Group Non-Negative Matrix Factorisation With Speaker And Session Similarity Constraints For Speaker Identification

Romain Serizel, Slim Essid, Gael Richard

GROUP NON-NEGATIVE MATRIX FACTORISATION WITH SPEAKER AND SESSION SIMILARITY CONSTRAINTS FOR SPEAKER IDENTIFICATION

Romain Serizel, Slim Essid, Gaël Richard

Institut Mines-Telecom, Telecom ParisTech, CNRS LTCI, Paris, FRANCE

ABSTRACT

This paper presents a feature learning approach for speaker identification that is based on non-negative matrix factorisation. Recent studies have shown that in methods such as non-negative matrix factorisation, the dictionary atoms can represent well the speaker identity and that Using speaker identity to induce group similarity can proven to improve further the performance. However, the approaches proposed so far focused only on speakers variability and not on sessions variability. However, this later point is a crucial aspect in the success of the I-vector approaches that is now the state-of-the-art in speaker identification.

This paper proposes an approach that relies on group-NMF and that is inspired that the I-vector training procedure. By doing so this approach intends to capture both the speaker variability and the session variability. Results on a small corpus prove the proposed approach to be competitive with the state-of-the-art I-vector approach.

Index Terms— Non-negative matrix factorisation, group similarity, spectrogram factorisation, 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 find his/her identity. Applications of speaker identification are numerous, among which speaker dependent automatic speech recognition and subject identification based on biometric information. In this later case, the sentence pronounced by the subject can be unknown and the recordings can be of various quality. Therefore the process of speaker identification can become highly challenging.

Since their emergence almost five years ago, the I-vectors [1] have become the state-of-the-art approach for speaker verification and by extension for speaker identification [2]. A typical speaker identification system is composed of I-vector extraction, I-vector normalisation [3, 4] and I-vector 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 verification systems have now reached a high level of performance on databases such as the National Institute of Standards and Technology (NIST) Speaker recognition evaluation (SRE) campaigns [2, 6].

On the other hand, recent studies have shown that approaches such as non-negative matrix factorisation [7] can be successfully applied to spectrogram factorisation [8, 9, 10] or to multimodal co-factorisation [11] to retrieve speaker identity. Therefore indicating that the activations of dictionary atoms can represent well the speaker identity [10]. Using speaker identity to induce group sparsity or groups similarity has then proven to improve further the performance of NMF-based approaches to speaker identification. NMF therefore offer a credible alternative to i-vectors that takes advantage of the intrinsic sparsity of speech [9, 12]. However, to our best knowledge, none of these approaches take the recording sessions information 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 that the I-vector training procedure. Given data measured with several subject, the key idea in group-NMF is to track inter-subjects and intra-subjects variations by constraining a set of common bases across subjects in the decomposition dictionaries [13]. The approach presented here extends this idea and proposes to capture inter-speakers and inter-sessions variabilities by constraining a set of speaker depend bases across sessions and a set of sessions dependent bases across speakers. This approach is inspired by 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 takes 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 dictionary.

The paper is organised as follows. The problem, the notations and the general NMF approach for speaker recognition are introduced, in Section 2. The proposed approach is described in Section 3. Experiment results on a toy example are presented in Section 4. Finally, conclusions are exposed in Section 5.

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

2. PROBLEM STATEMENT

2.1. Notations

We consider the (positive) 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 recordings 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 class and $\mathcal{S}$ the set of sessions.

$$c \in \mathcal{C} = [\![1\ ;\ C]\!] \text{ and } s \in \mathcal{S} = [\![1\ ;\ S]\!]$$

Let $\mathcal{C}^s$ denote the subset of speakers that appear in the session s and $\mathcal{S}^c$ the subset of session in which the speaker c appears.

$$\mathcal{S}^c \subset \mathcal{S} \text{ and } \mathcal{C}^s \subset \mathcal{C}$$

In the remainder of this paper, superscript c and s will denotes the current speaker and session, respectively.

