I-vector based Representation of Highly Imperfect Automatic Transcriptions

Mohamed Morchid[†], *Mohamed Bouallegue*[†], *Richard Dufour*[†],
Georgès Linares[†], *Driss Matrouf*[†] *and Renato De Mori*[†‡]

[†]LIA, University of Avignon, France
[‡]McGill University, School of Computer Science, Montreal, Quebec, Canada

{firstname.lastname}@univ-avignon.fr, rdemori@cs.mcgill.ca

Abstract

The performance of Automatic Speech Recognition (ASR) systems drops dramatically when used in noisy environments. Speech analytics suffer from this poor quality of automatic transcriptions. In this paper, we seek to identify themes from dialogues of telephone conversation services using multiple topic-spaces estimated with a Latent Dirichlet Allocation (LDA) approach. This technique consists in estimating several topic models that offer different views of the document. Unfortunately, such a multi-model approach also introduces additional variabilities due to the model diversity. We propose to extract the useful information from the full model-set by using an i-vector based approach, previously developed in the context of speaker recognition. Experiments are conducted on the DECODA corpus, that contains records from the call center of the Paris Transportation Company. Results show the effectiveness of the proposed representation paradigm, our identification system reaching an accuracy of 84.7%, with a gain of 3.3 points compared to the baseline.

Index Terms: human/human conversation, speech recognition, Latent Dirichlet Allocation, i-vectors, joint factor analysis

1. Introduction

Automatic Speech Recognition (ASR) systems frequently fail on noisy conditions and high Word Error Rates (WER) make difficult the analysis of the automatic transcriptions. Speech analytics suffer from these transcription issues that may be overcome by improving the ASR robustness or/and the tolerance of speech analytic systems to ASR errors. This paper proposes a new method to improve the robustness of speech analytics by combining a semantic multi-model approach and a nuisance reduction technique based on the i-vector paradigm.

This method is evaluated in the application framework of the RATP call centre (Paris Public Transportation Authority), focusing on the theme identification task [1].

Telephone conversation is a particular case of human/human interaction whose automatic processing encounters many difficulties, especially due to the speech recognition step required to obtain the transcription of the speech contents. First, the speaker behavior may be unexpected and the train/test mismatch may be very large. Second, speech signal may be strongly impacted by various sources of variability: environment and channel noises, acquisition devices…

Topics are related to the reason why the customer called. 8 classes corresponding to the main customer requests are considered (*lost and founds, traffic state, timelines…*). Additionally to the classical problems in such adverse conditions, the topic identification system has also to face issues due to class proximity. For example, a *lost & found* request is related to itinerary (*where the object was lost?*) or timeline (*when?*), that could appear in most of the classes. In fact, these conversations involve a relatively small set of basic concepts related to transportation issues.

An efficient way to tackle both ASR robustness and class ambiguity could be to map dialogues into a topic space abstracting the ASR outputs. Then, dialogue categorization is achieved in this topic space. Numerous unsupervised methods for topic-space estimate were proposed in the past. Latent Dirichlet Allocation (LDA) [2] was largely used for speech analytics; one of its main drawback 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…If the decision process is highly dependent on these features, the system performance could be quite unstable.

Our proposal is to estimate a large set of topic spaces by varying the LDA meta-parameters. The mapping of the document into each of the resulting spaces could be considered as a particular view of the spoken contents. In the topic identification context, this multiple representation of the same dialogue improves the tolerance of the identification system to the recognition errors [3, 4].

On the other hand, multi-view approaches introduce an additional variability due to the diversity of the views. We propose to reduce this variability by using a factor analysis technique, which was developed in the field of speaker identification. 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 considers multiple variabilities that may be cross-dependent. Thereby, JFA [5] representation allows to compensate the variability within session of a same speaker. It is an extension of the GMM-UBM (Gaussian Mixture Model-Universal Background Model) models [6]. In [7], the authors extract from the GMM super-vector, a compact super-vector named i-vector (i for *identification*). The aim of the compression process (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 was not yet used in the field of text classification.

