Total Variability Space for LDA-based multi-view text categorization

Mohamed Morchid, Mohamed Bouallegue, Richard Dufour, Georges Linarès, Renato de Mori

Compact Multi-View Representation of Documents based on the Total Variability Space

Article *in* IEEE/ACM Transactions on Audio, Speech, and Language Processing · August 2015

DOI: 10.1109/TASLP.2015.2431854

6 authors, including:

Mohamed Morchid

Université d´Avignon et des Pays du Vaucluse

28 PUBLICATIONS **40** CITATIONS

SEE PROFILE

Richard Dufour

Université d´Avignon et des Pays du Vaucluse

44 PUBLICATIONS **82** CITATIONS

SEE PROFILE

Georges Linarès

Université d´Avignon et des Pays du Vaucluse

153 PUBLICATIONS **465** CITATIONS

SEE PROFILE

Renato De Mori

McGill University and University of Avignon

295 PUBLICATIONS **3,609** CITATIONS

SEE PROFILE

Total Variability Space for LDA-based multi-view text categorization

Mohamed Morchid*, Mohamed Bouallegue*, Richard Dufour*,
Georges Linarès* and Renato De Mori*, *Fellow, IEEE*
*Laboratoire d'Informatique d'Avignon (LIA), University of Avignon, France
†McGill University, School of Computer Science, Montreal, Quebec, Canada

Mapping text document into LDA-based topic-space is a classical way to extract high level representation of text documents. Unfortunatly, LDA is higly sensitive to hyper-parameters related to class number or word and topic distribution, and there is not any systematic way to prior estimate optimal configurations. Morover, various hyperparameter configurations offer complementary views on the document.

In this paper, we propose a method based on a two-step process that, first, expands representation space by using a set of topic spaces and, second, compacts representation space by removing poorly relevant dimensions. These two steps are based respectivelly on multi-view LDA-based representation spaces and factor-analysis models. This model provides a view-independant representation of documents while extracting complementary information from a massive multi-view representation.

Experiments are conducted on the DECODA conversation corpus and Reuters-21578 textual dataset. Results show the effectiveness of the proposed multi-view compact representation paradigm. The proposed categorization system reaches an accuracy of 86.9% and 86.5% respectively with manual and automatic transcriptions of conversations, and a macro-F1 of 80% during a classification task of the well-known studied Reuters-21578 corpus, with a significant gain compared to the baseline (best single topic space configuration), as well as methods and document representations previously studied.

***Index Terms*—Latent Dirichlet Allocation, Factor analysis, C-vector, Classification**

I. Introduction

One of the most efficicent way to process noisy text documents consists in mapping word-level surface forms into semantic spaces, where documents are represented by meaningfull abstract features.

Numerous unsupervised methods for topic-space estimation were proposed in the past, mostly based on the extaction of interesting regularities in huge text corpus. The Latent Dirichlet Allocation (LDA) [1] was largely used for text mining, speech analytics or information retrieval tasks; one of its main drawbacks is the tuning of the model, that involves various meta-parameters such as the number of classes (that determines the model granularity), word distribution methods, temporal spans... Performance of systems that use topic models can then be quite unstable if the decision process is highly dependent on these meta-parameters.

Classically, this abstract representation involves the selection of a number of classes composing the topic space (n) as well as the LDA hyper-parameters (α and β). The hyper-parameters α and β control both the topic distribution for each document and the word distribution into each class of the topic space itself. The number of classes n contained into the topic space control the "granularity" of the model from a general topic-based representation (few number of classes into the model) to a relatively precise representation (large number of classes). Finding the optimal parameters is crucial since topic model perplexity, that expresses its quality, is highly dependent to these features. Moreover, the multi-theme context of the proposed study implies a more complex dialogue representation [2].

In this paper, we tackle these two drawbacks by using multiple topic spaces obtained by varying the LDA hyper-parameters α, β and the topic number n. Each of these space offers a specific view on the documents and our goal, at this point, is to extract relevant and complementary information for the large set of diferent views. A potential issue with such a massive multi-view approach is due to the diversity of views, wich introduces both a *relevant* variability needed to represent different contexts of the document, and a *noisy* variability related to topic spaces processing. Thus, a topic-based representation of a document is built from the document content itself and the mapping process of a document into several topic spaces generates a noisy variability related to the difference of the document and each class content. In the same way, the relevant variability is from the common content between the document and the classes composing the topic space.

We propose to reduce this noisy variability by compacting multiple views of documents using factor analysis technique. Factor analysis is a very old data-analysis method that was successfully applied fisrt to speaker identification and, latter, generalized to various speech and audio categorization tasks.

In this field, the factor analysis paradigm is used as a decomposition model that enables to separate the representation space into two subspaces containing respectively useful and useless information. The general Joint Factor Analysis (JFA) paradigm [3] considers multiple variabilities that may be cross-dependent. Thereby, JFA representation allows to compen-

Manuscript received December 1, 2012; revised December 27, 2012. Corresponding author: M. Morchid (email: mohamed.morchid@univ-avignon.fr).

sate the variability within sessions of a same speaker. This representation is an extension of the GMM-UBM (Gaussian Mixture Model-Universal Background Model) models [4]. The authors in [5] extract, from the GMM super-vector, a compact super-vector called an i-vector. The aim of the compression process (*i.e.* i-vector extraction) is to represent the super-vector variability in a low dimensional space. Although this compact representation is widely used in speaker recognition systems, this method has been little used in the field of text classification.

In this paper, we propose to apply factor analysis to compensate noisy variabilities due to the multiplication of LDA models when varying all LDA hyper-parameters. We also propose to evaluate this approach on two different classification tasks using respectively automatic transcriptions, obtained from an Automatic Speech Recognition (ASR) system, and usual textual documents. Two classification tasks are then considered: the theme identification of RATP call centre (Paris Public Transportation Authority) dialogues [6] and the Reuters-21578 (ModApte split) classification task [7].

The intuition behind this study is that varying LDA hyper-parameters α and β should allow us to obtain an *optimal* topic-based representation of the document, while the multiple views of a given document are obtained by varying the number of classes into the topic space. Indeed, when one varies LDA hyper-parameters, the topic space structure is not deeply modified, which is the case when the number of classes is changed.

Furthermore, a normalization approach to condition document representations (multi-model and i-vector) is proposed. The two methods showed improvements for speaker verification: within Class Covariance Normalization (WCCN) [5] and Eigen Factor Radial (EFR) [8]. The last one includes length normalization [9]. Both of these methods dilate the total variability space as the mean to reduce the within class variability. In our multi-model representation, the within class variability is redefined according to both document content (vocabulary) and topic space characteristics (words distribution among the topics). Thus, the speaker is represented by a theme, and the speaker session is a set of topic-based representations (frames) of a document (session).

The paper is organized as follows. Section II presents previous related works. The document representation is described in Section III. Section IV introduces the i-vector compact representation and presents its application to text documents. Sections V and VI report experiments and results. The last section concludes and proposes some perspectives.

II. Related work

Considerable research have been proposed to combine topic related information with n-gram models [10], [11], [12], [13]. The basic idea of these approaches is to exploit the differences of word n-gram distributions across topics. That is, first the whole training data is separated into several topic-specific clusters, and then topic-specific LM are built using the topic-specific data. One problem of this approach is data fragmentation, which results in the data sparseness problem. In order to remedy the data sparseness problem, linear interpolation (or LM mixture) has been applied.

Several methods have been proposed considering that word n-grams have different word probability distributions in different topics and represent a document with a mixture of topic language models. These methods demonstrated their performance on various tasks, such as sentence [14] or keyword [15] extraction.

Considering documents as bags-of-words [16], Latent Dirichlet Allocation (LDA) [1] was proposed as a new method for obtaining word probabilities as mixtures of word distributions in hidden topics. PLSA and LDA models have been shown to generally outperform LSA on IR tasks [17]. Furthermore, probabilities of a hidden topic given a document can be computed with LDA providing topic classification features that capture word dependencies related to the semantic contents of a given conversation.

Supervised LDA [18] has been proposed in the context of multi-label topic classification to estimate word and topic label probabilities given a training corpus annotated with-label data. In all the approaches considered above, the choice of the number of topics is empirical. Many studies have proposed suggestions for solving this problem. Authors in [19] proposed to use a Singular Value Decomposition (SVD) to represent the words of the vocabulary. This method has to be evaluated with the Kullback-Liebler divergence metric for each topic space. It is not rigorous and time consuming.

Authors in [20] proposed to use a Hierarchical Dirichlet Process (HDP) method to find the "right" number of topics by assuming that the data have a hierarchical structure. In [21], authors presented a method to *learn* the right depth of an ontology depending of the number of topics of LDA models. The study presented by [22] is quite similar to [20]. The authors consider as the right number of topics, the average correlation between pairs of topics at each stage of the process. All these methods assume that a document can have representations in only one hidden space which is limited by the fact that classes of a LDA hidden space are correlated [23]. Moreover, authors in [24] consider a class as a node of an acyclic graph and as a distribution over other classes contained in the same topic space.

We proposed some studies to show the contribution of a compact dialogue representation from multiple views, based on the i-vector framework [25], [26]. In [25], we proposed to represent a dialogue in a set of topic spaces learned from a LDA algorithm, by varying the number of classes contained into the LDA topic space. Then, this multiple representation of the dialogue is compacted with the use of the i-vector framework. We proposed in [26] to learn a set of LDA topic spaces by varying the α hyper-parameter, which controls topic distribution for each document contained into the training corpus as well as the documents contained into the validation set. This last distribution is obtained during the inference process with the use of the Gibbs Sampling algorithm [27]. These studies evaluate the impact of LDA hyper-parameters separately and use only noisy transcriptions (dialogues) obtained from an automatic speech recognition system (ASR).

