}}^{(cs)}$ are left unconstrained to represent characteristics that depend neither on the speaker nor on the session.

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

$$\min_{\mathbf{W},\mathbf{H}} J_{\mathrm{global}} + \lambda_1 J_{\mathrm{SPK}} + \lambda_2 J_{\mathrm{SES}} \quad \text{s.t.} \quad \mathbf{W} \geq 0,\ \mathbf{H} \geq 0 \quad (7)$$

which in turn leads to the multiplicative update rules for the dictionaries $\mathbf{W}_{\mathrm{SPK}}^{(cs)}$ and $\mathbf{W}_{\mathrm{SES}}^{(cs)}$ that are given in equations (8) and (9), respectively. We obtained these update rules using the well know heuristic which consists in expressing the gradient of the cost function (7) as the difference between a positive contribution and a negative contribution. The multiplicative update then has the form of a quotient of the negative contribution by the positive contribution. The update rules for $\mathbf{W}_{\mathrm{RES}}^{(cs)}$ are similar to the standard rules:

$$\mathbf{W}_{\mathrm{RES}}^{(cs)} \leftarrow \mathbf{W}_{\mathrm{RES}}^{(cs)} \odot \frac{\left[(\mathbf{W}^{(cs)}\mathbf{H}^{(cs)})^{-1} \odot \mathbf{V}^{(cs)}\right] {\mathbf{H}_{\mathrm{RES}}^{(cs)}}^T}{\mathbf{1}{\mathbf{H}_{\mathrm{RES}}^{(cs)}}^T}.$$

Note that the update rules for the activations ($\mathbf{H}^{(cs)}$) are left unchanged.

4. EXPERIMENTS

4.1. Experimental setup and corpus

The approach presented here is tested on a subset of the ESTER corpus [20], a radio broadcast corpus. Only speakers with at least 10 seconds of training data are selected from ESTER to compose the subset corpus. Speaker utterances are split in 10 seconds segments in order to obtain enough segments to train the back-end classifier. The amount of training data is limited to 6 minutes per speaker. When there is more than 6 minutes of speech for a speaker, 10 seconds segments are selected randomly to compose a 6 minutes subset. The resulting corpus is composed of 6 hours and 11 minutes of training data and 3 hours 40 minutes of test data both distributed among 95 speakers. The training data is extracted from the original ESTER training set and the test data is extracted from the original ESTER development set. This way, there is no overlapping session between the training set and the test set. The amount of training data per speaker ranges from 10 seconds to 6 minutes (Table 1). This small dataset is used for preliminary experiments and future work should include experiments with larger datasets such as NIST datasets.

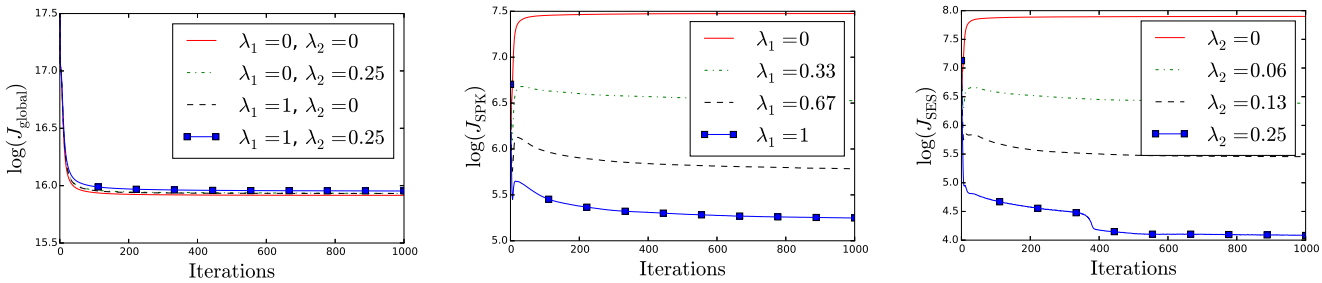
Duration	$<$ 1min	1min – 5min	$>$ 5min
Number of speakers	25	26	44

Table 1. Speakers repartition according to the amount of available training data.

A baseline I-vector-based system is trained with the LIUM speaker diarisation toolkit [21]. The acoustic features are computed with YAAFE [22]. They are 20 mel frequency cepstral coefficients (MFCC) [23], 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 features 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].