In this paper, we propose to apply factor analysis to compensate nuisible variabilities due to the multiplication of LDA models. Furthermore, a normalization approach, called c-vector (c for *classification*), to condition dialogue representa-

This work was funded by the SUMACC and ContNomina projects supported by the French National Research Agency (ANR) under contracts ANR-10-CORD-007 and ANR-12-BS02-0009.

tions (multi-model and i-vector) is proposed. This multiple representation of a transcription even if the purpose of the application is theme identification and an annotated train corpus is available, supervised LDA [8] is not suitable for the proposed approach since LDA is used only for producing different feature sets used for computing statistical variability models.

Two methods showed improvements for speaker verification: within Class Covariance Normalization (WCCN) [7] and Eigen Factor Radial (EFR) [9] (that includes length normalization [10]). 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 dialogue 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 dialogue (session).

The transcription representation is described in section 2. Section 3 introduces the i-vector compact representation. Sections 4 and 5 report experiments and results before concluding in section 6.

2. Multi-view representation of automatic transcriptions in a homogenous space

The approach considered in this paper focuses on modeling the variability between different views of a same transcription. For this purpose, it is important to select features that represent semantic contents relevant for this transcription. An attractive set of features for capturing possible semantically relevant word dependencies is obtained with LDA [2], a generative probabilistic model for collections of discrete data such as text corpora.

A transcription is then represented as a finite mixture over an underlying set of topics. Given a train set of transcriptions, a hidden topic space is derived and a transcription d is represented by its probability in each hidden space topic. 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. Section 3 describes how the i-vector representation substantiates this claim.

In order to estimate the parameters of different hidden spaces in a homogenous space, a vocabulary V of discriminative words is constructed as described in [11, 3, 4]. Several techniques have been proposed to estimate the LDA parameters, such as Variational Method [2], Expectation-propagation [12] or Gibbs Sampling [8, 13]. Gibbs Sampling is a special case of Markov-chain Monte Carlo (MCMC) [14] and gives a simple algorithm for approximate inference in high-dimensional models such as LDA [13]. This overcomes the difficulty to directly and exactly estimate parameters that maximize the likelihood of the whole data collection defined as: $P(W|\overrightarrow{\alpha}, \overrightarrow{\beta}) = \prod_{\overrightarrow{w} \in W} P(\overrightarrow{w}|\overrightarrow{\alpha}, \overrightarrow{\beta})$ for the whole data collection W knowing the Dirichlet parameters $\overrightarrow{\alpha}$ and $\overrightarrow{\beta}$.

Gibbs Sampling allows both to estimate the LDA parameters, in order to represent a new transcription d with the n^{th} topic space Γ^q_n of size q, and to obtain a feature vector $V^{z^n}_d$ of the topic representation of d. The k^{th} feature $V^{z^n_k}_d = P(z^n_k|d)$ (where $1 \leq k \leq q$) is the probability of topic z^n_k generated by the unseen transcription d in the n^{th} topic space of size q, and $V^{w_i}_{z^n_k} = P(w_i|z^n_k)$ is the vector representation of a word w_i into Γ^q_n.

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 $\{\Gamma^q_n\}^p_{n=1}$ 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 $\alpha_i = \frac{50}{q}$[8], 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 $\alpha_n = \frac{n}{p} \times \frac{50}{q}$.

The larger α_n ($\alpha_n \geq 1$) is, the more uniform $P(z|d)$ will be (see figure 1). 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 1 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 $\alpha_p = 1$).

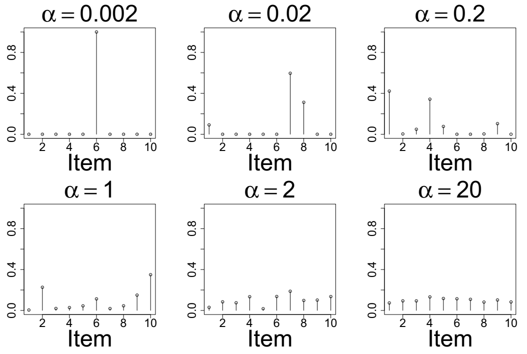

Figure 1: Dirichlet distribution with a varied α_n.