III. Multi-view representation of documents in a homogeneous space

The approach considered in this paper focuses on modeling the variability between different documents expressing the same theme t[1]. For this purpose, it is important to select relevant features that represent semantic contents for the theme of a document. An attractive set of features for capturing possible semantically relevant word dependencies is obtained with LDA [1], as described in Section II.

Given a training set of conversations D, a hidden topic space is derived and a conversation d is represented by its probability in each topic of the hidden space. Estimation of these probabilities is affected by a variability inherent to the estimation of the model parameters. If many hidden spaces are considered and features are computed for each hidden space, it is possible to model the estimation variability together with the variability of the linguistic expression of a theme by different speakers in different real-life situations. Even if the purpose of the application is theme identification and a training corpus annotated with themes is available, supervised LDA [27] is not suitable for the proposed approach. LDA is used only for producing different feature sets involved in statistical variability models.

In order to estimate the parameters of different hidden spaces, a set of discriminative words V is constructed as described in [2]. Each theme t contains a set of specific words. Note that the same word may appear in several discriminative word sets. All the selected words are then merged without repetition to form V. The entire application vocabulary is used for estimating the hidden spaces while only the words of the discriminative vocabulary are used for integrating the features obtained in the hidden spaces.

Several techniques, such as Variational Methods [1], Expectation-propagation [28] or Gibbs Sampling [27], have been proposed for estimating the parameters describing a LDA hidden space. Gibbs Sampling is a special case of Markov-chain Monte Carlo (MCMC) [29] and gives a simple algorithm for approximate inference in high-dimensional models such as LDA [30]. This overcomes the difficulty to directly and exactly estimate parameters that maximize the likelihood of the whole data collection defined as:

$$p(W|\overrightarrow{\alpha},\overrightarrow{\beta}) = \prod_{w \in W} p(\overrightarrow{w}|\overrightarrow{\alpha},\overrightarrow{\beta}) \quad (1)$$

for the whole data collection W knowing the Dirichlet parameters $\overrightarrow{\alpha}$ and $\overrightarrow{\beta}$.

Gibbs Sampling makes it possible to estimate the LDA parameters in order to represent a new document d with the r^{th} topic space of size n, and to obtain a feature vector $V_d^{z^r}$ of the topic representation of d. The j^{th} feature $V_d^{z_j^r} = P(z_j^r|d)$ (where $1 \leq j \leq n$) is the probability of topic z_j^r to be generated by the unseen document d in the r^{th} topic space of size n (see Figure 1) and $V_{z_j^r}^w = P(w|z_j^r)$ is the vector representation of a word into r.

[1]For comparaison, a set of textual documents from the well-known Reuters-21578 dataset is used during the experiments as well.

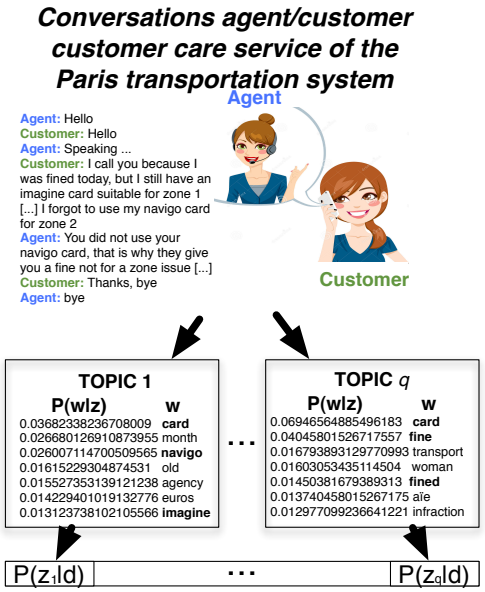

Fig. 1. Example of a document d mapped into a topic space of size q.

A. Variation of LDA model parameters for a document multi-view

Thus, a set of p topic spaces are learned using LDA, presented in its plate notation in Figure 2, by varying the hyper-parameters of the p (here, $p = 500$) topic spaces:

- the number of classes n into the topic space (Section III-A1),
- the α parameter (Section III-A2),
- and the β parameter (Section III-A3).

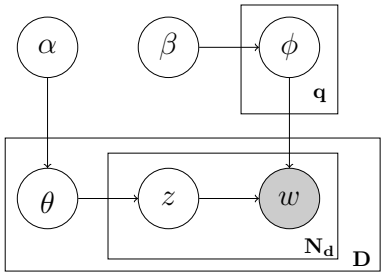

Fig. 2. Generative model for documents in plate notation for Latent Dirichlet Allocation (LDA).

1) Varying n: The number of topics is varied to obtain p topic spaces of size n. The number of topics n varies from 5 to 505. Thus, a set of 500 topic spaces is estimated. This is high enough to generate, for each document, many feature sets for estimating the parameters of a variability model.

2) Varying α: In the LDA technique, the topic z is drawn from a multinomial over θ which is drawn from a Dirichlet distribution over $\overrightarrow{\alpha}$. Thus, a set of p topic spaces of size q is learned using LDA by varying the topic distribution parameter $\overrightarrow{\alpha} = [\alpha_1, \ldots, \alpha_q]^t$.

The standard heuristic is $\alpha_0 = \frac{50}{q}$ [27], which for the setup

of the n^{th} topic space ($1 \leq n \leq p$) would be $\overrightarrow{\alpha_n}[\underbrace{\alpha_n, \ldots, \alpha_n}_{\text{q times}}]^t$ with:

$$\begin{aligned}\alpha_n &= \frac{n}{p} \times \alpha_0 \\ &= \frac{n}{p} \times \frac{50}{q} \ . \end{aligned} \tag{2}$$

The larger α_n ($\alpha_n \geq 1$) is, the more uniform $P(z|d)$ will be (see Figure 3). Nonetheless, this is not what we want: different transcriptions have to be associated with different topic distributions. In the meantime, the higher the α is, the more the draws from the Dirichlet will be concentrated around the mean (see Figure 3 with $\alpha = 20$), which, for a symmetric alpha vector, will be the uniform distribution over q. The number of topics q is fixed to 50, and 500 topic spaces are built ($p = 500$) in our experiments. Thus, α_n varies between a low value (sparse topic distribution $\alpha_1 = 0.002$) to 1 (uniform Dirichlet $\alpha_p = 1$).

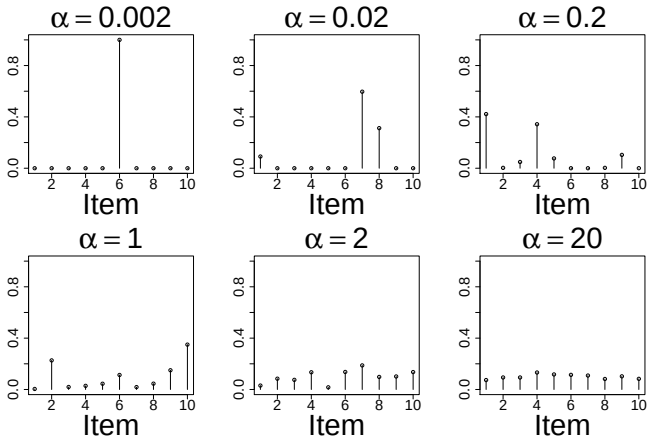

Fig. 3. Dirichlet distribution with a varied α_n.

3) Varying β: In the same way, the hyper-parameter β controls the sparsity of words distribution in each class in the topic space. Thus, the larger β is, the more uniform $P(w|z)$ will be. This means that the probability of each word contained into a class will be roughly the same, and therefore, the classes themselves will be thematically close. During the inference process which allows us to represent the document in the topic space, the distribution of topics for a given document have to be different, mostly if the documents are not labeled with the same theme. The ϕ matrix is drawn from a Dirichlet with β parameter. The results obtained with different values of α shown in Figure 3, could be considered for different values of β.

The standard heuristic is $\beta_0 = 0.1$ [27], which for the setup of the n^{th} topic space ($1 \leq n \leq p$) would be $\overrightarrow{\beta_n}[\underbrace{\beta_n, \ldots, \beta_n}_{|V|\text{times}}]^t$ with:

$$\begin{aligned}\beta_n &= \frac{n}{p} \times \beta_0 \\ &= \frac{n}{p} \times 0.1 \ . \end{aligned} \tag{3}$$

The number of topics q is fixed to 50, and 500 topic spaces are built ($p = 500$) in our experiments. Thus, β_n varies between a low value (sparse topic distribution $\alpha_1 = 0.0002$) to 1 (uniform Dirichlet $\beta_p = 0.1$).

The next process allows us to obtain a homogeneous representation of document d for the r^{th} topic space r. Section III-B presents the mapping of each thematic representation of a document into a homogenous space composed with a set of discriminant words.

B. Multiple representations in a homogenous space of discriminant words

The feature vector $V_d^{z^m}$ of d is mapped to the common vocabulary space V composed with a set of $|V|$ discriminative words [2] of size 166, to obtain a new feature vector $V_{d,r}^w = \{P(w|d)_r\}_{w \in V}$ of size $|V|$ for the r^{th} topic space r of size n where the i^{th} ($0 \leq i \leq |V|$) feature is:

$$\begin{aligned} V_{d,r}^{w_i} &= P(w_i|d) \\ &= \sum_{j=1}^{n} P(w_i|z_j^r) P(z_j^r|d) \\ &= \sum_{j=1}^{n} V_{z_j^r}^{w_i} \times V_d^{z_j^r} \\ &= \left\langle \overrightarrow{V_{z^r}^{w_i}}, \overrightarrow{V_d^{z^r}} \right\rangle \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ is the inner product, δ being the frequency of the term w_i in d, $V_{z_j^r}^{w_i} = P(w_i|z_j)$ and $V_d^{z_j^r} = P(z_j|d)$ evaluated using Gibbs Sampling in the topic space r.

