SOFT NONNEGATIVE MATRIX CO-FACTORIZATION WITH APPLICATION TO MULTIMODAL SPEAKER DIARIZATION

N. Seichepine[⋆] *S. Essid*[⋆] *C. Févotte*[†] *O. Cappé*[‡]

[⋆]Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI
[†]Laboratoire Lagrange (CNRS, OCA & University of Nice) [‡]CNRS LTCI, Télécom ParisTech

ABSTRACT

This paper presents a new method for bimodal nonnegative matrix factorization (NMF). This method is well-suited to situations where two streams of data are concurrently analyzed and are expected to be related by *loosely common* factors. It allows for a *soft co-factorization*, which takes into account the relationship that exists between the modalities being processed, but returns different factors for distinct modalities. There is no need that the data related with each modality live in the same feature space; there is also no need that they have the same dimensionality. The co-factorization is obtained via a majorization-minimization (MM) algorithm. The behavior of the method is illustrated on both synthetic and real-world data. In particular, we show that exploiting the correlation between audio and video modalities in edited talk-show videos improve speaker diarization results.

***Index Terms*—** Nonnegative matrix factorization, co-factorization, multimodality, speaker diarization

1. INTRODUCTION

This work is concerned with data analysis tasks where observations are available from two concurrent streams of information (modalities) that exhibit some relationship, typically a strong correlation, without necessarily being of the same nature (for example, observations in different modalities do not necessarily live in the same feature space). This is for instance the case in the task of *multimodal speaker diarization* [1, 2], that jointly exploits audio and video tracks to improve speaker diarization results. Speaker diarization consists in identifying homogeneous segments, according to the speaker identity: the objective is to find "who spoke when". In the specific case of edited videos, where a human meaningfully assembles audio and multiview video tracks, it is obvious that both tracks are related (the director *generally* chooses to show the current speaker). Classic speaker diarization methods generally rely on Gaussian mixture models (GMMs) and variants of Bayesian Information Criterion [3, 4]. However, it has been recently shown that both speaker diarization and onscreen person spotting tasks can be effectively performed using nonnegative matrix factorization (NMF) [5], opening the door to NMF-based multimodal speaker diarization.

Recent works have already exhibited some methods to jointly factorize different streams of information, in situations where each stream of information is naturally represented as an array. For example, [6, 7] propose solutions to jointly factorize data of different dimensionalities. Closer to our work, [8] alternatively solves and initializes two NMF problems, one per modality. Also [9] proposes a general framework to solve an arbitrary number of related NMF problems. All these methods have even been used to solve real-world problems [10, 11, 12], in which co-factorization is defined as a joint factorization, with a shared factor matrix, of different streams of information.

However, *all these methods* presuppose the existence of *common underlying factors*, possibly noisy, shared by all modalities. Therefore, they return common factors while factorizing distinct modalities. The proposed method, on the contrary, *informs* the factorization of each modality with the factorization of the other, but resulting factorizations do not include identical common factors; it is even possible to control *how much* each modality influences the other, or to use a soft ℓ_1-coupling. As such, our approach can be seen as an *unsupervised* version of a multi-task learning problem [13]. Yet, our approach offers some flexibility that is well-suited for our practical application, consisting in softly co-factorizing audio and video tracks of edited videos, for speaker diarization.

Section 2 describes our model and a majorization-minimization (MM) algorithm for estimation. Section 3 is devoted to the application: firstly, the behavior of the algorithm is illustrated on synthetic data, then we discuss tuning of the hyperparameters, and finally we apply the proposed algorithm to a real-world speaker diarization setting.

2. MODEL AND OPTIMIZATION

We first describe the novel penalized NMF framework that formalizes the soft co-factorization task, before an optimization algorithm is exposed.

C. Févotte acknowledges support of project ANR-09-JCJC-0073-01 TANGERINE (Theory and applications of nonnegative matrix factorization).

2.1. General framework

Given a matrix $V \in \mathbb{R}_+^{F\times N}$, the problem of NMF consists in finding two matrices $W \in \mathbb{R}_+^{F\times K}$ and $H \in \mathbb{R}_+^{K\times N}$ such that $V \simeq WH$. N is the number of observations (vectors) and F is the number of features describing each observation. Given a measure of fit D, NMF can be expressed as the minimization problem $\min\ D(V\,|\,WH)$, with respect to W and H, and under the nonnegativity constraints $W \geq 0$ and $H \geq 0$.

