Spoken Language Understanding in a Latent Topic-based Subspace

Mohamed Morchid, Mohamed Bouaziz, Waad Ben Kheder, Killian Janod,
Pierre-Michel Bousquet Bousquet, Richard Dufour, Georges Linares

INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Spoken Language Understanding in a Latent Topic-based Subspace

Mohamed Morchid[1], *Mohamed Bouaziz*[1,3], *Waad Ben Kheder*[1],
Killian Janod[1,2], *Pierre-Michel Bousquet*[1], *Richard Dufour*[1], *Georges Linarès*[1]

[1]LIA - University of Avignon (France)
{firstname.lastname}@univ-avignon.fr
[2]ORKIS - Aix en Provence (France)
kjanod@orkis.com
[3]EDD - Paris (France)
mbouaziz@edd.fr

Abstract

Performance of spoken language understanding applications declines when spoken documents are automatically transcribed in noisy conditions due to high Word Error Rates (WER). To improve the robustness to transcription errors, recent solutions propose to map these automatic transcriptions in a latent space. These studies have proposed to compare classical topic-based representations such as Latent Dirichlet Allocation (LDA), supervised LDA and author-topic (AT) models. An original compact representation, called c-vector, has recently been introduced to walk around the tricky choice of the number of latent topics in these topic-based representations. Moreover, c-vectors allow to increase the robustness of document classification with respect to transcription errors by compacting different LDA representations of a same speech document in a reduced space and then compensate most of the noise of the document representation. The main drawback of this method is the number of subtasks needed to build the c-vector space. This paper proposes to both improve this compact representation (c-vector) of spoken documents and to reduce the number of needed sub-tasks, using an original framework in a robust low dimensional space of features from a set of AT models called "Latent Topic-based Subspace" (LTS). In comparison to LDA, the AT model considers not only the dialogue content (words), but also the class related to the document. Experiments are conducted on the DECODA corpus containing speech conversations from the call-center of the RATP Paris transportation company. Results show that the original LTS representation outperforms the best previous compact representation (c-vector), with a substantial gain of more than 2.5% in terms of correctly labeled conversations.
Index Terms: author-topic model, factor analysis, c-vector, document clustering.

1. Introduction

Performance of spoken language understanding applications moves down when dealing with automatically transcribed speech documents in noisy conditions, several word transcription errors being encountered. This is the case of telephone conversations, human/human interactions where automatic processing faces many difficulties, especially due to the speech recognition step required to transcribe the speech contents: the speaker behavior may be unexpected, the mismatch between train/test conditions can be very large, speech signal could be strongly impacted by various sources of variability such as environment and channel noises, acquisition devices... Recent reviews for spoken conversation analysis, speech analytics, topic identification and segmentation can be found in [1, 2, 3, 4, 5] and [6] respectively. Some important problems in finding topic dependent segments are the detection of segment boundaries and modeling the fact that segments may overlap. An efficient way to improve the ASR robustness is to map the conversations in a topic space abstracting the ASR outputs to achieve classification of dialogues in this latent space. Numerous unsupervised topic-spaces were proposed to represent effectively the dialogue content such as Latent Dirichlet Allocation (LDA) [7] or Author-Topic (AT) model [8].

Authors in [9] and [10] have respectively proposed to overcome two drawbacks separately:

- efficiently choosing the size of a topic model by using multiple latent representations obtained by varying the size of the LDA topic space and compacting these representations with the factor analysis [11, 9] (different subprocesses are needed),
- building a topic model, called author-topic (AT) model [8, 10], to take into consideration all information contained into a document: the content itself (*i.e.* words), the label (*i.e.* class) and the relation between the distribution of words and the labels, considered as a latent relation.

Firstly, this paper proposes to jointly overcome these two drawbacks, the tricky choice of the "right" (*i.e.* optimal) size of a topic model and taking into account the label as well as the words contained in the document, by learning a set of topic spaces from an AT model, and then, extracting a compact feature vector from these representations with the factor analysis [11]. This approach requires multiple pre-processing tasks or mappings (deep neural network [12], UBM-GMM, normalization...), best performance being observed on very noisy document representations [13]. Nonetheless, this is not the case with a small representation such as AT model [9] that globally contains low noisy variability. Thus, this paper proposes to secondly consider the different AT spaces as a common homogeneous feature subspace, and to compact these multiple representations (super-vector) to directly extract a robust feature vector. The rest of this paper is organized as follows. The proposed approaches are described in Section 2. Section 3 presents the

This work was funded by the Gafes project supported by the French National Research Agency (ANR) - contract ANR-14-CE24-0022.

http://dx.doi.org/10.21437/Interspeech.2016-50

experimental protocol and reports the results. Finally, Section 5 concludes the work and gives some perspectives.