IV. Compact multi-view representation

The multi-view representations of each theme in a large number of hidden spaces may cause large discrepancies in the theme identification accuracy when different hidden space sizes are used. In this section, an i-vector-based method to represent automatic transcriptions is presented. Initially introduced for speaker recognition, i-vectors [3] have become very popular in the field of speech processing and recent publications show that they are also reliable for language recognition [31] and speaker diarization [32]. I-vectors are an elegant way of reducing the input space dimensionality while retaining most of the relevant information. The technique was originally inspired by the Joint Factor Analysis framework [33]. Hence, i-vectors convey the speaker characteristics among other information such as transmission channel, acoustic environment or phonetic content of speech segments. The next sections describe the i-vector extraction process, the application of this compact representation to textual documents (called c-vector), and the vector transformation with the EFR method and the Mahalanobis metric.

A. Total variability space definition

I-vector extraction could be seen as a probabilistic compression process that reduces the dimensionality of speech super-vectors according to a linear-Gaussian model. I-vectors is also an elegant way of mapping a high dimensional representation of an entity into a feature vector of reduced dimensions while retaining while preserving most if not all the relevant information content of the initial representation.

The speech (of a given speech recording) super-vector $\mathbf{m}_s$ of concatenated GMM means is projected in a low dimensionality space, named Total Variability space, with:

$$\mathbf{m}_{(h,s)} = m + \mathbf{T}\mathbf{x}_{(h,s)} \ , \tag{4}$$

where m is the mean super-vector of the UBM[2] and is constructed by concatenation the means of all the Gaussians in the UBM. $\mathbf{T}$ is a low rank matrix $(MD \times R)$, where M is the number of Gaussians in the UBM and D is the cepstral feature size, which represents a basis of the reduced total variability space. $\mathbf{T}$ is named *Total Variability matrix*; the components of $\mathbf{x}_{(h,s)}$ are the total factors which represent the coordinates of the speech recording in the reduced total variability space called i-vector (i for $i$dentification).

B. From i-vector speaker identification to c-vector textual document classification

The proposed approach uses i-vectors to model transcription representation through each topic space in a homogeneous vocabulary space. These short segments are considered as basic semantic-based representation units. Indeed, vector V_d^w represents a segment or a session of a transcription d. In the following, (d, r) will indicate the document representation d in the topic space r. In our model, the segment super-vector $\mathbf{m}_{(d,r)}$ of a transcription d knowing a topic space r is modeled:

$$\mathbf{m}_{(d,r)} = m + \mathbf{T}\mathbf{x}_{(d,r)} \tag{5}$$

where $\mathbf{x}_{(d,r)}$ contains the coordinates of the topic-based representation of the document in the reduced total variability space called c-vector (c for $c$lassification).

where $\mathbf{T}$ is a low rank matrix of dimensions $(J \times C)$, where J ($M \times |V|$, $|V|$ is the number of words contained into the discriminative words list V) is the number of elements of the super-vector and C is the number of elements in a reduced total variability space where the vector representing d is called c-vector (for $c$lassification vector). The c-vector is obtained as described below by adapting to theme identification of textual data an algorithm for computing I-vectors used for speaker verification.

$\mathbf{T}$ is named total variability matrix; the components of are the elements of the C-vector. The $\mathbf{T}$ matrix is estimated, as described below, using training data and an estimation $x_{(d,r)}$of is obtained with maximum posterior probability (MAP) estimation as described in the following. Let $\mathbf{N}_{(d,r)}$ and $\mathbf{X}_{(d,r)}$ be two vectors containing the zero order and first order document statistics respectively. The statistics are estimated against the UBM:

$$\mathbf{N}_r[g] = \sum_{t \in r} \gamma_g(t); \ \{\mathbf{X}_{(d,r)}\}_{[g]} = \sum_{t \in (d,r)} \gamma_g(t) \cdot t \tag{6}$$

where $\gamma_g(t)$ is the *a posteriori* probability of Gaussian g for the observation t. In the equation, $\sum_{t \in (d,r)}$ represents the sum over all the frames belonging to the document d.

Let $\overline{\mathbf{X}}_{(d,r)}$ be the state dependent statistics defined as follows:

$$\{\overline{\mathbf{X}}_{(d,r)}\}_{[g]} = \{\mathbf{X}_{(d,r)}\}_{[g]} - \mathbf{m}_{[g]} \cdot \sum_{(d,r)} \mathbf{N}_{(d,r)}[g] \tag{7}$$

Let $\mathbf{L}_{(d,r)}$ be a $R \times R$ matrix, and $\mathbf{B}_{(d,r)}$ a vector of dimension R, both defined as:

$$\begin{aligned} \mathbf{L}_{(d,r)} &= \mathbf{I} + \sum_{g \in \text{UBM}} \mathbf{N}_{(d,r)}[g] \cdot \{\mathbf{T}\}_{[g]}^t \cdot \mathbf{\Sigma}_{[g]}^{-1} \cdot \{\mathbf{T}\}_{[g]} \\ \mathbf{B}_{(d,r)} &= \sum_{g \in \text{UBM}} \{\mathbf{T}\}_{[g]}^t \cdot \mathbf{\Sigma}_g^{-1} \cdot \{\overline{\mathbf{X}}_{(d,r)}\}_{[g]}, \end{aligned} \tag{8}$$

By using $\mathbf{L}_{(d,r)}$ and $\mathbf{B}_{(d,r)}$, $\mathbf{x}_{(d,r)}$ can be obtained using the following equation:

$$\mathbf{x}_{(d,r)} = \mathbf{L}_{(d,r)}^{-1} \cdot \mathbf{B}_{(d,r)} \tag{9}$$

The matrix $\mathbf{T}$ can be estimated line by line, with $\{\mathbf{T}\}_{[g]}^i$ being the i^{th} line of $\{\mathbf{T}\}_{[g]}$ then:

$$\mathbf{T}_{[g]}^i = \mathbf{L}\mathbf{U}_g^{-1} \cdot \mathbf{R}\mathbf{U}_g^i, \tag{10}$$

where $\mathbf{R}\mathbf{U}_g^i$ and $\mathbf{L}\mathbf{U}_g$ are given by:

$$\begin{aligned} \mathbf{L}\mathbf{U}_g &= \sum_{(d,r)} \mathbf{L}_{(d,r)}^{-1} + \mathbf{x}_{(d,r)}\mathbf{x}_{(d,r)}^t \cdot \mathbf{N}_{(d,r)}[g] \\ \mathbf{R}\mathbf{U}_g^i &= \sum_{(d,r)} \{\overline{\mathbf{X}}_{(d,r)}\}_{[g]}^{[i]} \cdot \mathbf{x}_{(d,r)} \end{aligned} \tag{11}$$

Algorithm 1: Estimation of matrix $\mathbf{T}$ and latent variable $\mathbf{x}_{(d,r)}$.

```
For each document d mapped into the topic space r:
x_(d,r) ← 0, T ← random ;
Estimate statistics: N_(d,r), X_(d,r) (eq.6);
for i = 1 to nb_iterations do
    for all d and r do
        Center statistics: X̄_(d,r) (eq.7);
        Estimate L_(d,r) and B_(d,r) (eq.8);
        Estimate x_(d,r) (eq.9);
    end
    Estimate matrix T (eq. 10 and 11) ;
end
```

Algorithm 1 shows the pseudo-code for the method adopted to estimate the conversation multi-view variability matrix. A standard likelihood function can be used to assess the convergence as shown with more details in [34].

C-vector representation suffers from 3 raised issues. In the following, the application of these important constraints is discussed:

[2]The UBM is a GMM that represents all the possible observations.

- In theory c-vectors (Equation 5) should have normal distribution $\mathcal{N}(0, I)$.
- The so called radial effect should be removed.
- The full rank total factor space should be used to apply discriminant transformations.

The next section presents a solution to these 3 problems.

C. C-vector standardization

A solution to standardize c-vectors has been developed in [8]. The authors proposed to apply transformations for training and test transcription representations. The first step is to evaluate the empirical mean $\overline{\mathbf{x}}$ and covariance matrix $\mathbf{V}$ of the training c-vector. Covariance matrix $\mathbf{V}$ is decomposed by diagonalization into:

$$\mathbf{PDP}^T \tag{12}$$

where $\mathbf{P}$ is the eigenvector matrix of $\mathbf{V}$ and $\mathbf{D}$ is the diagonal version of $\mathbf{V}$. A training i-vector $\mathbf{x}_{(d,r)}$ is transformed in $\mathbf{x}'_{(d,r)}$ as follows:

$$\mathbf{x}'_{(d,r)} = \frac{\mathbf{D}^{-\frac{1}{2}}\mathbf{P}^T(\mathbf{x}_{(d,r)} - \overline{\mathbf{x}})}{\sqrt{(\mathbf{x}_{(d,r)} - \overline{\mathbf{x}})^T\mathbf{V}^{-1}(\mathbf{x}_{(d,r)} - \overline{\mathbf{x}})}} \tag{13}$$

The numerator is equivalent by rotation to $\mathbf{V}^{-\frac{1}{2}}(\mathbf{x}_{(d,r)} - \overline{\mathbf{x}})$ and the Euclidean norm of $\mathbf{x}'_{(d,r)}$ is equal to 1. The same transformation is applied to the test c-vectors, using the training set parameters $\overline{\mathbf{x}}$ and mean covariance $\mathbf{V}$as estimations of the test set of parameters.