Now, given two matrices V_1 and V_2, two measures of fit D_1 and D_2, a penalization function P, and weighting hyper-parameters β_1, β_2 and β_j, we propose to formalize the soft nonnegative matrix co-factorization (sNMcF) problem via the following program:

$$\begin{aligned}\min_{W_1,H_1,W_2,H_2} \beta_1 D_1(V_1\,|\,W_1H_1) + \beta_2 D_2(V_2\,|\,W_2H_2)\\ + \beta_j P\left(W_1,H_1,W_2,H_2\right),\\ \text{s.t.}\quad W_1\geq 0,\, H_1\geq 0,\, W_2\geq 0,\, H_2\geq 0. \quad (1)\end{aligned}$$

2.2. Proposed model

The proposed framework is very general (and stands for many problems) and choices must be made for D_1, D_2 and P for particular instantiations. Given the application-specifics of Section 3, we opt for the generalized Kullback-Leibler (KL) divergence $D_{KL}\left(x|y\right) = x\log\left({}^{x}/_{y}\right) - x + y$ as the measure of fit. The KL divergence is well-suited for multinomial distributions [14, 15], and hence for the histograms data used in Section 3. Furthermore, we consider that V_1 and V_2 are related through H_1 and H_2. W_1 and W_2 act as a dictionary of patterns characteristic of each modality, and the activation matrices H_1 and H_2 are assumed to be "similar". More precisely, we choose to penalize $H_1 - H_2$ by its ℓ_1-norm. This choice means that the differences between H_1 and H_2 are sparse (i.e., often zero).

As such, a naive implementation of our soft co-NMF would aim at the following:

$$\begin{aligned}\min_{W_1,H_1,W_2,H_2} C\left(W_1,H_1,W_2,H_2\right) = \beta_1 D_{KL}(V_1\,|\,W_1H_1)\\ + \beta_2 D_{KL}(V_2\,|\,W_2H_2) + \beta_j\left\|H_1 - H_2\right\|_1,\\ \text{s.t.}\quad W_1\geq 0,\, H_1\geq 0,\, W_2\geq 0,\, H_2\geq 0. \quad (2)\end{aligned}$$

However, this straight implementation is not a viable one, for the following reasons.

1. Because of the scale ambiguity between the dictionaries and the activation matrices, the minimization of the cost function leads to degenerate solutions.[1]
2. H_1 and H_2 have no reason to have similar scales; they are simply expected to have similar *shapes* and should therefore be rescaled prior to comparison.
3. It is not directly possible to handle situations where $H_1 \in \mathbb{R}^{K_1\times N}$ and $H_2 \in \mathbb{R}^{K_2\times N}$ with $K_1 \neq K_2$.

The situation where $K_1 \neq K_2$ is readily handled by ignoring the penalty term for rows of H_1 and H_2 that are unrelated. Thus, without loss of generality, we will use $K_1 = K_2 = K$ in the following. The two other difficulties are solved as follows. First, we introduce the diagonal matrices $\Lambda_1, \Lambda_2 \in \mathbb{R}^{K\times K}$ with k-th diagonal coefficient $\lambda_{1,k} = \sum_f w_{1,fk}, \lambda_{2,k} = \sum_f w_{2,fk}$ and the diagonal matrix $S \in \mathbb{R}_+^{K\times K}$ with k-th diagonal coefficient s_k. Then, we reformulate the program as:

$$\begin{aligned}\min_{W_1,H_1,W_2,H_2,S} C\left(W_1,H_1,W_2,H_2\right) = \beta_1 D_{KL}(V_1\,|\,W_1H_1)\\ + \beta_2 D_{KL}(V_2\,|\,W_2H_2) + \beta_j\left\|\Lambda_1H_1 - S\Lambda_2H_2\right\|_1,\\ \text{s.t.}\quad W_1\geq 0,\, H_1\geq 0,\, W_2\geq 0,\, H_2\geq 0. \quad (3)\end{aligned}$$

With these changes our approach becomes well conditioned: Λ_1 and Λ_2 prevent any cost improvement simply related with the scale, and S scales H_2 so that it can be compared to H_1. S can be obtained in closed form given iterates of H_1 and H_2 and does not incur extra difficulties.