The next process allows to obtain a homogeneous representation of the transcription d for the n^{th} topic space Γ^q_n. The feature vector $V^{z^n}_d$ of d is mapped into the common vocabulary space V composed with a set of $|V|$ discriminative words [11, 3, 4] to obtain a new feature vector [15] $V^w_{d,n} = \{P(w|d)_{\Gamma^q_n}\}_{w \in V}$ of size $|V|$ for the n^{th} topic space Γ^q_n of size q where the i^{th} ($0 \leq i \leq |V|$) feature is:

$$V^{w_i}_{d,n} = \sum_{k=1}^{q} P(w_i|z^n_k)P(z^n_k|d) = \sum_{k=1}^{q} V^{w_i}_{z^n_k} \times V^{z^n_k}_d$$

3. Compact representation

In this section, an i-vector-based method to represent automatic transcriptions, called c-vector, is presented. Initially introduced for speaker recognition, i-vectors [5] have become very popular in the field of speech processing and recent publications show that they are also reliable for language recognition [16] and speaker diarization [17]. I-vectors are an elegant way of reducing the large-dimensional input data to a small-dimensional feature vector while retaining most of the relevant information. The technique was originally inspired by the Joint Factor Analysis framework [18]. Hence, i-vectors convey the speaker characteristics among other information such as transmission channel, acoustic environment or phonetic content of speech segments. Next sections describe the c-vector extraction process, the vector transformation with the EFR method, and the Mahalanobis metric.

3.1. Total variability space definition

The 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. The speech (of given speech recording) super-vector $\mathbf{m}_s$ of concatenated GMM means is projected in a low dimensionality space, named Total Variability space, with $\mathbf{m}_s = m + \mathbf{T}\mathbf{x}_s$, where m is the mean super-vector of the UBM[1]. $\mathbf{T}$ 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 $\mathbf{x}_s$ are the total factors which represent the coordinates of the speech recording in the reduced total variability space called i-vector.

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

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

$$\mathbf{m}_{(d,\Gamma)} = m + \mathbf{T}\mathbf{x}_{(d,\Gamma)} \qquad (1)$$

3.3. C-vector conditioning

In [9], the authors proposed a solution to these 3 raised i-vector issues: (i) the i-vectors x of equation 1 have to be theoretically normally distributed among the normal distribution $\mathcal{N}(0, I)$, (ii) the "radial" effect should be removed, and (iii) the full rank total factor space should be used to apply discriminant transformations. To do so, they apply transformations for train and test transcription representations. The first step is to evaluate the empirical mean $\overline{x}$ and covariance matrix V of the training c-vector. The covariance matrix V is decomposed by diagonalization into PDP^t where P is the eigenvector matrix of V and D is the diagonal version of V. A train c-vector x is transformed to x' as follows:

$$x' = \frac{D^{-\frac{1}{2}} P^t (x - \overline{x})}{\sqrt{(x - \overline{x})^t V^{-1} (x - \overline{x})}} \qquad (2)$$

The numerator is equivalent by rotation to $V^{-\frac{1}{2}}(x - \overline{x})$ and the euclidean norm of x' is equal to 1. The same transformation is applied to the test c-vectors, using the training set parameters $\overline{x}$ and mean covariance Vas estimations of the test set of parameters. Figure 2 shows the transformation steps: figure 2-(a) is the original training set; figure 2-(b) shows the rotation applied to the initial training set around principal axes of the total variability when P^t is applied; figure 2-(c) shows the standardization of c-vectors when $D^{-\frac{1}{2}}$ is applied; and finally, figure 2-(d) shows the c-vector x' on the surface area of the unit hypersphere after a length normalization by a division of $\sqrt{(x - \overline{x})^t V^{-1} (x - \overline{x})}$.

4. Experimental Protocol

The proposed c-vector representation of automatic transcriptions is evaluated in the context of the theme identification of a human/human telephone conversation in the customer care service (CCS) of the RATP Paris transportation system. The Mahalanobis metric is used to associate a theme to a dialogue.