2.2. NMF with Kullback-Leibler divergence

The goal of NMF [7] is to find a factorisation of $\mathbf{V}$ of the form:

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

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

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

When considering audio signals, D is often chosen to be the Kullback-Leibler divergence (denoted D_{KL} here) [14]. The multiplicative update rules for the matrices $\mathbf{W}$ and $\mathbf{H}$ can then be expressed as follows [15, 16]:

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

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

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 features learning in speaker recognition

In this paper, NMF is used for feature learning in a speaker recognition framework. The factorisation is first learnt on a training set and activations are used as input feature 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] is inspired by exemplar-based approaches [8, 9]. The idea of a decomposition across speaker was originally used in Saeidi *et al.* [10] but session variability was not considered since.

3.1. NMF on speaker utterances for speaker recognition

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

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

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. This later basis is 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 in Lee *et al.* [13]. The matrix $\mathbf{W}^{(cs)}$ can indeed be arbitrarily decomposed as follows:

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

with

$$K_{\text{SPK}} + K_{\text{SES}} + K_{\text{RES}} = K$$

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

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

The second target is to capture session variability. This in turn is similar to finding vectors for the sessions bases $\mathbf{W}_{\text{SES}}^{(cs)}$

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

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

$$\mathbf{W}_{\text{RES}}^{(cs)} \leftarrow \mathbf{W}_{\text{RES}}^{(cs)} \odot \frac{\left[(\mathbf{W}^{(cs)}\mathbf{H}^{(cs)})^{-1} \odot \mathbf{V}^{(cs)}\right] {\mathbf{H}_{\text{RES}}^{(cs)}}^T}{\mathbf{1}{\mathbf{H}_{\text{RES}}^{(cs)}}^T} \tag{10}$$

that are as close as possible across each session s across all the speaker that speaks in the session, leading to the constraint:

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

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

Minimizing the global divergence (4) subject to constraints (5) and (6) results in the following dual problem:

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

which in turn leads to the multiplicative update rules for the dictionaries ($\mathbf{W}^{(cs)}$) that are given in equations (8-10). Note that the update rules for the activations ($\mathbf{H}^{(cs)}$) are left unchanged.

4. EXPERIMENTS

4.1. Experimental setup and corpus

This paper is intended mainly as a proof of concept. The approach presented here is tested on a toy example: a subset of the ESTER corpus [17]. Only speaker with at least 10 seconds of training data are selected from ESTER to compose the subset corpus. Speakers 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 amount of training data per speaker ranges from 10 seconds to 6 minutes (Table 1).

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-vectors system is trained with LIUM speaker diarisation toolkit [18]. The acoustic features are 20 mel frequency cepstral coefficients (MFCC) [19], including the energy coefficient. They are computed on 32ms frames with 16ms overlap. The MFCC are augmented with their first and second derivatives to form a 60-dimensional features vector. They are computed with Yaafe [20]. An 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. Eigen factor radial normalisation is applied on I-vectors before classification [4].

The acoustic features for NMF based systems are 64 mel-spectrum coefficients computed on 32ms frames with 16ms overlap. NMF and group-NMF are initialised randomly six times and trained independently for 1000 iterations. In each case, the factorisation with the lowest cost function at the end of the training is selected to extract features. All NMF-based systems are trained on GPGPU with an in-house software[1] based on Theano toolbox [21]. The number of components is set to $K = 100$ and to ($K_{\text{SPK}} = 4,\ K_{\text{SES}} = 2, K_{\text{RES}} = 2$) for the NMF and the group-NMF, respectively. There are 236 unique couples (speaker, session) so the dimension of the features vectors extracted with the group-NMF is 1888. The weights λ_1 and λ_2 are normalised by the values of the cost functions (4), (5) and (6) at convergence for the unconstrained case. This way for $\lambda_1 = 1$ the contribution from (4) and (5) to (7) are equivalent, respectively $\lambda_2 = 1$ and the contributions from (4) and (6). It does not make sense to apply EFR to features extracted from NMF, therefore these features are

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

λ_1 \ λ_2	0	0.06	0.12	0.25
0	77.8%	76.5%	76.0%	76.7%
0.33	75.6%	80.2%	78.9%	79.7%
0.67	74.1%	77.3%	77.4%	75.1%
1.33	76.6%	74.7%	79.4%	80.5%

Table 2. Weighted F1-scores obtained for different values of λ_1 and λ_2.

only scaled to unit variance before classification.