2.3. Optimization algorithm

Optimization of function (3) is not straightforward and we resort to a block-coordinate MM algorithm [16, 17] that updates H_1, H_2, W_1 and W_2 sequentially. The MM framework relies on the construction of an easier-to-minimize upper-bound of the original cost function that is tight at the current parameter value. The upper bound is minimized in lieu of the original cost function, which is in turn decreased at each iteration.

For the sake of conciseness, we here report the update of only one of the activation matrices, H_1, which, to the best of our knowledge, is a novel problem. The update of H_2 is essentially identical, while the updates of W_1 and W_2 amount to special cases of the literature on NMF, see, e.g., [5]. Full derivations as well as MATLAB codes are given online.[2]

Let us first introduce the following notation: $\Psi_1 = \tilde{H}_1 \cdot *(W_1^T(V \cdot /(W_1\tilde{H}_1)))$ is a matrix, with coefficients $\psi_{1,kn}$. The symbols $\cdot *$ and $\cdot /$ refer to element-wise multiplication and division, respectively, and $\tilde{H}_1$ refers to the current value of H_1. Thanks to the convexity of $D_{KL}\left(V\,|\,.\right)$, and using Jensen's inequality, we can obtain the following majoring function of $C(W_1,H_1,W_2,H_2)$:

$$\begin{aligned}G(H_1|\tilde{H}_1) = \sum_{k=1}^{K}\sum_{n=1}^{N}\left(-\psi_{1,kn}\log h_{1,kn} + \lambda_{1,k}h_{1,kn}\right)\\ + \beta_j\sum_{k=1}^{K}\sum_{n=1}^{N}\left|\lambda_{1,k}h_{1,kn} - s_k\lambda_{2,k}h_{2,kn}\right| + \text{cst}. \quad (4)\end{aligned}$$

[1] For any $0<\alpha<1$, $C\left(W_1/\alpha, \alpha H_1, W_2/\alpha, \alpha H_2\right)$ is always lower than $C\left(W_1,H_1,W_2,H_2\right)$.

[2] http://perso.telecom-paristech.fr/~seichepi/icassp2013

For $h_{1,kn} \neq 0$, the derivative of G w.r.t to $h_{1,kn}$ writes

$$\nabla_{h_{1,kn}} G(H_1|\tilde{H}_1) = \frac{-\psi_{1,kn}}{h_{1,kn}} + \lambda_{1,k} + \beta_j \lambda_{1,k} \text{sign}\left(\lambda_{1,k} h_{1,kn} - s_k \lambda_{2,k} h_{2,kn}\right). \quad (5)$$

The gradient is cancelled out by a unique solution $\bar{h}_{1,kn}$, corresponding to the minimum of the auxiliary function, and therefore giving the update rule for $h_{1,kn}$. The resolution involves the critical point $h_c = \frac{s_k \lambda_{2,k} h_{2,kn}}{\lambda_{1,k}}$, and we must take several possibilities into account, depending on the behavior of ∇G on the left and right of h_c. Simple algebra leads to

$$\begin{cases} \bar{h}_{1,kn} = h_c, \text{if } \nabla G(h_c^-) \leq 0 \;\&\; \nabla G(h_c^+) \geq 0 \\ \bar{h}_{1,kn} = \frac{\psi_{1,kn}}{\lambda_{1,k}(1+\beta_j((\nabla G(h_c^+)<0)-(\nabla G(h_c^-)>0)))}, \text{otherwise.} \end{cases} \quad (6)$$

where $\nabla G(h_c)$ is a shortand for $\nabla_{h_{1,kn}} G$ evaluated at $h_{1,kn} = h_c$, and h_c^- and h_c^+ denote left and right neighborhoods of h_c, respectively. The notations $(\nabla G(h_c^+) < 0)$ and $(\nabla G(h_c^-) > 0)$ denote the value 0 or 1 resulting from the boolean test.

3. APPLICATIONS

In this section, our proposed algorithm is first tested on synthetic data, then applied to the joint audio/video diarization task.