Figure 4 shows the transformation steps: Figure 4-(a) is the original training set; Figure 4-(b) shows the rotation applied to the initial training set around the principal axes of the total variability when $\mathbf{P}^T$ is applied; Figure 4-(c) shows the standardization of c-vectors when $\mathbf{D}^{-\frac{1}{2}}$ is applied; and finally, Figure 4-(d) shows the c-vector $\mathbf{x}'_{(d,r)}$ on the surface area of the unit hypersphere after a length normalization by a division of $\sqrt{(\mathbf{x}_{(d,r)} - \overline{\mathbf{x}})^T\mathbf{V}^{-1}(\mathbf{x}_{(d,r)} - \overline{\mathbf{x}})}$.

V. Experimental Protocol

The proposed c-vector representation is evaluated in the context of the theme identification of automatic human-human telephone conversation transcriptions and of the categorization of textual newswire collection. This representation is built from a set of feature vectors. Each one is composed with scores of discriminative words. Then, the metric used to associate a document to a class is the Mahalanobis metric.

A LDA model allowed us to elaborate 500 topic spaces by varying LDA hyper-parameters (see Section III-A). A topic space having less than 5 topics is not suitable for large corpus such as those used during our experiments (see Section V-B). For each theme or category, a set of $|V|$ specific words is identified as explained in Section V-D. All the selected words are then merged without repetition to compose V. The topic spaces are made with the LDA Mallet Java implementation[3].

Next sections describe the datasets used for the experiments, the Mahalanobis distance between two vectors, the metrics to evaluate the system performance, and finally a study is given to find out the best number of discriminative words for each configuration.

[3]http://mallet.cs.umass.edu/

A. Mahalanobis distance

Given a new observation x, the goal of the task is to identify the theme (or category) belonging to x. Probabilistic approaches ignore the process by which c-vectors were extracted and they pretend instead they were generated by a prescribed generative model. Once a c-vector is obtained from a document, its representation mechanism is ignored and it is regarded as an observation from a probabilistic generative model. The Mahalanobis scoring metric assigns a document d with the most likely theme C. Given a training dataset of documents, let $\mathbf{W}$ denote the within document covariance matrix defined by:

$$\begin{aligned}\mathbf{W} &= \sum_{k=1}^{K} \frac{n_t}{n}\mathbf{W_k} \\ &= \frac{1}{n}\sum_{k=1}^{K}\sum_{i=0}^{n_t}\left(x_i^k - \overline{x_k}\right)\left(x_i^k - \overline{x_k}\right)^t\end{aligned} \tag{14}$$

where $\mathbf{W_k}$ is the covariance matrix of the k^{th} theme C_k, n_t is the number of utterances for the theme C_k, n is the total number of documents, and $\overline{x_k}$ is the centroid (mean) of all documents x_i^k of C_k.

Each document does not contribute to the covariance in an equivalent way. For this reason, the term $\frac{n_t}{n}$ is introduced in equation 14. If homoscedasticity (equality of the class covariances) and Gaussian conditional density models are assumed, a new observation x from the test dataset can be assigned to the most likely theme $C_{k_{\text{Bayes}}}$ using the classifier based on the Bayes decision rule:

$$\begin{aligned}C_{k_{\text{Bayes}}} &= \arg\max_k \left\{\mathcal{N}\left(x \mid \overline{x_k}, \mathbf{W}\right)\right\} \\ &= \arg\max_k \left\{-\frac{1}{2}\left(x - \overline{x_k}\right)^t \mathbf{W}^{-1}\left(x - \overline{x_k}\right) + a_k\right\}\end{aligned}$$

where $\mathbf{W}$ is the within theme covariance matrix defined in equation 14; $\mathcal{N}$ denotes the normal distribution and $a_k = \log\left(P(C_k)\right)$. It is noted that, with these assumptions, the Bayesian approach is similar to Fisher's geometric approach: x is assigned to the class of the nearest centroid, according to the Mahalanobis metric [35] of $\mathbf{W}^{-1}$:

$$C_{k_{\text{Bayes}}} = \arg\max_k \left\{-\frac{1}{2}||x - \overline{x_k}||^2_{\mathbf{W}^{-1}} + a_k\right\}$$

B. Datasets

To evaluate the effectiveness of the proposed compact version of a multi-granularity representation of a document, the experiments are conduced by using both documents from automatic transcriptions (DECODA corpus presented in Section V-B1) and classical textual documents (Reuters-21578 corpus presented in Section V-B2).

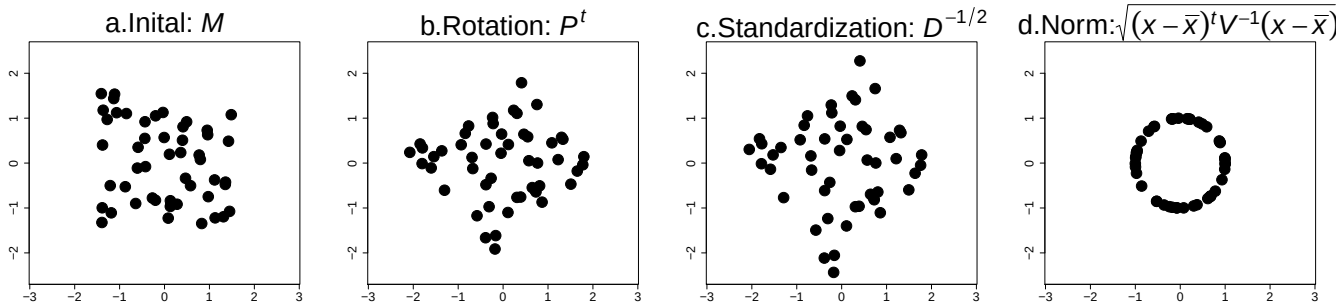

Fig. 4. Effect of the standardization with the EFR algorithm.

1) DECODA corpus: The first corpus is a set of human-human telephone conversations in the customer care service (CCS) of the RATP Paris transportation system. This corpus comes from the DECODA project [6] and is used to perform experiments on the conversation theme identification. It is composed of 1,242 telephone conversations, corresponding to about 74 hours of signal, split as described in Table I.

TABLE I
DECODA DATASET.

Class label	*Number of samples*		
	training	**development**	**testing**
problems of itinerary	145	44	67
lost and found	143	33	63
time schedules	47	7	18
transportation cards	106	24	47
state of the traffic	202	45	90
fares	19	9	11
infractions	47	4	18
special offers	31	9	13
Total	**740**	**175**	**327**

To extract textual content of dialogues from DECODA corpus, an Automatic Speech Recognition (ASR) system is needed. The LIA-Speeral ASR system [36] is used for the experiments. Acoustic model parameters were estimated from 150 hours of speech in telephone conditions. The vocabulary contains 5,782 words. A 3-gram language model (LM) was obtained by adapting a basic LM with the training set transcriptions. A "stop list" of 126 words[4] was used to remove unnecessary words (mainly function words) which results in a Word Error Rate (WER) of 33.8% on the training, 45.2% on the development, and 49.5% on the test. These high WER are mainly due to speech disfluencies and to adverse acoustic environments (for example, calls from noisy streets with mobile phones)

2) Reuters-21578 dataset: To evaluate the relevance of the proposed compact representation of a document, the categorization task of top-10 classes of the Reuters-21578 ModApte splitcorpora [7] is used. Table II presents the number of documents of both training, testing and development sets for each of the 10 classes of Reuters corpus [37], [38].

[4]http://code.google.com/p/stop-words/

TABLE II
TOP-10 CLASSES OF REUTERS-21578 DATASET.

Class label	*Number of samples*		
	training	**development**	**testing**
earn	2,590	287	1,087
acq	1,485	165	719
money-fx	484	54	179
grain	390	43	149
crude	350	39	189
trade	332	37	117
interest	312	35	131
ship	177	20	89
wheat	191	21	71
corn	163	18	56
Total	**6,474**	**719**	**2,787**

C. Metrics

The Mahalanobis distance allows us to evaluate the similarity between two vectors (here, the document representation and the centroid of each class) and to label a document (or a dialogue in the DECODA corpus) with a certain class. At the end of this process, an efficient metric have to be chosen to evaluate the performance of the categorization system proposed in this paper. This section presents two metrics: the accuracy, for DECODA theme identification task, and the Macro-F1, for automatic labeling process of Reuters-21578 documents. The accuracy is the metric chosen during the previous studies concerning the DECODA theme identification task. In the same way, to compare the results obtained with automatic transcription (DECODA) and textual documents (Reuters), the accuracy is also used as an evaluation metric for the Reuters corpus.

This last one is usually evaluated with the Macro-F1 in previous studies. For this reason, this metric is employed to evaluate the proposed compact representation of textual documents (Reuters) in comparison to previous studies. Next sections describe these two metrics.