Normalised I-vectors and features vectors extracted with NMF are classified with a multinomial logistic regression. The logistic regression is preferred to PLDA as the later is known to perform quite poorly when the number of samples becomes small compared to the features 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 [22] 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-scores are performed with the scikit learn toolkit [23]. Variations in identification performance were validated using the McNemar test [24] considered significance levels were .01 and .001.

4.2. Discussion

The first experiment is to control that the constraints imposed on the speaker bases $\mathbf{W}_{\text{SPK}}^{(cs)}$ and the sessions bases $\mathbf{W}_{\text{SES}}^{(cs)}$ does not affect the stability of the NMF algorithm. Imposing constraints on the costs function (7) does not seem to affect the convergence of the global KL-divergence (Figure 1 (a)). However, the constraints are effective at reducing the distance between the speaker bases (Figure 1 (b)) and the sessions bases(Figure 1 (c)), respectively.

In a second experiment the proposed approach is tested for different value of the weight applied to the constraints. Weighted F1-score performance is presented in Table 2. A few trends appear on this table. Firstly it seems clear now that imposing constraint on the speaker bases $\mathbf{W}_{\text{SPK}}^{(cs)}$ and the sessions bases $\mathbf{W}_{\text{SES}}^{(cs)}$ does have an impact on the performance of the speaker identification. Secondly, it appears that there is a trade-off between the weight λ_1 and λ_2. Indeed, for a fixed λ_1, the performance reaches a maximum for a particular value of λ_2. Increasing λ_2 beyond this value results in a performance degradation.

Finally, the systems described above have been tested on the subset the ESTER. Table 3 present the performance of the systems. Two different configurations are considered for the group-NMF approach. The first configuration is fully unconstrained ($\lambda_1 = 0$ and $\lambda_2 = 0$). Both constraints are active in the second configuration ($\lambda_1 = 0.33$ and $\lambda_2 = 0.06$).

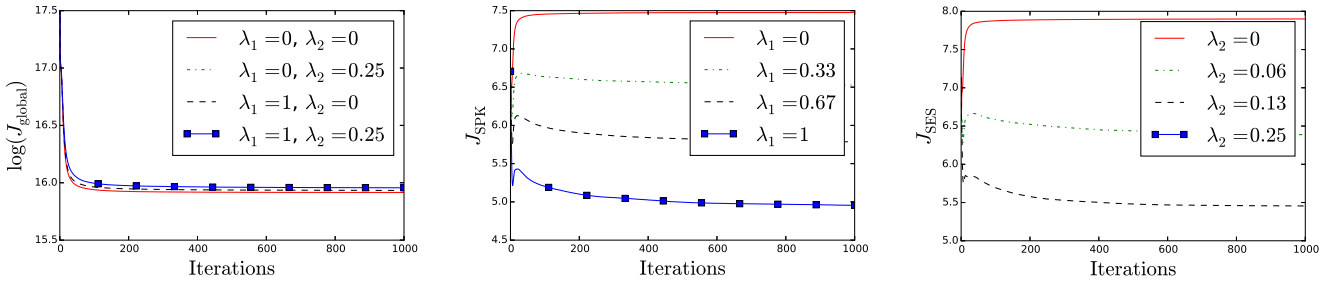
Features	I-vector	NMF	Group-NMF	
			$\lambda_1 1 = 0$ $\lambda_2 = 0$	$\lambda_1 = 0.33$ $\lambda_2 = 0.07$
F1-score	76.1%	70.7%	77.8%	80.2%

Table 3. Weighted F1-scores obtained for a classification with multinomial logistic regression.