3.1. Behavior on synthetic data

We first verify on a toy example that the algorithm behaves as expected. This also gives an illustration of the possibilities of the proposed soft co-factorization. Experiments have been carried out as follows. Firstly, we generate H_1 and $H_2 \in \mathbb{R}^{2\times 240}$ using patterns of zeros and ones. Then, we generate W_1 and $W_2 \in \mathbb{R}^{20\times 2}$ with uniform noise on $[1, 11]$. We finally generate $V_1 = W_1 H_1$ and $V_2 = W_2 H_2$. W_1 and W_2 are here fixed, and the algorithm is initialized with random values for H_1 and H_2. Figure 1 shows that when the coupling parameter β_j is increased, the algorithm returns activation patterns for the first modality that move away from its ground truth and become closer to the ground truth of the second modality, and conversely. It has also been verified that the activation patterns associated with the modality with the highest weight (for instance, β_1) are less distorted than the activation patterns associated with the modality with the lowest weight (for instance, β_2). These are desirable behaviors, that allow us to obtain activation patterns for one modality, *arbitrarily influenced* by the other modality.

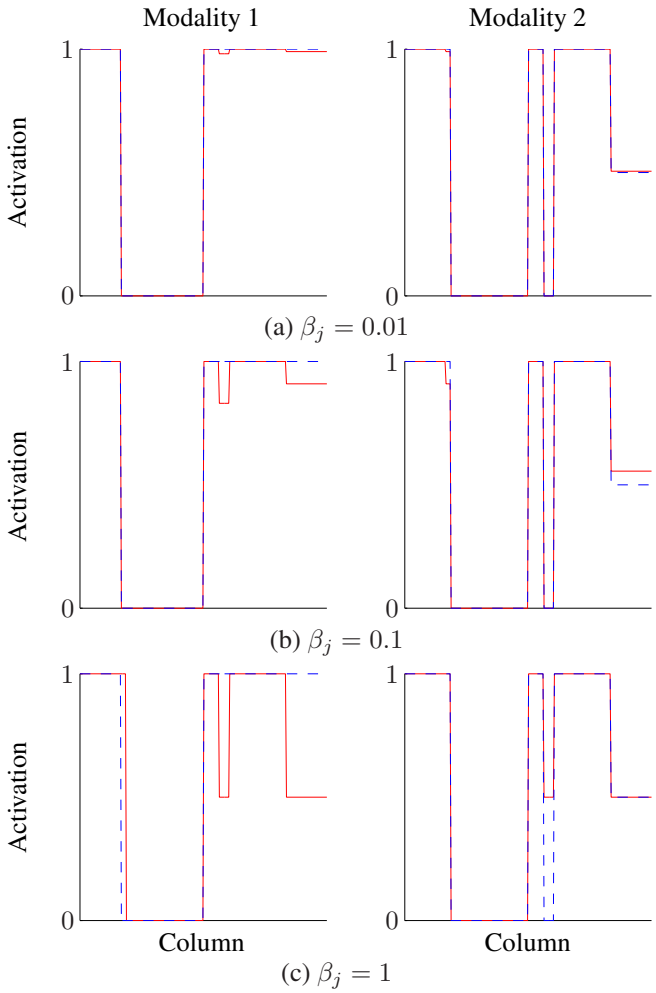

Fig. 1. Influence of the coupling: left and right columns correspond to the first and the second modality, respectively. Continuous lines are the activation patterns returned by the algorithm, while dashed lines correspond to the ground truth. For simplicity, only one row per modality is displayed.

3.2. Hyperparameters

The cost function (3) comprises three hyperparameters. However, the parametrization as presented until now is redundant and we can choose $\beta_1 = 1$ without loss of generality. Then, β_2 simply represents the weight given to the second modality, controls the respective importance of the modalities, and in particular decide whether one modality is more important than the other. However, D_{KL} is not a scale-invariant measure. Therefore, if (for example) $\|V_1\|_1 = 2\|V_2\|_1$ we must correct the weighting parameters and (here) double β_2 to balance the scale effect. Besides, it is difficult to decide *a priori* whether two modalities will be loosely or highly related, and β_j should typically be estimated by cross-validation or from development data.