1) Macro-F1 metric: This well-known dataset categorization task is usually evaluated using the macro-F1 metric. F1-measure is computed for each class within the dataset and then, the average over all of the classes is obtained. Hence, equal weight is assigned to each class regardless of the class frequency [39]. Computation of Macro-F1 can be formulated

as:

$$\text{Macro-F1} = \sum_{k=1}^{K} \frac{F_k}{K} \ , \ F_k = \frac{2 \times p_k \times r_k}{p_k + r_k} \ , \qquad (15)$$

where p_k and r_k are respectively the precision and the recall of the class k among the K classes, determined as follow :

$$p_k = \frac{TP_k}{TP_k + FP_k} \quad \text{and} \quad r_k = \frac{TP_k}{TP_k + FN_k} \ . \qquad (16)$$

FP_k represents the number of documents that do not belong to the class k but are classified to this class incorrectly (*i.e.* false positives); TP_k is the number of documents correctly classified as class k (*i.e.* true positives); FN_k represents the number of documents that belong to class k but which are not classified to this class (*i.e.* false negatives).

2) Accuracy metric: Theme identification task in the DECODA project consists in associating the most likely theme to a dialogue between an agent and a customer. To evaluate the effectiveness of the proposed method, the authors in [40], [41], [42], [43], [44] used only the accuracy defined as:

$$\begin{aligned} accuracy(d) &= \frac{1}{K} \sum_{k=1}^{K} r_k \\ &= \frac{1}{K} \sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k} \end{aligned} \qquad (17)$$

One can find more about evaluation metrics including macro- and micro-metrics in [45].

D. Size of discriminative words set

The method proposes to build a compact representation of a given document from a set of feature vectors $V_d^{w_i^r}$. This vector is composed with the score of each word contained into a discriminative set of words. The mapping step of the thematic representation $V_d^{z_j^r}$ is needed to obtain a homogenous and with equal size representation of the document.

Figure 5 shows theme identification accuracies obtained with different configurations (Train./Dev./Test) and different discriminative words set size $|V|$ for DECODA corpus and macro-F1 for Reuters dataset. Note that, for the DECODA corpus, ASR corresponds to automatic transcriptions of dialogues. The number of classes contained into the topic space is fixed to 50. The main remark is that the best accuracy or macro-F1 is roughly achieve with a set of 20 discriminant words for each configuration of training and development sets. We can also point out that the larger the size of discriminative words set is, the lower the accuracy is. Thus, a set size of 20 discriminative words seems to be the most effective for theme identification task. Other experiments (not presented here due to space considerations) with topic spaces having different numbers of classes (not only 50), show that the best number of discriminative words is around 20 words. Thus, the following experiments are performed with a vocabulary V of size 20.

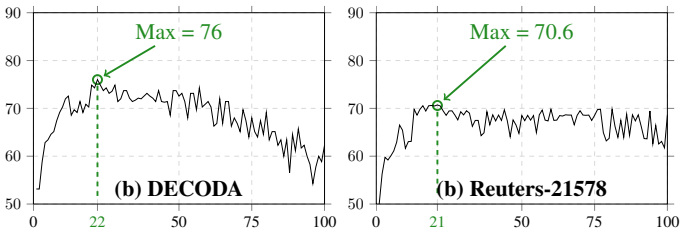

Fig. 5. Theme classification accuracies (%) of development set from DECODA data set and macro-F1 for Reuters-21578 (b) development set using various discriminative words sets size $|V|$ ($1 \leq |V| \leq 100$) (x-axis).

TABLE III
EXPERIMENTAL CONFIGURATIONS FOR MULTI-GRANULARITY REPRESENTATIONS.

	Parameters configuration		
	n	α	β
n	$\mathbf{5 \leq n \leq 505}$	$\alpha = \frac{50}{50} = 1$	$\beta = 0.1$
α	$n = 50$	$\mathbf{0.002 \leq \alpha \leq 1}$	$\beta = 0.1$
β	$n = 50$	$\alpha = \frac{50}{50} = 1$	$\mathbf{0.0002 \leq \beta \leq 0.1}$

VI. EXPERIMENTS AND RESULTS

The proposed c-vector approach is applied to the same classification task and corpus proposed in [2] (state-of-the-art in text classification). Experiments are conducted using the multiple topic spaces estimated with a LDA approach. From these multiple topic spaces, a classical way is to find the one that reaches the best performance [2]. The first experiments presented in Section VI-A are conduced with the DECODA dataset composed with dialogues between an agent and a customer presented in Section V-B1. The compact vector of a textual document is then used to represent documents from Reuters-21578 corpus [7] (see Section V-B2) in a categorization task.

For both tasks, the multiple topic spaces are built by varying one parameter among α, β, and the number of classes into the topic space itself as shown in Table III.

A. Compact representation of highly imperfect automatic transcriptions

Figures 6 present the theme classification performance obtained on the development and test sets using various topic-based representation configurations with the EFR normalization algorithm (*baseline*) for ASR datasets from DECODA project.

First of all, we can see that this baseline approach reached best classification accuracies with multi-view representation of highly imperfect transcription when varying the parameter α, that controls sparsity of topics distribution in documents of training set, for ASR DECODA datasets. These results achieve 86.9% and 80.1% respectively on the development and the test sets.

Then, the variation of the hyper-parameter β obtains the second best accuracy in terms of theme identification with 82.9% and 74.0% on the development and test sets.

We can point out that these results are close to those obtained with a multiple representation by varying the number

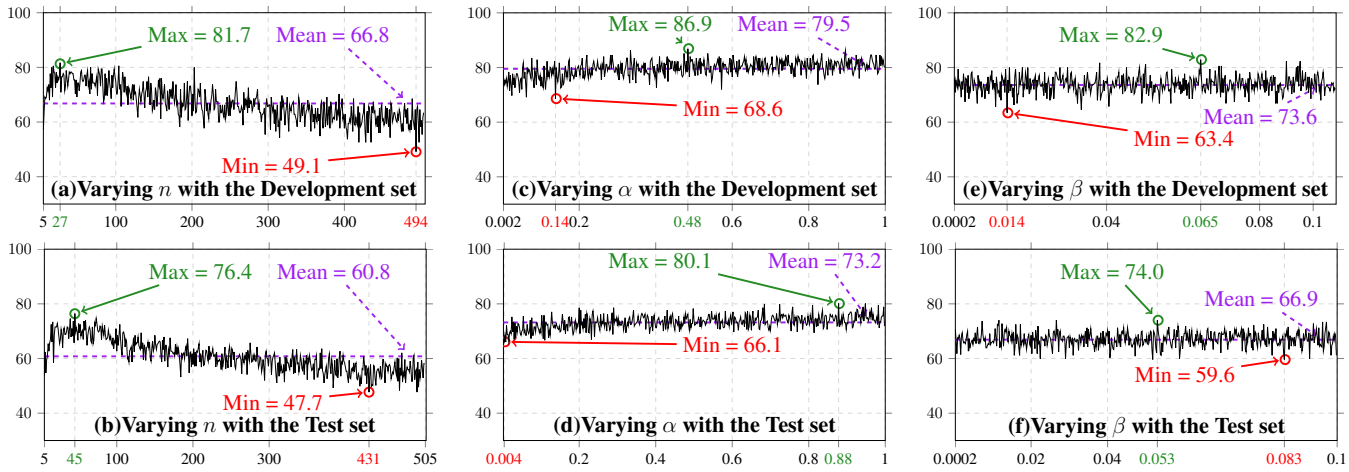

Fig. 6. Theme classification accuracies (%) using various topic-based representations with EFR normalization (baseline) on the development and test sets from Decoda corpus with different experimental configurations with both Training and Validation datasets from ASR. X-axis represents represents the varying parameter: the number n of classes contained into the topic space (a) and (b); α (c) and (d) and β (e) and (f).

of classes n into the topic space (81.7% and 76.4% for development and test sets). Since difference on the development set is not negligible ($82.9 - 81.7 = 1.2$ points), the result achieved with the test set is quit similar for both β and n multiple representations (74.0% and 76.4%).

Nonetheless, we note that the classification performance is rather unstable, and may completely change from a topic space configuration to another. The gap between the lower and the higher classification results is also important, with a difference of 32.5 points on the development set (same tendency is observed on the test set) when the parameter n is varying. As a result, finding the best topic space size seems crucial for this classification task, particularly in the context of highly imperfect automatic dialogue transcriptions containing more than one theme.

Then, the main remark regarding the accuracies obtained by varying the LDA hyper-parameters, is that the best results are clearly achieved by varying the parameter α, while the two others parameters (β and n) obtain roughly the same accuracies. This is non-intuitive: we expected that the number of topics n would have a higher impact on the topic space statistical structure than the hyper-parameter α. Nonetheless, this remark is effective and relevant when the goal is to map a document in a single topic space. Thus, the inference process is sensitive to the topic distribution of an unseen document into the topic space which is controlled by the parameter α. The purpose here is to build different views of the same document to avoid the complex choice of LDA model hyper-parameters, and to consider different "views" of the same document. Thus, the number of topics, which controls the granularity of topic spaces, allows us to better represent the multiple views of a document and compact this multiple representation by compensating the noise variability.

Tables **??** and IV present accuracies obtained with the proposed c-vector approach coupled with the EFR algorithm with different c-vector sizes and a different number of Gaussians into the GMM-UBM for DECODA ASR datasets. The results presented in Table V show the theme identification accuracies

TABLE IV
THEME CLASSIFICATION ACCURACIES (%) WITH DIFFERENT c-VECTORS AND GMM-UBM SIZES FOR TRAINING SET FROM ASR → TEST FROM ASR BY VARYING THE NUMBER OF TOPICS n.