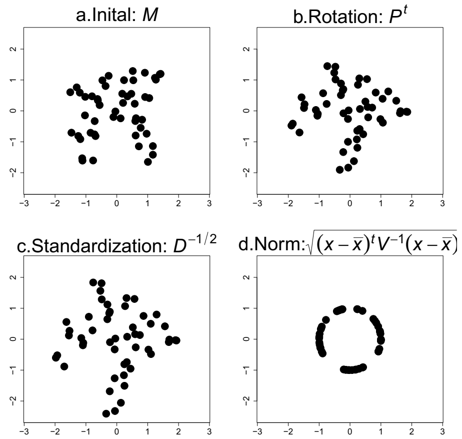

Figure 2: Effect of the standardization with the EFR algorithm.

4.1. Theme identification task

The DECODA project corpus [1] was used to perform experiments on the conversation theme identification. It is composed of 1,514 telephone conversations, corresponding to about 74 hours of signal, split into a train set (740 dialogues), a development set (447 dialogues) and a test set (327 dialogues), and manually annotated with 8 conversation themes: *problems of itinerary*, *lost and found*, *time schedules*, *transportation cards*, *state of the traffic*, *fares*, *infractions* and *special offers*.

A LDA model allowed to elaborate 500 topic spaces with 50 topics by varying the topic distribution parameter $\overrightarrow{\alpha}$. For each theme $\{C_i\}_{i=1}^{8}$, a set of 50 theme specific words is identified. The same word may appear in more than one theme vocabulary selection. All the selected words are then merged without repetition to form V made of 166 words. The topic spaces are made with the LDA Mallet Java implementation[2].

The ASR system used for the experiments is LIA-Speeral [19]. 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 train set transcriptions. This system reaches an overall Word Error Rate (WER) of 45.8%, 59.3%, and 58.0%, respectively on the train, development and on test sets. These high WER are mainly due to speech disfluencies and to adverse acoustic environments (for example, calls from noisy streets with mobile phones). A "stop list" of 126 words[3] was used to remove unnecessary words (mainly function words) which results in a WER of 33.8% on the train, 45.2% on the development, and 49.5% on the test.

4.2. Mahalanobis metric

Given a new observation x, the goal of the task is to identify the theme belonging to x. The 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 dialogue, its representation mechanism is ignored and is regarded as an observation from a

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

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

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

probabilistic generative model. The Mahalanobis scoring metric assigns a dialogue d with the most likely theme C. Given a training dataset of dialogues, let $\mathbf{W}$ denote the within dialogue covariance matrix defined by:

$$\mathbf{W} = \sum_{k=1}^{K} \frac{n_t}{n} \mathbf{W_k} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=0}^{n_t} \left(x_i^k - \overline{x_k}\right) \left(x_i^k - \overline{x_k}\right)^t \quad (3)$$

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 dialogues, and $\overline{x_k}$ is the mean of all dialogues x_i^k of C_k.

Each dialogue does not contribute to the covariance in an equivalent way: the term $\frac{n_t}{n}$ is then introduced in equation 3. 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:

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

where $a_k = \log\left(P(C_k)\right)$. It is noted that, with these assumptions, the Bayesian approach is similar to the Fisher's geometric approach: x is assigned to the nearest centroid's class, according to the Mahalanobis metric [20] of $\mathbf{W}^{-1}$:

$$C_{k_{\text{Bayes}}} = \arg\max_k \left\{ -\frac{1}{2} ||x - \overline{x_k}||^2_{\mathbf{W}^{-1}} + a_k \right\} \quad (4)$$

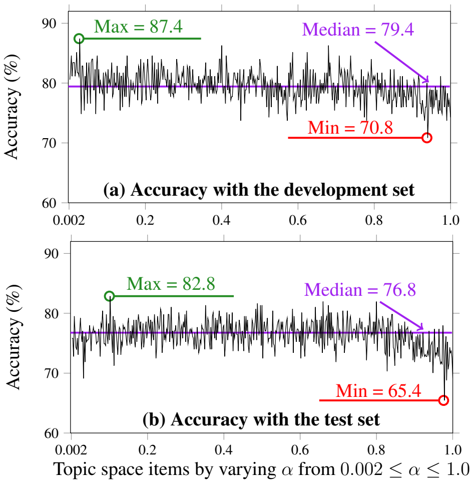

Figure 3: Theme classification rates using various topic-based representations with EFR normalization on the dev and test sets.