The first remarks is that all systems perform reasonably well even if standard NMF is clearly behind the other approaches ($p < .001$). The unconstrained NMF and the I-vector approach perform similarly (the difference is not statistically significant). Imposing constraints on both the speaker bases $\mathbf{W}_{\text{SPK}}^{(cs)}$ and the sessions bases $\mathbf{W}_{\text{SES}}^{(cs)}$ 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 recognition have shown that dictionary atoms in a 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 speakers variability and sessions variability. The central idea is to impose similarity constraints on speaker bases and sessions bases in the decomposition dictionaries. The proposed approach as proven to be competitive with I-vector on a small corpus and future works should include extensive tests on larger corpora and on a wider range of configurations.



(a) KL-divergence (b) Speakers distance ($\lambda_2 = 0$) (c) Sessions distance ($\lambda_1 = 0$)

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

6. REFERENCES

[1] Najim Dehak, Patrick J Kenny, Reida Dehak, Pierre Dumouchel, and Pierre 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] Craig S Greenberg, Désiré Bansé, George R Doddington, Daniel Garcia-Romero, John J Godfrey, Tomi Kinnunen, Alvin F Martin, Alan McCree, Mark Przybocki, and Douglas A Reynolds, "The NIST 2014 Speaker Recognition i-Vector Machine Learning Challenge," *In Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[3] D. Garcia-Romero and Carol Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 249–252, 2011.

[4] Pm Bousquet, D Matrouf, and Jf Bonastre, "Intersession Compensation and Scoring Methods in the i-vectors Space for Speaker Recognition.," *Interspeech*, 2011.

[5] Simon Prince, Peng Li, Yun Fu, Umar Mohammed, and James Elder, "Probabilistic Models for Inference about Identity.," *IEEE transactions on pattern analysis and machine intelligence*, , no. 1, pp. 144–157, May 2011.

[6] Craig S Greenberg, Vincent M Stanford, Alvin F Martin, Meghana Yadagiri, George R Doddington, John J Godfrey, Jaime Hernandez-Cordero, and Fort Meade, "The 2012 NIST Speaker Recognition Evaluation," in *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] Antti Hurmalainen, Rahim Saeidi, and Tuomas Virtanen, "Noise Robust Speaker Recognition with Convolutive Sparse Coding," 2015.

[9] Antti Hurmalainen, Rahim Saeidi, and Tuomas Virtanen, "Similarity induced group sparsity for non-negative matrix factorisation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2015, pp. 4425–4429, IEEE.

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

[11] Nicolas Seichepine, Slim Essid, Cedric Fevotte, and Olivier Cappe, "Soft nonnegative matrix co-factorization," *IEEE Transactions on Signal Processing*, vol. PP, no. 99, 2014.

[12] Antti Hurmalainen, Rahim Saeidi, and Tuomas Virtanen, "Group Sparsity for Speaker Identity Discrimination in Factorisation-based Speech Recognition," *INTERSPEECH*, , no. 2, pp. 2–5, 2012.

[13] H Lee and S Choi, "Group nonnegative matrix factorization for EEG classification," *International Conference on Artificial ...*, 2009.

[14] Solomon Kullback and Richard A Leibler, "On information and sufficiency," *The annals of mathematical statistics*, pp. 79–86, 1951.

[15] Daniel D Lee and H Sebastian Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

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

[17] 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.," *Proc. Journees d'Etude sur la Parole (JEP)*, 2004.

[18] Mickael Rouvier, G Dupuy, Paul Gay, and Elie Khoury, "An open-source state-of-the-art toolbox for broadcast news diarization," ..., p. 5, 2013.

[19] 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, 1980.

[20] Benoit Mathieu, Slim Essid, Thomas Fillon, Jacques Prado, and Gaël Richard, "Yaafe, an easy to use and efficient audio feature extraction software," in *Proceedings of the 11th International Society for Music Information Retrieval Conference*, Utrecht, The Netherlands, August 9-13 2010, pp. 441–446.

[21] Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, and Yoshua Bengio, "Theano: new features and speed improvements," Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

[22] C. J. Van Rijsbergen, "Information Retrieval," Jan. 1979.

[23] 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.

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