(c) Variation of the number of classes n

size of c-vector	*DEV* Number of Gaussians in GMM-UBM					*TEST* Number of Gaussians in GMM-UBM				
	32	*64*	*128*	*256*	*512*	*32*	*64*	*128*	*256*	*512*
60	82.9	82.9	88.0	84.0	86.3	85.0	83.2	83.2	83.8	84.7
80	85.1	81.7	84.6	82.9	86.3	82.9	79.5	83.2	84.7	82.0
100	82.4	82.9	85.1	**89.7**	88.0	84.1	81.0	84.7	**86.5**	83.5
120	83.4	84.0	81.7	87.4	85.1	82.9	83.2	85.0	84.4	81.6
140	81.1	84.6	87.4	84.6	83.4	85.3	82.9	84.1	82.9	82.9
160	78.9	82.3	82.9	83.4	82.9	86.5	84.7	80.7	82.9	81.7
180	81.1	81.7	80.6	83.4	82.3	85.0	80.7	79.8	78.6	79.2
200	82.3	82.9	84.6	81.1	81.7	83.5	81.3	79.8	79.8	76.1

(a) Variation of parameter α

size of c-vector	*DEV* Number of Gaussians in GMM-UBM					*TEST* Number of Gaussians in GMM-UBM				
	32	*64*	*128*	*256*	*512*	*32*	*64*	*128*	*256*	*512*
60	81.1	82.9	78.9	78.9	77.1	82.9	78.6	72.8	80.1	71.0
80	82.9	84.6	75.4	80.6	76.0	81.3	79.2	73.4	78.3	69.1
100	85.7	**86.9**	79.4	80.0	72.6	78.9	**80.4**	80.1	75.5	63.9
120	80.0	82.9	75.5	74.3	73.7	83.2	80.4	71.0	70.6	60.2
140	79.4	80.0	79.4	76.0	68.6	74.3	73.1	75.0	69.4	59.6
160	74.3	80.6	73.1	78.3	68.0	77.7	75.2	70.3	68.2	61.2
180	76.6	83.4	72.0	73.7	66.3	72.2	71.9	61.2	66.1	57.5
200	72.0	74.3	68.0	66.9	67.4	70.3	70.3	68.8	66.4	55.4

(b) Variation of parameter β

size of c-vector	*DEV* Number of Gaussians in GMM-UBM					*TEST* Number of Gaussians in GMM-UBM				
	32	*64*	*128*	*256*	*512*	*32*	*64*	*128*	*256*	*512*
60	84.0	84.0	74.9	76.0	73.7	73.0	77.3	70.8	75.1	67.5
80	76.6	83.4	79.4	74.3	76.0	73.6	75.1	72.4	72.7	69.9
100	71.4	**85.7**	74.3	76.0	76.6	77.8	**75.4**	68.7	68.5	69.9
120	74.3	77.7	71.4	76.6	65.7	80.4	70.5	68.7	75.7	66.6
140	66.6	78.3	63.4	64.6	65.1	73.7	69.0	67.8	67.2	62.6
160	69.4	76.0	68.6	61.1	65.1	70.3	74.5	70.2	63.2	69.3
180	70.9	77.1	68.6	65.7	62.9	70.9	72.4	68.3	61.4	66.0
200	68.6	73.7	60.6	66.3	56.0	69.7	63.9	65.0	63.8	68.0

obtained with ASR DECODA dataset and LDA parameter variations (α, β and n). The three last columns present best accuracies obtained with both the development and test sets, and the last column presents results obtained when the best configuration found with the development set is applied to the test set (real configuration since, in a real case of dialogue categorization, the test label is unknown).

The configuration is whether TRS or ASR, for both training and test sets. We can firstly note that this compact representation allows it to outperform results obtained with the best topic space configuration (*baseline*), with a gain of 7.9 points[5] on the development data and of 7.7 points[6] on the test data.

TABLE V
BEST AND REAL THEME CLASSIFICATION ACCURACY (%) OF DECODA DATASET WITH DIFFERENT METHODS AND DIFFERENT CONFIGURATIONS.

Method employed	Variation parameter	DataSet		Best	Best	Real
		TRAIN	TEST	DEV	TEST	TEST
TbT	n	ASR	ASR	81.7	76.4	70.6
c-vector	n	ASR	ASR	89.7	**86.5**	**86.5**
TbT	α	ASR	ASR	86.9	80.1	76.1
c-vector	α	ASR	ASR	86.9	**83.2**	**80.4**
TbT	β	ASR	ASR	82.9	74.0	68.5
c-vector	β	ASR	ASR	85.7	**80.4**	**75.4**

Some results obtained with different document representations as well as different classification methods, are presented in Table VI. Indeed, throughout these previous studies, several approaches were proposed to puzzle out this categorization task of such noisy documents. The most classical one is a term frequency representation coupled with a support vector machines [2] (SVM) classification approach (TF-IDF-Gini+SVM). This method is applied for ASR dataset configurations and obtain good results (73.5% for ASR).

This basic representation of a document with the term-frequency reveals little in way of intra- or inter-documents statistical structure. For this reason, a set of more abstract features from a topic space (LDA) are used to represent the document. This representation is coupled with a SVM classification approach and a Gaussian classifier (Mahalanobis) in [2] which allows us to improve the results obtained (LDA+Maha.). This classifier, based on a decision rule, was also used with different representations of the same document, depending on the speaker [44] with an accuracy of 87.2% and 84.1% for development and test sets respectively.

Another issue related to document structure (words distribution into the document), is the position of each occurrence of a word in a document. Indeed, in the conversation context, the agent have to decide the theme of the dialogue during the first sentences and have to follow a strict protocol. Thus, the position of words impact the document labeling task. Moreover, a same dialogue may contain more than one theme. For these reasons, a theme in a dialogue may change from a segment to another. Thus, a document representation which considers the word occurrence positions in the document is proposed as well. This representation takes the form of a hyper-complexe named *quaternion* [40], and contains 4 elements. In each element, authors in [40] insert the term frequency of the word in a particular segment of the dialogue. This representation is extremely dependent to the document transcription quality.

[5]Lower difference when the baseline system is applied with the development set (83.8%) and the c-vector representation with the development set by varying the parameter n (91.7%).

[6]Lower difference when the baseline system is applied with the test set (83.2%) and the c-vector representation with the test set by varying respectively the parameter α and n (89.3%).

The proposed c-vector compact representation obtains the best results. This method is from the Joint Factor Analysis framework (JFA). Table VI presents some interesting results obtained with other methods from JFA such as the semantic variability compensation (filter) [42] or the Subspace Gaussian Mixture Models (SGMM) [43]. These methods are applied with the ASR configuration to compensate noise variability (automatic transcription process + semantic variability due to multiple document mapping in several topic spaces). The SGMM method is evaluated with different adaptation algorithms (EM, MAP and JFA). One can notice that the method, which achieves the best results in terms of theme identification and noise variability compensation, is the compact representation c-vector.

TABLE VI
COMPARISON BETWEEN THE PROPOSED COMPACT c-VECTOR REPRESENTATION AND DIFFERENT DOCUMENT REPRESENTATIONS, CATEGORIZATION ALGORITHMS IN TERMS OF CLASSIFICATION ACCURACY (%) OF DECODA DATASET.

DECODA dataset of automatic transcriptions				
[43] GMM-EM	Maha.	ASR	ASR	63.5
TbT (β)	Maha.	ASR	ASR	68.5
TbT (n)	Maha.	ASR	ASR	70.6
[2] TF-IDF-Gini	SVM	ASR	ASR	73.5
[40] Quaternion	kNN	ASR	ASR	73.9
TbT (α)	Maha.	ASR	ASR	76.1
[43] GMM-MAP	Maha.	ASR	ASR	77.9
[43] SGMM (JFA)	Maha.	ASR	ASR	78.8
[42] Filtrage (JFA)	Maha.	ASR	ASR	80.0
[2] LDA	SVM	ASR	ASR	81.4
[2] LDA	Maha.	ASR	ASR	83.3
[44] LDA Speaker	Maha.	ASR	ASR	84.1
c-vector (n) (JFA)	Maha.	ASR	ASR	**86.5**

B. Compact representation of textual documents

Figure 7 shows macro-F1 obtained on the classification task of Reuters-21578 dataset with different compact representations obtained by varying LDA parameters (n, α and β). The first remark, is that the best results are obtained when the parameter is α varies (91.1% for the development set and 76.8% for the test set), then, when the parameter β varies (87.6% for development set and 74.1% for test set), and finally when the number of classes into a topic space varies (82.9% for development set and 72.6% for test set). Even if the difference Δ of macro-F1 is significative for the development set from a parameter to another ($3.5 \leq \Delta \leq 8.2$), this difference is reduced when the test set is considered ($1.5 \leq \Delta \leq 4.2$).

Table VII presents accuracies obtained with the proposed c-vector approach coupled with the EFR algorithm with different c-vector sizes and a different number of Gaussians into the GMM-UBM for the Reuters-21578 dataset. The results presented in Table VIII show the categorization macro-F1 scores obtained with different LDA parameter variations (α, β and n). The three last columns present best accuracies obtained with both the development and test sets, and the last column presents results obtained when the best configuration found with the development set is applied to the test set (real configuration since, in a real case of dialogue categorization, the test label is unknown).