5. Results

Classification approaches applied on the same classification task and corpus are proposed in [3] (state-of-the-art in text classification). The best configuration (LDA representation + SVM classification) reaches an accuracy of 81.4%. We propose to consider this work as a baseline system (baseline BL_SVM).

Experiments are conducted using the multiple topic spaces estimated with a LDA approach. From these multiple topic spaces, the classical approach is to find the one that reaches the best performance. Figure 3-(a) presents the theme classification performance obtained on the development and test sets using various topic-based representation configurations with the EFR normalization algorithm (baseline BL_TBR).

First of all, we can see that the baseline BL_TBR reached a classification accuracy of 87.4% on the development set. 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 16.6 points. As a result, finding the best topic space configuration seems crucial for this classification task, particularly in the context of highly imperfect automatic transcriptions. Finally, when comparing results obtained on the development and test sets (figures 3-(a) and (b)), we can see that the best operating point is different: if the one estimated on the development set would be applied to the test set (best operating point), the classification accuracy would reach 75.2% (best development accuracy is reached with $\alpha = 0.024$), while the best potential classification result reaches 82.8%.

Table 1 presents the original c-vector approach coupled with the EFR normalization algorithm. We can firstly note that this compact representation allows to outperform results obtained with the best topic space configuration, with a gain of 1.7 points on the development and of 1.9 points on the test data. The inconsistency of the classification performance is not observed with this approach. Indeed, the configuration that obtained the best accuracy on the dev. set is also the same on the test set. Moreover, if we consider the different c-vector configurations, the gap between accuracies is much smaller: classification accuracy does not go below 82.3%, while it reached 70.8% for the worst topic-based configuration (see figure 3-(a)).

Table 1: Theme classification accuracy (%) using c-vectors.

	DEV			*TEST*		
	Number of Gaussians in GMM-UBM					
c-vector size	*32*	*64*	*128*	*32*	*64*	*128*
60	82.8	88.6	83.4	76.7	83.1	77.0
80	87.4	86.3	87.4	83.4	82.8	74.3
100	82.3	**89.1**	85.1	81.0	**84.7**	72.2
120	82.3	83.0	83.4	78.3	81.3	76.1

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

6. Conclusions

This paper presents an original multi-view representation of highly imperfect dialogue transcriptions, and a fusion process with the use of Factor Analysis. The effectiveness of the proposed approach is evaluated in the task of theme identification. Thus, the architecture of the system identifies conversation themes using an i-vector approach. Originally developed for speaker recognition, we showed that this compact representation can be applied to a text classification task. Indeed, this solution allowed to obtain a better classification accuracy than the use of the classical best topic space configuration. In fact, we highlighted that this original compact version of all topic-based representations of dialogues, named c-vector, coupled with the EFR normalization algorithm, is a better solution to deal with dialogue variabilities (high word error rates, bad acoustic conditions, unusual word vocabulary...). Finally, the classification accuracy reached 84.7% with a gain of 9.5 points with the same configuration (best BL_TBR operating point 75.2%), 1.9 points with best topic space size (82.8%), and 3.7 points with the baseline BL_SVM (81.4%).

7. References

[1] 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.

[2] 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.

[3] 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 *ICASSP*, 2014.

[4] M. Morchid, R. Dufour, and G. Linarès, "A LDA-based topic classification approach from highly imperfect automatic transcriptions," in *LREC'14*, 2014.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] M. Morchid, G. Linarès, M. El-Beze, and R. De Mori, "Theme identification in telephone service conversations using quaternions of speech features," in *INTERSPEECH*, 2013.

[12] 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.

[13] G. Heinrich, "Parameter estimation for text analysis," *Web: http://www. arbylon. net/publications/text-est. pdf*, 2005.

[14] 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.

[15] M. Morchid, R. Dufour, and G. Linarès, "Thematic representation of short text messages with latent topics: Application in the twitter context," in *PACLING*, 2013.

[16] D. Martınez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in ivectors space," *INTERSPEECH*, pp. 861–864, 2011.

[17] 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.

[18] 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.

[19] 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.

[20] 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.