β_j	0.01	0.1	1
Mean score	22.2	20.3	42.2

Table 1. Mean results for different values of β_j (training set).

3.3. Experimental setup for speaker diarization

In edited videos, current speakers are *generally* onscreen. Identifying onscreen persons therefore provides relevant information for the audio speaker diarization task. While factorizing audio features V_{audio} with NMF, each row of H_{audio} will match a speaker and activation patterns will represent speech segments. For video features V_{video}, each row of H_{video} will match one person and activation patterns will represent onscreen appearances. Therefore, H_{audio} and H_{video} will clearly be related, but not equal. We then simply want to take H_{video} into account to help in situations where the speaker diarization task is difficult, and our algorithm is precisely suited to this task.

We use the 33 first videos of *Canal9 political debates database* [18]. This dataset is made of several broadcasts, featuring a moderator and 2 to 4 guests debating a political question. Diarization is tested on one 8-minute long video segment per video.[3] All scores are computed using the NIST scoring script for speaker diarization evaluation [19]. The evaluation metric is hence the Diarization Error Rate (DER), which is roughly a measure of the fraction of speaker time that is not attributed to the right speaker. Matrices of features V_{audio} and V_{video} are built according to [5]. Each column of V_{video} corresponds to an histogram of visual words, while columns of V_{audio} are made up of histograms of audio states, inferred from short-term Mel Frequency Cepstral Coefficients. Only slight modifications have been made: histograms are built using 50 states per speaker, and an aggregation window of 2 seconds is used.

We define $V_1 = V_{video}$, $V_2 = V_{audio}$, and set without loss of generality $\beta_1 = 1$, see Section 3.2. We set $\beta_2 = 5$, which gives priority to the audio factorization. Finally, the value $\beta_j = 0.1$ is chosen after testing different values on 10 development videos, see Table 1. These videos have been randomly selected[4] among the 33 considered, and are not used for test. Since the optimized cost function is not convex, results may vary over initializations. Hence the tests have been made using 15 random initializations for each video, which is sufficient in practice to have representative results; only results associated with the lowest end cost function value are considered; the score values over all videos are then averaged. Finally, denoting by Q the number of speakers – assumed known, a sensible assumption for TV contents, that typically provide integrated subtitles or teletext – V_{audio} is factorized with $K_{audio} = Q$, while V_{video} is factorized with $K_{video} = Q + 1$. The latter accounts for one component per speaker plus a component for wide shots, see [5].

[3] Starting at 3 minutes and 30 seconds to avoid opening credits.

[4] Videos 06-11-15, 06-06-07, 05-11-23, 05-10-12, 06-04-19, 06-02-08, 06-10-18, 06-11-29, 05-12-07 and 06-10-04.

Method	Audio only	Stack $K = Q$	Stack $K = Q + 1$	sNMcF
Mean score	21.4	25.1	18.9	16.8

Table 2. Mean results of the different methods (test set). Lower values indicate better performance.

3.4. Results

Table 2 reports the speaker diarization scores (using NIST reference scoring script) obtained using 1) NMF of the audio track only, 2) a naive method that factorizes the matrix $V_{stacked}$ form of the vertical concatenation of $\beta_1 V_{video}$ and $\beta_2 V_{audio}$, using either $K = Q$ or $K = Q + 1$, and 3) our soft nonnegative matrix co-factorization (sNMcF) method.

The results are as follows: it is better to factorize audio over $K + 1$ channels rather than restraining the video factorization to K channels if simply stacking audio and video features. Both stacked features and soft co-factorization yield better results than the analysis of only the audio track. Finally, soft co-factorization yields an improvement over other methods.

4. CONCLUSION

In this paper, we presented a new soft nonnegative matrix co-factorization (sNMcF) paradigm for heterogeneous but related modalities. Our paradigm relax the assumption that the modalities are explained by an *identical* common underlying factor, and gives full control over the expected correlation between the factorizations. We illustrated our approach on a real-world speaker diarization problem, and observed an improvement over reference methods. These results have been obtained with a limited use of prior knowledge, both to construct the cost function and initialize the proposed MM algorithm.