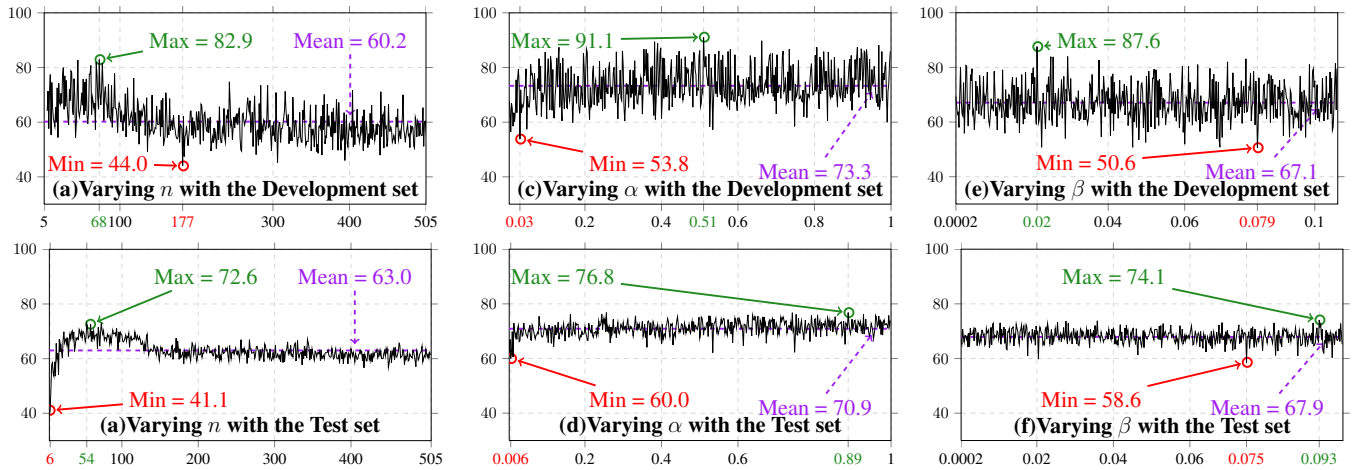

Fig. 7. Theme classification Macro-F1 (%) using various topic-based representations with EFR normalization (baseline) on the development and test sets from the Reuters corpus. X-axis represents represents the varying parameter: the number n of classes contained into the topic space (a) and (b); α (c) and (d) and β (e) and (f).

TABLE VII
THEME CLASSIFICATION MACRO-F1 (%) WITH DIFFERENT c-VECTORS AND GMM-UBM SIZES FOR THE REUTERS DATASET BY VARYING THE PARAMETER α, β AND THE NUMBER OF CLASSES n CONTAINED IN THE TOPIC SPACE.

(a) Variation of the number of classes n

size of	*DEV*					*TEST*				
	Number of Gaussians in GMM-UBM									
c-vector	*32*	*64*	*128*	*256*	*512*	*32*	*64*	*128*	*256*	*512*
60	81.6	84.7	84.0	79.4	80.6	80.6	79.8	80.7	79.3	80.2
80	82.6	77.5	83.4	87.4	75.5	81.6	76.7	80.3	80.3	80.4
100	80.0	81.1	**86.7**	78.8	75.8	81.0	78.3	**80.0**	79.0	79.6
120	72.9	82.7	86.1	76.6	74.4	81.1	79.1	78.9	78.2	79.0
140	74.5	80.1	82.3	82.0	85.1	80.7	79.5	76.5	77.0	79.8
160	81.7	77.4	74.4	78.6	77.7	79.9	79.9	78.6	75.2	78.7
180	72.1	80.3	72.3	71.4	70.8	78.9	80.3	74.3	73.7	77.2
200	73.3	79.8	74.7	79.8	77.6	79.7	81.5	75.8	73.5	76.7

(b) Variation of parameter α

size of	*DEV*					*TEST*				
	Number of Gaussians in GMM-UBM									
c-vector	*32*	*64*	*128*	*256*	*512*	*32*	*64*	*128*	*256*	*512*
60	83.3	75.2	81.3	80.1	74.2	75.2	78.4	77.9	78.2	79.6
80	79.8	64.8	75.0	70.3	80.8	77.1	76.8	78.0	77.2	78.9
100	76.4	75.8	63.2	69.1	70.9	77.7	77.1	76.9	78.0	78.2
120	75.0	81.3	67.9	77.1	71.9	75.1	77.3	77.7	73.1	77.3
140	79.5	**86.2**	65.1	72.0	74.7	74.1	**74.5**	76.7	74.4	75.7
160	71.5	76.7	72.0	78.1	63.4	78.6	79.5	73.3	73.3	71.4
180	70.0	78.0	78.6	71.1	58.0	74.3	76.2	73.7	73.3	69.5
200	62.6	84.0	73.4	64.3	58.2	69.1	79.6	74.4	72.0	69.1

(c) Variation of parameter β

size of	*DEV*					*TEST*				
	Number of Gaussians in GMM-UBM									
c-vector	*32*	*64*	*128*	*256*	*512*	*32*	*64*	*128*	*256*	*512*
60	74.3	65.2	71.7	77.7	79.4	79.3	77.1	79.1	79.7	78.6
80	66.3	**80.8**	67.1	68.2	75.7	73.4	**78.6**	75.6	75.7	77.4
100	62.3	69.0	59.0	67.6	67.1	75.9	75.4	73.3	70.8	77.2
120	64.3	67.8	62.6	66.1	70.8	74.3	77.8	68.1	70.0	75.0
140	65.8	63.0	68.7	68.3	69.8	77.3	76.4	67.1	66.1	73.6
160	61.7	72.0	66.5	73.3	61.9	79.5	79.4	63.5	67.7	72.1
180	61.8	71.1	62.7	68.8	66.2	79.5	73.4	65.0	69.2	64.5
200	61.5	68.3	66.5	75.5	65.4	81.5	74.4	63.1	67.2	61.9

The first remark is that results obtained with the proposed compact version of a multi-view representation of a document (81.6% and 80.0% for test in the best and real conditions respectively by varying parameter n), outperform those obtained with the best configuration of LDA parameters (76.8% and 70.9% for test in the best and real conditions respectively by varying parameter α) with a gain of 4.8 and 9.1 points on the test set in the best and real conditions respectively.

One can point out that the best results are obtained by varying α parameter for the baseline, while the best macro-F1 score is obtained by varying the granularity of the topic space (n) with the compact representation c-vector proposed in this paper. These results confirm the initial intuition that the α parameter control the inference process of an unseen document and the multi-view representation requires a set of topic spaces with different granularities (n).

TABLE VIII
BEST AND REAL MACRO-F1 SCORES (%) FOR THEME CLASSIFICATION OF REUTERS CORPORA WITH DIFFERENT METHODS AND DIFFERENT CONFIGURATIONS.

Method employed	Variation parameter	Best DEV	Best TEST	Real TEST
TbT	n	82.9	72.6	63.6
c-vector	n	86.7	**81.6**	**80.0**
TbT	α	91.1	76.8	70.9
c-vector	α	86.2	**79.6**	**74.5**
TbT	β	87.6	74.1	69.1
c-vector	β	80.8	**79.5**	**78.6**

Nonetheless, Table VIII shows that the macro-F1 obtained with the best test (penultimate column), are not as distant for both TbT or c-vector representation as in the "real" conditions (best configuration of the development set applied for the test set). Moreover, the Δ difference of the macro-F1 of the TbT representation when parameters α and β vary for the test set is small ($\Delta = 2.5$ and $\Delta = 1.8$ points for best and real conditions respectively). This difference is more visible with the compact representation (c-vector) with $\Delta = 4.1$ points for test set in "real"condition between representations when α and β vary.

Table IX presents macro-F1 obtained with different document representations and different classification methods. Outcomes obtained with the TbT baseline approach when the number of classes n varies are the less performant (63.62%).

The baseline TbT, with varying parameter β or α, obtains

good results (69.13% and 70.87% respectively) compared to those obtained with an Hybrid Feature Selection (HFS) coupled with either a SVM (63.66%) or a decision tree classification (66.19%) method [37]. HFS consists on two successive selections of relevant features: filter and wrapper selection stages. The filter select features (words) by using the document frequency (DF number of document containing this term [46], [47], [48]), the mutual information [47], the Chi-square which examine the independence of 2 terms [46] and the information gain [47], [48] which performs the importance of a term for a given class. The wrapper method is a genetic algorithm-based selection [49].

A more classical approach based on term frequency (TF-IDF) with or without the Gini criteria (TF-Gini) coupled with a kNN [50] classification algorithm obtains better results (67.15% and 67.93 for TF-IDF and TF-Gini respectively) compared to HFS methods, but less good than those obtained with TbT β or α.

All these approaches are based on a single representation of the document (TF-IDF, TF-Gini, TbT, HFS) and do not take into account different views of this document. The best outcomes in terms of macro-F1 is obtained with the compact representation of the document ($c$vector) with a gain of 9.5 points.

TABLE IX
COMPARISON BETWEEN THE PROPOSED COMPACT c-VECTOR REPRESENTATION AND DIFFERENT DOCUMENT REPRESENTATIONS AND CATEGORIZATION ALGORITHMS IN TERMS OF CLASSIFICATION MACRO-F1 (%) OF REUTERS DATASET.

representation Document	**algorithm categorization**	**TEST Acc. (%)**
TbT (n)	Maha.	63.62
[37] HFS	SVM	63.66
[37] HFS	Decision Tree	66.19
[38] TF-IDF	kNN	67.15
[38] TF-Gini	kNN	67.93
TbT (β)	Maha.	69.13
TbT (α)	Maha.	70.87
c-vector (n)	Maha.	**80.04**

We can conclude that this original c-vector approach allows to better handle variabilities contained in document: in a classification context, a better accuracy can be obtained and the results can be more consistent when varying the c-vector size and the number of Gaussians.