2. Proposed approach

The proposed original approach, called Latent Topic-based Subspace (LTS), is compared with the classical representation named c-vector. Both learn a set of AT-based spaces detailed in section 2.1, then map each document in each topic space, and finally compress these representations, the c-vector based representation with the factor analysis and the LTS with the Eigen Values Decomposition (EVD). Section 2.2 describes the c-vector approach illustrated in Figure 1-(a)-(b)-(c), while the second approach (LTS) is presented in Section 2.3. In the new LTS technique, the multiple topic spaces are considered as a homogeneous latent subspace, and then avoids us to map the documents in the GMM. Moreover, the super-vectors (concatenation of the representation of the document in each topic space) compose the LTS and are compressed with a straightforward EVD to extract a robust representation of the document. These methods are described in the next sections.

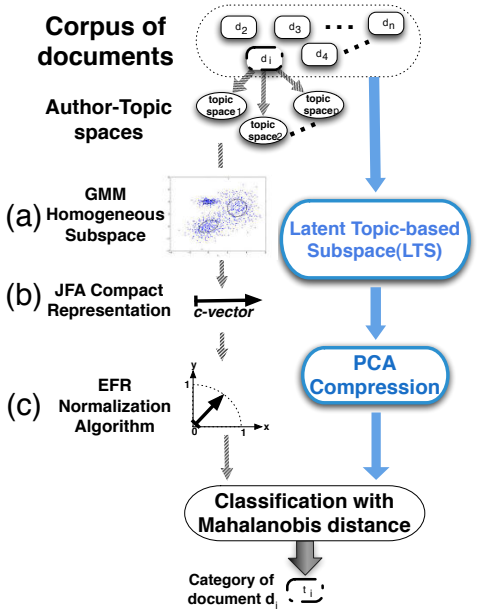

Figure 1: JFA + GMM subspace ((a)-(c)) and the LTS compression (in blue) approaches.

2.1. Author-topic (AT) model

The Author-topic (AT) [8] model codes both the document content (words distribution) and the authors (authors distribution). In our considered application, a document d is a human/human conversation between an agent and a customer. The agent has to label this dialogue with one of the 8 defined themes, a theme being considered as an author. Thus, each dialogue d is composed with a set of words w and a theme a. In this model, each author is associated with a distribution over topics (θ), chosen from a symmetric Dirichlet prior ($\vec{\alpha}$) and a weighted mixture to select a topic z. A word is then generated according to the distribution ϕ corresponding to the topic z. This distribution ϕ is drawn from a Dirichlet ($\vec{\beta}$). Thus, this model allows one to code statistical dependencies between dialogue content (words w) and label (theme a) through the distribution of the latent topics z in the dialogue. Gibbs Sampling allows us to estimate the AT model parameters, in order to represent an unseen dialogue d with the r^{th} author topic space of size T, and to obtain a feature vector $V_d^{a_k^r} = P(a_k|d)$ of the topic representation of an unseen dialogue d with the r^{th} author topic space Δ_r^n of size T. The k^{th} $(1 \leq k \leq A)$ feature is:

$$V_{d,r}^{a_k^r} = \sum_{i=1}^{N_d} \sum_{j=1}^{T} \theta_{j,a_k}^r \phi_{j,i}^r \quad (1)$$

where A is the number of themes; $\theta_{j,a_k}^r = P(a_k|z_j^r)$ is the probability of theme a_k to be generated by the topic z_j^r in the r^{th} topic space of size T. $\phi_{j,i}^r = P(w_i|z_j^r)$ is the probability of the word w_i (N_d is the vocabulary size of d) to be generated by the topic z_j^r.

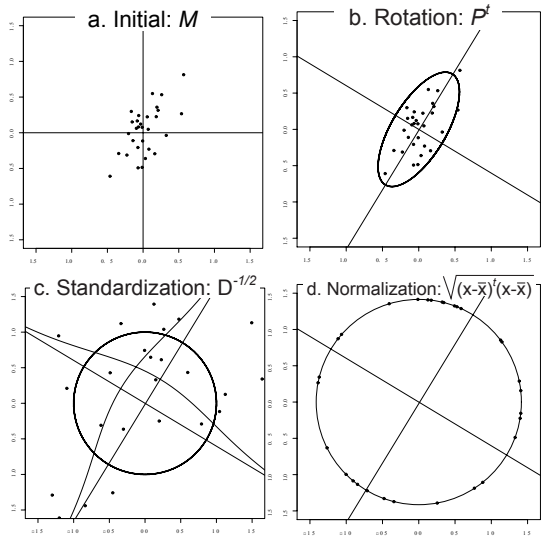

Figure 2: Effect of the standardization with the EFR algorithm.