In future variants, we could easily modify the optimized cost function to embed additional structure such as smoothness. Initializations based on Support Vector Machines precomputed factorizations could spare oneself the trouble of knowing the number of speakers beforehand, and help the selection of best results. Hence nonnegative matrix soft co-factorization holds the potential for further improved results.

5. REFERENCES

[1] Félicien Vallet, Slim Essid, and Jean Carrive, "A multimodal approach to speaker diarization on TV talk-shows," *IEEE Transactions on Multimedia*, 2012.

[2] Athanasios Noulas, Gwenn Englebienne, and Ben Krose, "Multimodal Speaker Diarization," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 34, no. 1, pp. 79–93, 2011.

[3] Xavier Anguera Miro, Simo Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals, "Speaker Diarization: A Review of Recent Research," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 20, no. 2, pp. 356–370, 2012.

[4] Mathieu Ben, Michaël Betser, Frédéric Bimbot, and Guillaume Gravier, "Speaker Diarization using bottom-up clustering based on a Parameter-derived Distance between adapted GMMs," in *Proc. 8th International Conference on Spoken Language Processing (ICSLP)*, Jeju Island, Korea, 2004, number 2.

[5] Slim Essid and Cédric Févotte, "Smooth nonnegative matrix factorization for unsupervised audiovisual document structuring," *IEEE Transactions on Multimedia*, vol. 15, no. 2, pp. 415–425, Feb. 2012.

[6] Yusuf Kenan Yilmaz, Ali Taylan Cemgil, and Umut Simsekli, "Generalized Coupled Tensor Factorization," in *Proc. 15th Advances in Neural Information Processing Systems (NIPS)*, 2011.

[7] Evrim Acar, Tamara Kolda, and Daniel Dunlavy, "All-at-once Optimization for Coupled Matrix and Tensor Factorizations," *Computing Research Repository (CoRR)*, 2011.

[8] Naoto Yokoya, Takehisa Yairi, and Akira Iwasaki, "Coupled Nonnegative Matrix Factorization Unmixing for Hyperspectral and Multispectral Data Fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 2, pp. 528–537, 2012.

[9] Yanhua Chen, Lijun Wang, and Ming Dong, "Non-Negative Matrix Factorization for Semisupervised Heterogeneous Data Coclustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1459–1474, 2010.

[10] Yi Fang and Luo Si, "Matrix co-factorization for recommendation with rich side information and implicit feedback," in *Proc. 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec)*, New York, New York, USA, 2011, pp. 65–69, ACM Press.

[11] Jiho Yoo, Minje Kim, Kyeongok Kang, and Seungjin Choi, "Nonnegative matrix partial co-factorization for drum source separation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 2010, pp. 1942–1945, IEEE.

[12] Jiho Yoo and Seungjin Choi, "Matrix co-factorization on compressed sensing," in *Proc. 22nd International joint Conference on Artificial Intelligence (ICAI)*, Toby Walsh, Ed., Barcelona, Catalonia, Spain, 2011, pp. 1595–1602, AAAI Press.

[13] Rich Caruana, "Multitask Learning: A Knowledge-Based Source of Inductive Bias," in *Proc. 10th International Conference on Machine Learning (ICML)*. 1993, pp. 18–41, Morgan Kaufmann.

[14] Jonathon Shlens, "Notes on Kullback-Leibler Divergence and Likelihood Theory," 2007.

[15] Cédric Févotte and Ali Taylan Cemgil, "Nonnegative matrix factorisations as probabilistic inference in composite models," in *Proc. 17th European Signal Processing Conference (EUSIPCO)*, Glasgow, Scotland, 2009, pp. 1913–1917.

[16] David Hunter and Kenneth Lange, "A Tutorial on MM Algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[17] Cédric Févotte and Jérôme Idier, "Algorithms for Nonnegative Matrix Factorization with the β-Divergence," *Neural Computation*, vol. 2456, pp. 2421–2456, 2011.

[18] Alessandro Vinciarelli, Alfred Dielmann, Sarah Favre, and Hugues Salamin, "Canal9: A database of political debates for analysis of social interactions," in *IEEE International Workshop on Social Signal Processing*, Amsterdam, 2009, Ieee.

[19] NIST, "The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan," 2009.