VII. CONCLUSIONS

This paper presents an original multi-view representation of textual documents or automatic speech dialogue transcriptions, and a fusion process with the use of a factor analysis method called i-vector. The first step of the proposed method is to represent a document in multiple topic spaces of different sizes (*i.e.* number of topics), α or β. Then, a compact representation of the document from the multiple views is processed to compensate the vocabulary and the variability of the topic-based representations. The effectiveness of the proposed approach is evaluated in a classification task of theme dialogue identification and document clustering from Reuters-21578.

Thus, the architecture of the system identifies document class using the i-vector approach. This compact representation was initially developed for speaker recognition and we showed that it can be successfully applied to a text classification task. Indeed, this solution allowed the system to obtain better classification accuracy than with the use of the classical best topic space configuration or others well-known methods. In fact, we highlighted that this original compact version of all topic-based representations of documents, called c-vector in this work, coupled with the EFR normalization algorithm, is a better solution to deal with document variabilities (high word error rates or bad acoustic conditions for transcriptions, unusual word vocabulary, etc). This promising compact representation allows us to effectively solve both the difficult choice of the right number of topics and the multi-theme representation issue of particular textual documents. Finally, the classification accuracy reached 86.5% with a gain of 2.4 points compared to the best previous accuracies observed (LDA Speaker). In the case of textual document from Reuters, the gain of 9.5 points is more promising in terms of macro-F1.

In a future work, we plan to evaluate this new representation of textual documents in other information retrieval tasks, such as keyword extraction or automatic summarization systems.

ACKNOWLEDGMENT

This work was funded by the SUMACC and DECODA projects supported by the French National Research Agency (ANR) under contracts ANR-10-CORD-007 and ANR-09-CORD-005.

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
[2] M. Morchid, R. Dufour, P.-M. Bousquet, M. Bouallegue, G. Linarès, and R. De Mori, "Improving dialogue classification using a topic space representation and a gaussian classifier based on the decision rule," in *International Conference on Acoustic, Speech and Signal Processing (ICASSP) 2014*. IEEE, 2014.
[3] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.
[4] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.
[5] 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, 2011.
[6] F. Bechet, B. Maza, N. Bigouroux, T. Bazillon, M. El-Beze, R. De Mori, and E. Arbillot, "Decoda: a call-centre human-human spoken conversation corpus." LREC'12, 2012.
[7] A. Asuncion and D. Newman, "Uci machine learning repository," 2007.
[8] P.-M. Bousquet, D. Matrouf, and J.-F. Bonastre, "Intersession compensation and scoring methods in the i-vectors space for speaker recognition." in *INTERSPEECH*, 2011, pp. 485–488.
[9] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems." in *INTERSPEECH*, 2011, pp. 249–252.
[10] J. R. Bellegarda, "Exploiting both local and global constraints for multi-span statistical language modeling," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2. IEEE, 1998, pp. 677–680.
[11] P. Clarkson and R. Rosenfeld, "Statistical language modeling using the cmu-cambridge toolkit." in *Eurospeech*, vol. 97, 1997, pp. 2707–2710.

[12] R. Iyer and M. Ostendorf, "Relevance weighting for combining multi-domain data for¡ i¿ n¡/i¿-gram language modeling," *Computer Speech & Language*, vol. 13, no. 3, pp. 267–282, 1999.
[13] R. Kneser, J. Peters, and D. Klakow, "Language model adaptation using dynamic marginals." in *Eurospeech*, 1997.
[14] J. R. Bellegarda, "Exploiting latent semantic information in statistical language modeling," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1279–1296, 2000.
[15] Y. Suzuki, F. Fukumoto, and Y. Sekiguchi, "Keyword extraction using term-domain interdependence for dictation of radio news," in *17th international conference on Computational linguistics*, vol. 2. ACL, 1998, pp. 1272–1276.
[16] G. Salton, "Automatic text processing: the transformation," *Analysis and Retrieval of Information by Computer*, 1989.
[17] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 42, no. 1, pp. 177–196, 2001.
[18] T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers, "Statistical topic models for multi-label document classification," *Machine Learning*, vol. 88, no. 1-2, pp. 157–208, 2012.
[19] R. Arun, V. Suresh, C. Veni Madhavan, and M. Narasimha Murthy, "On finding the natural number of topics with latent dirichlet allocation: Some observations," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2010, pp. 391–402.
[20] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Sharing clusters among related groups: Hierarchical dirichlet processes." in *NIPS*, 2004.
[21] E. Zavitsanos, S. Petridis, G. Paliouras, and G. A. Vouros, "Determining automatically the size of learned ontologies," in *ECAI*, vol. 178, 2008, pp. 775–776.
[22] J. Cao, T. Xia, J. Li, Y. Zhang, and S. Tang, "A density-based method for adaptive lda model selection," *Neurocomputing*, vol. 72, no. 7, pp. 1775–1781, 2009.
[23] D. M. Blei and J. Lafferty, "Correlated topic models," *Advances in neural information processing systems*, vol. 18, p. 147, 2006.
[24] W. Li and A. McCallum, "Pachinko allocation: Dag-structured mixture models of topic correlations," 2006.
[25] M. Morchid, M. Bouallegue, R. Dufour, G. Linarès, D. Matrouf, and R. De Mori, "I-vector based approach to compact multi-granularity topic spaces representation of textual documents," in *the Conference of Empirical Methods on Natural Lnguage Processing (EMNLP) 2014*. SIGDAT, 2014.
[26] ——, "I-vector based representation of highly imperfect automatic transcriptions," in *Conference of the International Speech Communication Association (INTERSPEECH) 2014*. ISCA, 2014.
[27] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5228–5235, 2004.
[28] T. Minka and J. Lafferty, "Expectation-propagation for the generative aspect model," in *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2002, pp. 352–359.
[29] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, pp. 721–741, 1984.
[30] G. Heinrich, "Parameter estimation for text analysis," *Web: http://www. arbylon. net/publications/text-est. pdf*, 2005.
[31] D. Martınez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in ivectors space," *INTERSPEECH*, pp. 861–864, 2011.
[32] J. Franco-Pedroso, I. Lopez-Moreno, D. T. Toledano, and J. Gonzalez-Rodriguez, "Atvs-uam system description for the audio segmentation and speaker diarization albayzin 2010 evaluation," in *FALA VI Jornadas en Tecnologa del Habla and II Iberian SLTech Workshop*, 2010, pp. 415–418.
[33] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.
[34] D. Matrouf, N. Scheffer, B. G. Fauve, and J.-F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification." in *INTERSPEECH*, 2007, pp. 1242–1245.
[35] E. P. Xing, M. I. Jordan, S. Russell, and A. Ng, "Distance metric learning with application to clustering with side-information," in *Advances in neural information processing systems*, 2002, pp. 505–512.
[36] G. Linarès, P. Nocéra, D. Massonie, and D. Matrouf, "The lia speech recognition system: from 10xrt to 1xrt," in *Text, Speech and Dialogue*. Springer, 2007, pp. 302–308.
[37] S. Gunal, "Hybrid feature selection for text classification," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 20, no. 2, pp. 1296–1311, 2012.
[38] W. Zhu and Y. Lin, "Using gini-index for feature weighting in text categorization," *Journal of Computational Information Systems*, vol. 9, no. 14, pp. 5819–5826, 2013.
[39] R. R. Larson, "Introduction to information retrieval," 2010.
[40] M. Morchid, G. Linarès, M. El-Beze, and R. De Mori, "Theme identification in telephone service conversations using quaternions of speech features," in *Conference of the International Speech Communication Association (INTERSPEECH) 2013*. ISCA, 2013.
[41] M. Morchid, R. Dufour, and G. Linarès, "A lda-based topic classification approach from highly imperfect automatic transcriptions," in *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC) 2014*, 2014.
[42] M. Bouallegue, M. Morchid, R. Dufour, M. Driss, G. Linarès, and R. De Mori, "Factor analysis based semantic variability compensation for automatic conversation representation," in *Conference of the International Speech Communication Association (INTERSPEECH) 2014*. ISCA, 2014.
[43] ——, "Subspace gaussian mixture models for dialogues classification," in *Conference of the International Speech Communication Association (INTERSPEECH) 2014*. ISCA, 2014.
[44] M. Morchid, R. Dufour, M. Bouallegue, G. Linarès, and R. De Mori, "Theme identification in human-human conversations with features from specific speaker type hidden spaces," in *Conference of the International Speech Communication Association (INTERSPEECH) 2014*. ISCA, 2014.
[45] V. Van Asch, "Macro-and micro-averaged evaluation measures [[basic draft]]," 2013.
[46] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge university press Cambridge, 2008, vol. 1.
[47] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *ICML*, vol. 97, 1997, pp. 412–420.
[48] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *The Journal of machine learning research*, vol. 3, pp. 1289–1305, 2003.
[49] S. Gunal, O. N. Gerek, D. G. Ece, and R. Edizkan, "The search for optimal feature set in power quality event classification," *Expert Systems with Applications*, vol. 36, no. 7, pp. 10 266–10 273, 2009.
[50] L. Yongmin, Z. Weidong, and S. Wenqian, "Improvement of the decision rule in knn text categorization," *Journal of Computer Research and Development*, vol. 42, pp. 378–382, 2005.

PLACE PHOTO HERE

Mohamed Morchid Biography text here.

PLACE PHOTO HERE

Mohamed Bouallegue Biography text here.

Richard Dufour Biography text here.

Georges Linarès Biography text here.

Renato De Mori Biography text here.