A Comparison of Normalization Techniques Applied to Latent Space Representations for Speech Analytics

Mohamed Morchid, Richard Dufour, Driss Matrouf

LIA - University of Avignon (France)

{mohamed.morchid, richard.dufour, driss.matrouf}@univ-avignon.fr

Abstract

In the context of noisy environments, Automatic Speech Recognition (ASR) systems usually produce poor transcription quality which also negatively impact performance of speech analytics. Various methods have then been proposed to compensate the bad effect of ASR errors, mainly by projecting transcribed words in an abstract space. In this paper, we seek to identify themes from dialogues of telephone conversation services using latent topic-spaces estimated from a latent Dirichlet allocation (LDA). As an outcome, a document can be represented with a vector containing probabilities to be associated to each topic estimated with LDA. This vector should nonetheless be normalized to condition document representations. We propose to compare the original LDA vector representation (without normalization) with two normalization approaches, the Eigen Factor Radial (EFR) and the Feature Warping (FW) methods, already successfully applied in speaker recognition field, but never compared and evaluated in the context of a speech analytic task. Results show the interest of these normalization techniques for theme identification tasks using automatic transcriptions The EFR normalization approach allows a gain of 3.67 and 3.06 points respectively in comparison to the absence of normalization and to the FW normalization technique.

Index Terms: human-human conversation, speech analytics, latent Dirichlet allocation, vector normalization

1. Introduction

Automatic Speech Recognition (ASR) systems, used in noisy acoustic conditions on conversations between humans speaking a spontaneous way, usually produce poor transcription quality. Speech analytics can be impacted by these transcription issues that may be overcome by improving the ASR robustness or/and the tolerance of speech analytics systems to ASR errors.

An efficient way to improve the robustness to ASR errors is to map transcriptions into a topic space abstracting the ASR outputs. Then, this topic space can be used instead of directly using words from transcriptions, for example, in a categorization task [1]. In a previous work, we proposed to use a latent Dirichlet allocation (LDA) [2] topic space approach estimated from automatic transcriptions in order to identify the main theme of human-human conversations [3]. The relevance of two assumptions about the automatic transcription of dialogues has been demonstrated: the Gaussianity of the theme classes (normal distribution) and the equality of the class covariances. Nonetheless, no study about the impact of normalization methods on highly imperfect automatic transcriptions from human-human conversations has been considered.

This work was funded by the ContNomina project supported by the French National Research Agency contract ANR-12-BS02-0009.

In this paper, we propose a comparison of different normalization methods to "Gaussianize" the transcriptions in order to improve the robustness of speech analytics to ASR errors during a theme identification task. Our proposal is to first estimate a topic space from a LDA. Then, each automatic transcription of each dialogue is mapped into this topic space to obtain as an outcome a vectorial representation. At this step, this vectorial representation have to be normalized. We then propose to evaluate different normalization methods successfully applied in the context of speaker verification but never on speech analytics. These normalization methods have shown impressive improvements for speaker verification: Feature Warping Normalization (FW) [4] and Eigen Factor Radial (EFR) [5] (that includes length normalization [6]). This last method dilates the space as the mean to reduce the within theme variability. Experiments are conducted in the application framework of the RATP call-centre (Paris Public Transportation Authority), focusing on the theme identification task [7]. To find out the most related theme to a given dialogue, the Mahalanobis metric [8] is computed. For sake of comparison, experiments will be performed using manual and automatic transcriptions.

Related work is firstly presented in Section 2. The topic-based representation of the transcriptions is described in Section 3. Section 4 introduces the two normalization approaches. Finally, Section 5 reports experiments and results before concluding in Section 6.

2. Related work

The classical Term Frequency-Inverse Document Frequency (TF-IDF) [9] has been widely used for extracting discriminative words. Improvements are observed with the Gini purity criteria [10]. Nonetheless, in the context of speech analytics based on noisy transcriptions, this classical term representation is not robust enough [1]. Other approaches, which proposed to consider the document as a mixture of latent topics, are more suitable to deal with highly imperfect transcriptions. These methods, such as Latent Semantic Analysis (LSA) [11, 12], Probabilistic LSA (PLSA) [13] or latent Dirichlet allocation (LDA) [2], build a higher-level representation of the document in a topic space. Documents are then considered as a bag-of-words [14] where the word order is not taken into account. Latent Dirichlet allocation (LDA) [2] was largely used for speech analytics [15], many previous studies highlighted its high level performance on a theme identification task of conversations [3, 16, 17, 18]. In pattern classification, the problem of undesired variability can be handled by using compensation or normalization. The compensation can be used when the effect of the noise is mathematically known. When there is no knowledge about the effect of the noise, the normalization techniques can be adopted. For example, most state-of-the-art

speech and speaker recognition systems use cepstral mean subtraction to normalize with respect to the channel variability. In speaker recognition domain, one of the first normalization technique which was largely used to compensate cepstral features is Cepstral Mean Subtraction (CMS) [19]. Two normalization techniques have recently been successfully proposed to replace the CMS approach. The first one, called Feature Warping [4], operates in cepstral domain. The second one, called Eigen Factor Radial [5] normalization, operates in i-vector domain.