2.2. C-vector based representation

This approach, initially proposed in [14], uses i-vectors to model dialogue representation through each AT space in a homogeneous space. These short segments are considered as basic semantic-based representation units. In our model, the segment super-vector $\mathbf{m}_{(d,r)}$ of concatenated Gaussian Mixture Model (GMM) means of the representation V_d^a of a transcription d knowing a topic space r is modeled with:

$$\mathbf{m}_{(d,r)} = m + \mathbf{T}\mathbf{x}_{(d,r)} \quad (2)$$

where $\mathbf{x}_{(d,r)}$ contains the coordinates of the AT-based representation of the dialogue in the reduced total variability space called c-vector; m is the mean super-vector of the UBM[1]. $\mathbf{T}$ is the *Total Variability matrix* of low rank $(MD \times R)$, where M is the number of Gaussians in the UBM and D is the feature size. C-vector representation suffers from 3 raised issues: (i) the c-vectors x of equation 2 have to be theoretically distributed among the normal distribution $\mathcal{N}(0, I)$, (ii) the "radial" effect should be removed, and (iii) the full rank total factor space

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

should be used to apply discriminant transformations. The solution to raise these 3 problems has been developed in [13] named "Eigen Factor Radial" (EFR) algorithm by standardizing the c-vectors as described in Figure 2.

2.3. Latent Topic-based Subspace (LTS)

The c-vector representation needs to map dialogues into a UBM-GMM to obtain a super-vector of high dimension (size of the topic-based representation multiplied by the number of Gaussians in the UBM). The Latent Topic-based Subspace (LTS) is composed with a set of latent spaces, and considers each latent-space as a sub area where each document is mapped. Thus, all topic-based representations of a document share a common latent structure. These shared latent parameters define the latent topic-based subspace. Each super-vector s_d of a given document d from the document dataset of size N, is partially associated with a small subset of latent features and the residual part of this document representation is mapped in a global features space shared by all representations that define the latent subspace. The super-vector $\mathbf{s}_d$ of a given dialogue d, is obtained by concatenating the AT-based representations $V_{d,r}^{a_k^r}$ for all r topic spaces. Thus, the matrix of super-vectors $\mathbf{S} = [s_0, \ldots, s_d, \ldots, s_N]$ represents the documents in the LTS. This matrix $\mathbf{S}$ of super-vectors $\mathbf{s}_d$ is then compressed with an EVD to obtain, as an outcome, a short representation $\mathbf{h}_d$ in a low dimensional space with a size depending on the number of eigenvalues e considered:

$$\mathbf{S} = \mathbf{P}\Delta\mathbf{V}^T \tag{3}$$

where $\mathbf{P}$ is a $MD \times N$ matrix of left singular vectors, $\mathbf{V}$ is the $N \times N$ ($N << MD$) matrix of right singular vectors and $\mathbf{\Delta}$ is the diagonal matrix of singular values. N is the rank of the matrix $\mathbf{S}$. More information about EVD can respectively be found in [15] and in [16]. The compact representation $\mathbf{h}_{(d,e)}$ of size e (number of eigenvalues considered) of a super-vector $\mathbf{s}_d$ from $\mathbf{S}$, is defined as follows:

$$\mathbf{h}_{(d,e)} = (\mathbf{s}_d - \overline{s}) \, . \mathbf{V}_e^T \tag{4}$$

where $\mathbf{V}_e$ is the reduced eigenvectors matrix with respect to the e highest eigenvalues contained in the diagonal matrix Δ, and $\overline{s}$ is the centroid (mean) of all super-vectors of the documents contained in the dataset. Moreover, this compact representation of a document based on the LST does not need to: 1) learn a common space such as UBM-GMM, the topic spaces being our homogeneous features space (Figure 1-(a) and (b)); 2) normalize the super-vector with the EFR algorithm (or any other normalization technique) (Figure 1-(c)).