NMF-based systems are trained on GPGPU with an in-house software[1] based on the Theano toolbox [24]. The acoustic features are 64 mel-spectrum coefficients computed on 32 ms frames with 16 ms overlap. To cope with the

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



(a) KL-divergence

(b) Intra-speaker distance across sessions($\lambda_2 = 0$)

(c) Intra-session distance across speakers ($\lambda_1 = 0$)

Fig. 1. Convergence of the different criteria depending on the weights λ_1 and λ_2

well-known non-uniqueness of the NMF solution, NMF and group-NMF are initialised randomly 6 times and trained independently for 1000 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 is set to $K = 100$. The number of component for each bases of the group-NMF is set to $K = 8$ ($K_{\text{SPK}} = 4,\ K_{\text{SES}} = 2, K_{\text{RES}} = 2$) such that the size of the concatenated basis remains reasonably low. There are 236 unique couples (speaker, session) so the dimension of the feature vectors extracted with the group-NMF is 1888. The weights λ_1 and λ_2 are scaled such that for $\lambda_1 = 1$ the contributions from (4) and (5) to (7) are equivalent, respectively for $\lambda_2 = 1$ the contributions from (4) and (6) to (7) are equivalent. The features extracted with NMF are scaled to unit variance before classification.

Normalised I-vectors and feature vectors extracted with NMF are classified with a multinomial logistic regression. 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. In order to mitigate the effect of the imbalance between speakers in the test set, the classification performance is measured with weighted F1-score [25] where the F1-score is computed for each class separately and weighted by the number of utterances in the class. Both logistic regression and F1-scoring are performed with the scikit-learn toolkit [26]. Variations in identification performance are validated using the McNemar test [27] with significance levels .01 and .001.

4.2. Discussion

The first important test is to control that the constraints imposed on the speaker bases and the session bases do not degrade the stability of the NMF algorithm. Indeed, convergence can quickly become problematic when imposing constraints on NMF. The KL-divergence still varies uniformly even with constraints on the cost function (7) (Figure 1 (a)). Yet the constraints are effective at reducing the distance between the speaker bases (Figure 1 (b)) and between the sessions bases (Figure 1 (c)).

Features	I-vector	NMF	Group-NMF	
			$\lambda_1 1 = 0$ $\lambda_2 = 0$	$\lambda_1 = 0.33$ $\lambda_2 = 0.06$
F1-score	76.1%	70.7%	77.8%	**80.2%**

Table 2. Weighted F1-scores obtained for a classification with multinomial logistic regression.

In a second experiment, the systems described above and the I-vector baseline are compared on the subset of ESTER (Table 2). The group NMF has been tested for different values of the weight applied to the constraints and two different configurations have been selected. The first configuration is fully unconstrained ($\lambda_1 = 0$ and $\lambda_2 = 0$) and both constraints are active in the second configuration ($\lambda_1 = 0.33$ and $\lambda_2 = 0.06$). The first remark is that all systems perform reasonably well even if standard NMF is clearly behind the other approaches ($p < .001$). The unconstrained group-NMF and the I-vector approach perform similarly (the difference is not statistically significant). Imposing constraints on both the speaker bases and the session bases improves significantly the performance compared to the I-vector approach and the unconstrained group-NMF ($p < .01$ in both cases).

5. CONCLUSIONS

This paper introduced a new feature learning approach for speaker identification that is based on NMF. Recent works on exemplar based speaker identification have shown that dictionary atoms in an NMF system can represent well speaker identity. Capitalising on this statement, the authors proposed an approach based on group-NMF that is inspired by the state-of-the-art I-vector approach and tries to capture both speaker variability and session variability. The central idea is to impose similarity constraints on speaker-dependent bases and session-dependent bases in the decomposition dictionaries. The proposed approach has proven to be competitive with I-vectors on a small corpus and future works should include extensive tests on larger corpora and on a wider range of configurations.

6. 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] C. S. Greenberg, V. M. Stanford, A. F. Martin, M. Yadagiri, G. R. Doddington, J. J. Godfrey, J. Hernandez-Cordero, and F. Meade, "The 2012 NIST Speaker Recognition Evaluation," in *Proc. of Interspeech*, 2013, pp. 1971–1975.

[7] 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.

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

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

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

[11] 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.

[12] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Group Sparsity for Speaker Identity Discrimination in Factorisation-based Speech Recognition," in *Proc. of Interspeech*, 2012, pp. 2–5.

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

[14] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 1435–1447, 2007.

[15] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 2, no. 1, pp. 79–86, 1951.

[16] Fumitada Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, no. 1, pp. 67–72, 1975.

[17] N. Gillis, "The why and how of nonnegative matrix factorization," in *Regularization, Optimization, Kernels, and Support Vector Machines*, M. Signoretto J.A.K. Suykens and A. Argyriou, Eds., Machine Learning and Pattern Recognition Series, pp. 257 – 291. Chapman & Hall/CRC, 2014.

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

[19] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duche, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

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