3. Semantic dialogue representation

The purpose of the considered application is the identification of the major theme of a human-human telephone conversation in the customer care service (CCS) of the RATP Paris transportation system. The approach considered in this paper focuses on modeling the variability between different dialogues expressing the same theme t. For this purpose, it is important to select relevant features that represent semantic content for the theme of a dialogue. An attractive set of features for capturing possible semantically relevant word dependencies is obtained with a latent Dirichlet allocation (LDA) [2], as described in section 2.

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. LDA is used only for producing different feature sets used for computing statistical variability models.

Several techniques, such as Variational Methods [2], Expectation-propagation [20] or Gibbs Sampling [21], have been proposed for estimating the parameters describing a LDA hidden space. Gibbs Sampling is a special case of Markov-chain Monte Carlo (MCMC) [22] and gives a simple algorithm for approximate inference in high-dimensional models such as LDA [23]. This overcomes the difficulty to directly and exactly estimate parameters that maximize the likelihood of the whole data collection defined as: $P(W|\overrightarrow{\alpha},\overrightarrow{\beta}) = \prod_{w\in W} P(\overrightarrow{w}|\overrightarrow{\alpha},\overrightarrow{\beta})$ for the whole data collection W knowing the Dirichlet parameters $\overrightarrow{\alpha}$ and $\overrightarrow{\beta}$.

Gibbs Sampling allows us both to estimate the LDA parameters, in order to represent a new dialogue d with the r^{th} topic space r^n of size n, and to obtain a feature vector $x_d^{z^r}$ of the topic representation of d. The j^{th} feature:

$$x_d^{z_j^r} = P(z_j^r|d)\ , \tag{1}$$

where $1 \leq j \leq n$ and $P(z_j^r|d)$ is the probability of topic z_j^r generated by the unseen dialogue d in the r^{th} topic space of size n (see Figure 1).

4. Normalization methods

The aim of the paper is to provide a robust representation of highly imperfect transcriptions of a given dialogue. In the next subsections, we will describe the manner that two techniques, which was initially designed for speaker verification, are applied to the classification field: the Feature Warping (FW) and Eigen Factor Radial (EFR) normalizations.

4.1. Feature Warping

Feature Warping [4] seeks to map the vectors obtained from the topic-based representation over a specified interval so that accumulated distribution is similar to a normal distribution. This

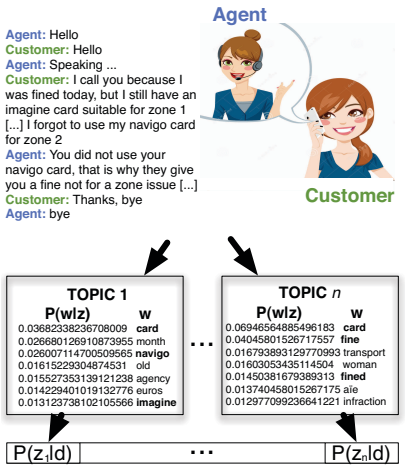

Figure 1: Example of a dialogue d mapped into a topic space of size n.

method allows us to transform the distribution of a given dialogue as a standard normal distribution. Each component in the topic-based representation vector is normalized independently of the others. First, each component in the data vectors is sorted from 1 to N (N is the number of dialogues). Then, a value x with rank R is transformed to obtain m as follows:

$$\frac{N+\frac{1}{2}-R}{N} = \int_{z=-\infty}^{m} h(z)dz \tag{2}$$

where the distribution of a normal curve is given by:

$$h(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \tag{3}$$

The target is to map the distribution of the automatic transcription of a given dialogue to a particular form such that the resulting feature distribution becomes normally distributed. One can refer to [4] for more details about Feature Warping approach.