3. Experimental Protocol

The effectiveness of the proposed compact representation in the Latent Topic-based Subspace (LTS) is evaluated in the application framework of the DECODA corpus [17]. 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 (175 dialogues) and a test set (327 dialogues), and manually annotated with 8 ($A = 8$) conversation themes (or authors a in the AT model): *problems of itinerary*, *lost and found*, *time schedules*, *transportation cards*, *state of the traffic*, *fares*, *infractions* and *special offers*.

Transcription of dialogues has been made by the LIA-Speeral ASR system [18]. 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. A "stop list" of 126 words[2] 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. These high WER are mainly due to speech disfluencies and adverse acoustic environments (for example, calls from noisy streets with mobile phones).

A classification approach based on Mahalanobis distance [19] is performed to find out the main theme of a given dialogue. This probabilistic approach ignores the process by which vectors were extracted. Once a compact 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 to the most likely theme C. Given a training dataset of documents, let $\mathbf{W}$ denote the within-document covariance matrix defined by:

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

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 documents, and $\overline{x_k}$ is the centroid (mean) of all documents x_i^k of C_k. Not every document contributes to the covariance in an equivalent way. For this reason, the term $\frac{n_t}{N}$ is 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:

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

where $a_k = \log\left(P(C_k)\right)$. 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 [20] of $\mathbf{W}^{-1}$:

$$C_{k_{\text{Bayes}}} = \arg\max_k \left\{ -\frac{1}{2} ||x - \overline{x_k}||^2_{\mathbf{W}^{-1}} + a_k \right\} \tag{7}$$

4. Experiments and Results

Section 4.1 presents the results obtained with two different dialogue representations based on AT models and the c-vector technique. Then, the proposed original compact representation based on a Latent Topic-based Subspace (LTS) is compared to this c-vector representation in Section 4.2.

4.1. Impact of c-vector compression

Experiments are conducted using 500 AT spaces by varying the number of topics from 5 to 505 (step of 1 topic). From these multiple topic spaces, a classical way is to find the one that reaches the best performance. Figure 3 presents the theme classification performance obtained on the development (Figure 3-(a)) and test (Figure 3-(b)) sets using various AT-based representation configurations (*baseline*).

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

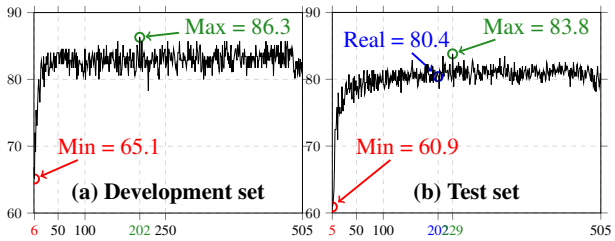

Figure 3: Theme classification accuracies (%) using various author topic-based representations on the development and test sets with different experimental configurations. X-axis represents the number n of classes contained into the topic space ($5 \leq n \leq 505$).

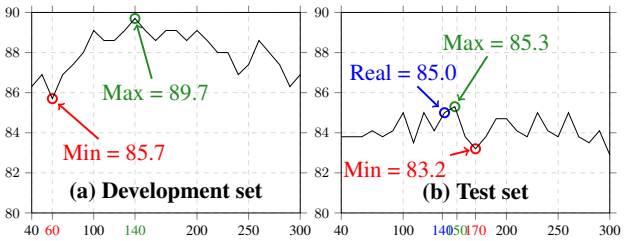

Figure 4: Theme classification accuracies (%) using a compact vector from the LTS on the development and test sets with different experimental configurations. X-axis represents the number of eigenvalues m considered ($40 \leq m \leq 300$).

Firstly, we can see that the baseline approach reached an accuracy of 86.3% and 83.8% on the development and test sets respectively. Nonetheless, one can 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 21.2 points. As a result, finding the best author-topic (AT) space configuration seems crucial for this classification task, particularly in the context of highly imperfect automatic transcriptions. Note that if the operating point estimated on the development set would be applied to the test set (best operating point), the classification accuracy would reach 80.4% on the test set (best development accuracy is reached with $n = 202$ topics), while the best potential classification result reaches 83.8%.

Table 1 presents the c-vector representation coupled with the EFR normalization algorithm [19]. We can firstly notice that this compact representation allows us to outperform results obtained with the best AT model, with a gain of 1.9 points on the test set. The inconsistency of the classification performance is not observed with this approach, as already observed in a previous work [9]. 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 78.9% with the test set, while it reached 60.9% for the worst AT configuration (see Figure 3-(b)).