4.2. Eigen Factor Radial Normalization

The issue worked out in this paper is the fact that the vector representation of a given dialogue has to be distributed among the normal distribution $\mathcal{N}(0, I)$. To do so, we 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 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 vector x is transformed to x' as follows:

$$x' = \frac{D^{-\frac{1}{2}}P^t(x-\overline{x})}{\sqrt{(x-\overline{x})^t V^{-1}(x-\overline{x})}} \tag{4}$$

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 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 vectors when $D^{-\frac{1}{2}}$ is applied; and finally, Figure 2-(d) shows the vector x' on the surface area of the unit hypersphere after a length normalization by a division of $\sqrt{(x-\overline{x})^t V^{-1}(x-\overline{x})}$.

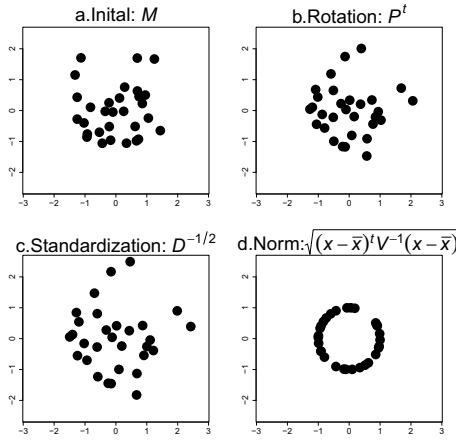

Figure 2: Effect of the standardization with the EFR algorithm.

5. Experiments

5.1. Experimental protocol

The 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 [7] and is used to perform experiments on conversation theme identification. It is composed of 1,242 telephone conversations, which corresponds to about 74 hours of signal. The data set was split as described in Table 1.

Table 1: DECODA dataset.

Class	*Number of samples*		
label	**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**

The ASR system used for the experiments is LIA-Speeral [24]. 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[1] 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. Experiments on manual transcriptions (TRS) will also be performed to better see the impact of the normalization methods on highly imperfect automatic transcriptions (ASR).

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

To find the best operating point (*i.e.* the best topic space configuration), 500 topic spaces are elaborated with a LDA by varying the number n of topics for each topic space from 5 to 505 and using the LDA Mallet Java implementation[2].

To find out the most related theme to a given dialogue, the Mahalanobis metric [8] is computed. In details, the goal of the task is to identify the theme belonging to a new dialogue d. The probabilistic approaches ignore the process by which the vectorial representations x were extracted and they pretend instead they were generated by a prescribed generative model. Once a topic-based representation is obtained from a dialogue, its representation mechanism is ignored and is regarded as an observation from a 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:

$$\mathbf{W} = \sum_{k=1}^{K} \frac{n_t}{n} \mathbf{W_k} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=0}^{n_t} \left(x_i^k - \overline{x_k}\right) \left(x_i^k - \overline{x_k}\right)^t \quad (5)$$

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 vectorial representations x_i^k of C_k.

Each dialogue d does not contribute to the covariance in an equivalent way: the term $\frac{n_t}{n}$ is then introduced in equation 5. 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 [8] of $\mathbf{W}^{-1}$:

$$C_{k_{\text{Bayes}}} = \arg\max_k \left\{ -\frac{1}{2} ||x - \overline{x_k}||^2_{\mathbf{W}^{-1}} + a_k \right\} \quad (6)$$

5.2. Results

Figure 3 shows the accuracies reached with different LDA topic spaces by varying the number of classes contained in these topic spaces from 5 to 505. The curves from 3(a) to (d) represent the accuracies obtained with the original representation without any post-processing normalization phase. Figures 3(e) to (h) show the accuracies using the FW normalization approach on the topic-space vectorial representation. Finally, Figures 3(i) to (l) represent the accuracies resulting from the EFR normalization technique.

Regardless the normalization approach, the first remark is that the accuracies obtained with manual transcriptions (TRS) are always better than those obtained using automatic transcriptions (ASR) for both development and test data sets. As expected, the theme identification performance is quite better on the development set, due to the relative small size of this data set (175 dialogues) comparatively to the test set (327 dialogues).

One can also point out that results are quite unstable from a topic space configuration (number of classes) to another. This

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

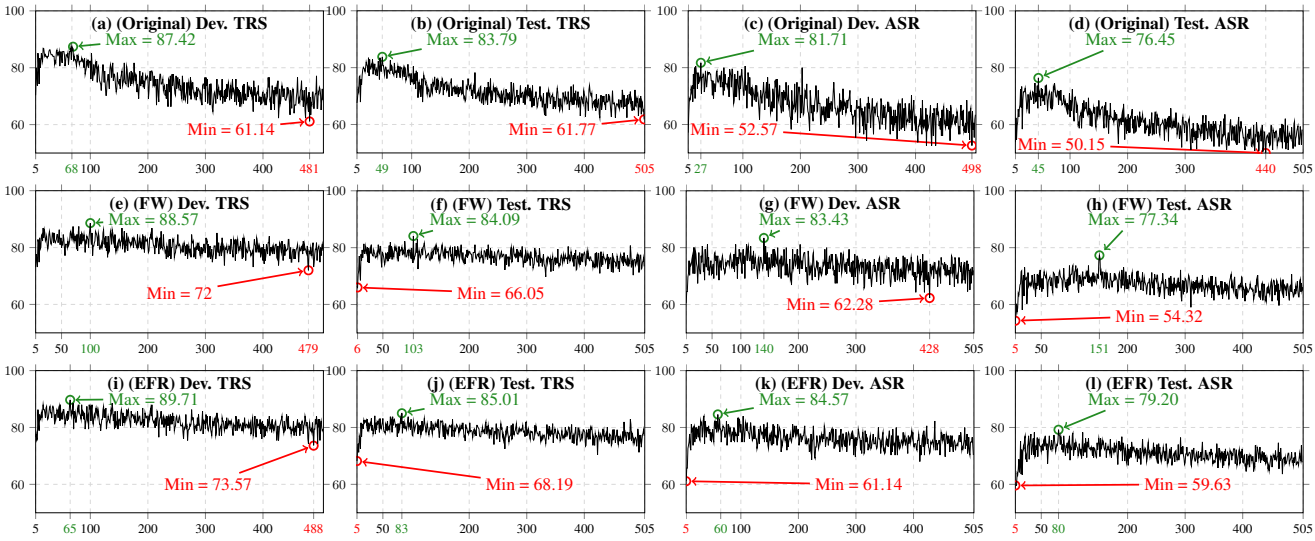

Figure 3: Theme classification accuracies (%) using various topic-based configurations when no normalization is applied (Original - first row), and when FW (second row) and EFR normalizations (last row) are used on the development and test sets. X-axis represents the number n of classes contained into the topic space ($5 \leq n \leq 505$).

phenomenon is due to the LDA initialization phase. Indeed, if the goal is to evaluate the quality of a LDA topic space (with the perplexity for example), the curves will be smoothed. In the case of theme identification task, the "quality" of the words distribution among all classes in the topic model (perplexity) is not correlated with the theme identification accuracy.

To better analyze the results, Table 2 reports robustness of theme identification accuracies. The mean and the standard deviation (std) of the obtained accuracies allow us to evaluate the robustness of each representation (Original dataset, and FW and EFR normalizations). It is straightforward to notice that the most robust representation is the one normalized using the EFR algorithm, with a std of 2.8 and 2.38 for manual and automatic transcriptions respectively. We can particularly figure out the gains (in terms of robustness) observed using automatic transcription (ASR) configuration. These std gains are of 3.16 (2 times smaller) compared to the absence of normalization and 0.57 compared to the FW normalization approach.

Table 2: Mean and standard deviation for Original dataset (OD), Feature Warping (FW), and EFR Standardization algorithms.

Standar. Algo.	DATASET		Dev		Test	
	Train	Test	Mean	Std	Mean	Std
OD	TRS	TRS	74.50	5.63	71.66	4.59
OD	ASR	ASR	66.86	6.19	60.98	5.54
FW	TRS	TRS	76.09	3.05	76.73	3.32
FW	ASR	ASR	70.94	4.96	67.06	2.95
EFR	TRS	TRS	80.55	2.77	82.04	**2.80**
EFR	ASR	ASR	73.31	3.43	78.54	**2.38**

Table 3 sums up the accuracies obtained during the theme identification task of dialogues presented in Figure 3. One can easily notice that the best accuracies are observed with the EFR normalization method using manual transcriptions (TRS) (82.56%) with a respective gain of 4.58 and 2.75 points on the original representation (OD) and the FW normalization approach. The same observation for automatic transcriptions (ASR) (74.31%) can be made with a respective gain of 3.67 and 3.06 points on the absence of normalization (OD) and the FW normalization.

Table 3: Theme classification accuracies (%) for Original dataset (OD), Feature Warping (FW), and EFR Standardization algorithms. **Best** corresponds to the best operating point obtained on the test data, while **Real** corresponds to the one estimated on the development set and applied to the test set.