Table 1: Theme classification accuracy (%) in the total variability space with different UBM and c-vector sizes.

size of the c-vector	*DEV*			*TEST*		
	Number of Gaussians in the GMM-UBM					
	32	*64*	*128*	*32*	*64*	*128*
80	80.6	82.3	83.1	79.2	81.0	80,4
100	81.7	**84.6**	83.1	78.9	**82.3**	80.4
120	84.0	81.7	82.3	80.4	79.2	81.8

4.2. Impact of LTS compression

Results obtained using the original Latent Topic-based Subspace (LTS) representation are shown in Figure 4. In order to better compare performance obtained by all approaches (AT/c-vectors/LTS), best results are reported in Table 2. It is worth emphasizing that these results are given in "real" application condition, *i.e.* the best configuration (number of topics contained into the topic space) being chosen with the development set. As a result, a better operating point could exist in the test set, which could explain the performance gap between results reported in Table 2, and Figures 3 and 4. We can firstly point out that the LTS representation outperforms both the AT baseline and c-vector representations, no matter the corpus studied (development in Figure 4-(a) or test in Figure 4-(b)) with an accuracy of 89.7% (+2.7 points) and 85.3% (+4.6 points) for development and test sets respectively. Another interesting point is the stability and robustness of the LTS model curves, comparatively to the c-vector representation. Indeed, the gap between the lowest and highest values is equal to 3.3 points for the c-vector and 2.1 points for the LTS representation on the test set.

Table 2: Theme classification accuracy (%) using best configuration from development set applied to test set.

Document representation	**Dev.**		**Test**
	size	**acc. %**	**acc. %**
AT-Model (baseline)	202	86.3	80.4
AT-Model + c-vector	100	84.6	82.3
AT-Model + LTS	140	89.7	**85.0**

5. Conclusion

ASR systems performance strongly depends to the recording environment, spoken language understanding tasks being impacted by transcription quality. This paper proposes an elegant way to deal with ASR errors by mapping a dialogue into a robust features subspace called Latent Topic-based Subspace (LTS). Experiments conducted on a classification task of conversations showed the effectiveness of the proposed LTS model in comparison to the use of the classic c-vector and AT model representations. This high-level representation allows us to significantly improve the performance of the theme identification task compared to the previous best results obtained, with a gain of more than 3 and 2 points respectively using the AT model and the c-vector based representations. The LTS model is combined with a PCA compression process, due to the small size of the corpus (740 dialogues in the training corpus). In a future work, it will be interesting to evaluate this promising representation with larger datasets than the set of dialogues from the DECODA project. Thus, other compression methods such as auto-encoder or Probabilistic PCA could give better results during different speech analytics tasks.

6. References

[1] J. Eisenstein and R. Barzilay, "Bayesian unsupervised topic segmentation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 2008, pp. 334–343.

[2] K. Lagus and J. Kuusisto, "Topic identification in natural language dialogues using neural networks," in *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 95–102. [Online]. Available: http://www.aclweb.org/anthology/W02-1014

[3] G. Tur and R. De Mori, *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons, 2011.

[4] T. Hazen, "Topic identification," *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pp. 319–356, 2011.

[5] I. Melamed and M. Gilbert, "Speech analytics," *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pp. 397–416, 2011.

[6] M. Purver, "Topic segmentation," *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pp. 291–317, 2011.

[7] 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.

[8] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. AUAI Press, 2004, pp. 487–494.

[9] M. Morchid, M. Bouallegue, R. Dufour, G. Linarès, D. Matrouf, and R. De Mori, "Compact multiview representation of documents based on the total variability space," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 8, pp. 1295–1308, 2015.

[10] M. Morchid, R. Dufour, M. Bouallegue, and G. Linarès, "Author-topic based representation of call-center conversations," in *International Spoken Language Technology Workshop (SLT) 2014*. IEEE, 2014.

[11] 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.

[12] C. May, F. Ferraro, A. McCree, J. Wintrode, D. Garcia-Romero, and B. Van Durme, "Topic identification and discovery on text and speech," 2015.

[13] 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.

[14] M. Morchid, M. Bouallegue, R. Dufour, G. Linarès, D. Matrouf, and R. De Mori, "I-vector based representation of highly imperfect automatic transcriptions," in *Conference of the International Speech Communication Association (Interspeech) 2014*. ISCA, 2014.

[15] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[16] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," *Numerische mathematik*, vol. 14, no. 5, pp. 403–420, 1970.

[17] 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.

[18] 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.

[19] 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.

[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.