Standar. Algo.	DATASET		Dev		Test	
	Train	Test	#z	Best	Best	Real
OD	TRS	TRS	*68*	87.42	83.79	77.98
OD	ASR	ASR	*27*	81.71	76.45	70.64
FW	TRS	TRS	*100*	88.57	84.09	79.81
FW	ASR	ASR	*140*	83.43	77.34	71.25
EFR	TRS	TRS	*65*	89.71	85.01	**82.56**
EFR	ASR	ASR	*60*	84.57	79.20	**74.31**

6. Conclusion

In this paper, a comparaison of different normalization approaches to evaluate the relevance of the assumptions made in [3] was presented. We firstly showed that these assumptions are supported, and secondly, the best classification performance being achieved with the Eigen Factor Radial (EFR) normalization approach. Indeed, the classification accuracies reached 82.56% using manual transcriptions and 74.31% using automatic transcriptions, which corresponds to a respective gain of 4.58 and 3.67 points when no normalization is employed (OD), and a respective gain of 2.75 and 3.06 when the Feature Warping (FW) normalization approach is used.

Finally, we showed that the EFR transformation allows us to obtain a more robust representation with a standard deviation (std) of 2.8 and 2.38 points for manual and automatic transcriptions with a respective gain of 0.52 and 0.57 point in comparison to the FW normalization method.

These promising results prompt us to evaluate EFR normalization approach in different tasks using automatic transcriptions, particularly if the representations are highly sensitive to parameters modification such as Support Vector Machines (SVM) or Deep Neural Networks (DNN) based representations.

7. References

[1] Mohamed Morchid, Richard Dufour, and Georges Linares, "A lda-based topic classification approach from highly imperfect automatic transcriptions," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, 2014, pp. 1309–1314.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[3] Mohamed Morchid, Richard Dufour, Pierre-Michel Bousquet, Mohamed Bouallegue, Georges Linarès, and Renato 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.

[4] Jason Pelecanos and Sridha Sridharan, "Feature warping for robust speaker verification," in *A Speaker Odyssey - The Speaker Recognition Workshop*, 2001.

[5] Pierre-Michel Bousquet, Driss Matrouf, and Jean-Franccois Bonastre, "Intersession compensation and scoring methods in the i-vectors space for speaker recognition.," in *INTERSPEECH*, 2011, pp. 485–488.

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

[7] Frederic Bechet, Benjamin Maza, Nicolas Bigouroux, Thierry Bazillon, Marc El-Beze, Renato De Mori, and Eric Arbillot, "Decoda: a call-centre human-human spoken conversation corpus," LREC'12, 2012.

[8] Eric P Xing, Michael I Jordan, Stuart Russell, and Andrew Ng, "Distance metric learning with application to clustering with side-information," in *Advances in neural information processing systems*, 2002, pp. 505–512.

[9] Stephen Robertson, "Understanding inverse document frequency: on theoretical arguments for idf," *Journal of Documentation*, vol. 60, no. 5, pp. 503–520, 2004.

[10] Tao Dong, Wenqian Shang, and Haibin Zhu, "An improved algorithm of bayesian text categorization," *Journal of Software*, vol. 6, no. 9, pp. 1837–1843, 2011.

[11] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[12] Jerome R Bellegarda, "A latent semantic analysis framework for large-span language modeling," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[13] Thomas Hofmann, "Probabilistic latent semantic analysis," in *Proc. of Uncertainty in Artificial Intelligence, UAI ' 99*. Citeseer, 1999, p. 21.

[14] Gerard Salton, "Automatic text processing: the transformation," *Analysis and Retrieval of Information by Computer*, 1989.

[15] Thomas Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 42, no. 1, pp. 177–196, 2001.

[16] Mohamed Morchid, Richard Dufour, Georges Linarès, and Youssef Hamadi, "Latent topic model based representations for a robust theme identification of highly imperfect automatic transcriptions," in *16th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING) 2015*, 2015.

[17] Yannick Estève, Mohamed Bouallegue, Carole Lailler, Mohamed Morchid, Richard Dufour, Georges Linarès, Driss Matrouf, and Renato De Mori, "Integration of word and semantic features for theme identification in telephone conversations," in *International Workshop Series on Spoken Dialogue Systems Technology (IWSDS) 2015*, 2015.

[18] Mohamed Morchid, Mohamed Bouallegue, Richard Dufour, Georges Linarès, Driss Matrouf, and Renato De Mori, "An 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.

[19] Sadaoki Furui, "Cepstral analysis technique for automatic speaker verification," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 29, no. 2, pp. 254–272, 1981.

[20] Thomas Minka and John 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.

[21] Thomas L Griffiths and Mark 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.

[22] Stuart Geman and Donald 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.

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

[24] Georges Linarès, Pascal Nocéra, Dominique Massonie, and Driss Matrouf, "The lia speech recognition system: from 10xrt to 1xrt," in *Text, Speech and Dialogue*. Springer, 2007, pp. 302–308.