IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

In-domain Adaptation Solutions for the RTVE 2018 Diarization Challenge

Ignacio Viñals, Pablo Gimeno, Alfonso Ortega, Antonio Miguel, Eduardo Lleida

ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Spain
{ivinalsb, pablogj, ortega, amiguel, lleida}@unizar.es

Abstract

This paper tries to deal with domain mismatch scenarios in the diarization task. This research has been carried out in the context of the Radio Televisión Española (RTVE) 2018 Challenge at IberSpeech 2018. This evaluation seeks the improvement of the diarization task in broadcast corpora, known to contain multiple unknown speakers. These speakers are set to contribute in different scenarios, genres, media and languages. The evaluation offers two different conditions: A closed one with restrictions in the resources to train and develop diarization systems, and an open condition without restrictions to check the latest improvements in the state-of-the-art.

Our proposal is centered on the closed condition, specially dealing with two important mismatches: media and language. ViVoLab system for the challenge is based on the i-vector PLDA framework: I-vectors are extracted from the input audio according to a given segmentation, supposing that each segment represents one speaker intervention. The diarization hypotheses are obtained by clustering the estimated i-vectors with a Fully Bayesian PLDA, a generative model with latent variables as speaker labels. The number of speakers is decided by comparing multiple hypotheses according to the Evidence Lower Bound (ELBO) provided by the PLDA, penalized in terms of the hypothesized speakers to compensate different modeling capabilities.

Index Terms: adaptation, diarization, broadcast, i-vector, PLDA, Variational Bayes

1. Introduction

The production of broadcast content has progressively augmented along the last years, becoming more and more necessary the tools to process and label all these new data. One of the required tasks is diarization, the indexation of some audio according to the active speaker. Hence the goal of diarization is the differentiation among the speakers by means of generic labels, leaving the identification of each speaker for further work, if necessary. Originally developed for telephone conversations, new domains such as broadcast audio and meetings are suitable to be interested in this technique, adding new challenging drawbacks not present in the original scenario.

Multiple approaches have been proposed to the diarization problem since its origins, most of them following two main strategies: The Top-Down philosophy, which obtains the correct labels by dividing an initial hypothesis with only one speaker, and the Bottom-Up strategy, which initially divides the input audio into acoustic segments containing only one speaker each, and combining them afterwards. Further information is available in [1][2]. Both philosophies need to characterize the speakers in the different parts of the audio to make any decision. For this reason diarization applies many methods developed for speaker recognition. Successful diarization systems considering these technologies are: Agglomerative Hierarchical Clustering (AHC)[3] with ΔBIC [4], streams of eigenvoices[5] resegmented with HMMs [6], i-vectors [7] clustered with PLDA [8], [9] in [10]. Neural Networks are also contributing, firstly providing more reliable acoustic information [11] and more recently a new representation: the embeddings such as xvectors [12].

When moving from telephone data to other scenarios, such as broadcast or meetings, new difficulties arise. Specially relevant are the estimation of the number of speakers and the domain mismatch. The first problem is caused by the presence of an unknown number of speakers in the audio. This difficulty increases if the contributions per speaker are significantly unbalanced. Our proposed solution to deal with this problem is [13], which makes use of a penalized version of the Evidence Lower Bound (ELBO) from a Variational Bayes solution, as reliability metric. Another important problem is the domain mismatch. Broadcast data consists of several shows, belonging to multiple genres and many differences in terms of locations, audio quality or postprocessing details. This large variability in the audio makes that one system is likely to lack in precision to cover the whole range of possibilities. However specific systems are also unfeasible for practical reasons. Our approach[14] combines both strategies: A single system is unsupervisedly adapted to the different shows to diarize with the same data to evaluate, obtaining the diarization labels afterwards.

The paper is organized as follows: Section 2 describes the evaluation and the available data. ViVoLab system is presented in Section 3. Section 4 is dedicated to present the obtained results. Finally, some conclusions are included in Section 5.

2. RTVE 2018 Challenge

The RTVE 2018 Challenge is part of the 2018 edition of the Albayzin [15], [16], [17], [18], [19] evaluations. These evaluations are designed to promote the evolution of speech technologies in Iberian languages. In particular, RTVE 2018 Challenge is focused on the extraction of relevant information from Broadcast data in Spanish language. This information, such as the identity of the person on screen and his speech, is intended to help describing and labeling the multimedia data for further work. To accomplish all these goals, the evaluation provides around 500 hours of shows from the Spanish Public TV corporation Radio Televisión Española (RTVE). The considered audio tries to cover the widest possible range of Spanish variability, including varieties of Spanish from Spain and Latin America. In addition to the provided audio some metadata is provided, with different levels of reliability.

The database is divided into 4 subsets, with different functionality:

This work has been supported by the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the 2015 FPI fellowship, the project TIN2017-85854-C4-1-R and Gobierno de Aragón /FEDER (research group T36_17R).

10.21437/IberSPEECH.2018-45

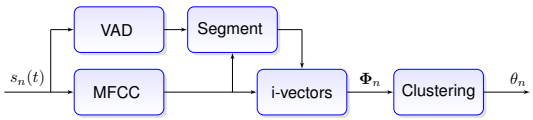

Figure 1: *Schematic of ViVoLab diarization system*

- **Train** 423 episodes from 16 shows. The included metadata consists of the broadcasted subtitles.
- **Dev1** 47 episodes from 5 shows, manually transcribed.
- **Dev2** 12 episodes from 2 shows, manually transcribed and with speaker time references.
- **Test** 61 episodes from 9 shows. Regarding the Diarization task, only 40 episodes from 4 shows are involved.

RTVE dataset is complemented with data collected for previous Albayzin evaluations. These data are:

- 3/24 TV channel [20]. Approximately 84 hours of Broadcast TV from 3/24 channel in Catalan language. Ground truth speaker marks are provided as metadata.
- CARTV. 20 Hours of radio broadcast, obtained from Corporación Aragonesa de Radio y Televisión broadcasted signal. The corresponding metadata consists of manually transcribed speaker time marks.

3. System description

ViVoLab submission is based on the diarization system firstly developed in [10]. This system, based on a bottom-up strategy, firstly divides the input audio into acoustic homogeneous segments, clustered afterwards by means of a Fully Bayesian PLDA. Its schematic is shown in Fig 1.

In the following lines we describe in details each part of the schematic.

3.1. Feature Extraction

From the input audio $s_n(t)$ 20 MFCC features vectors are extracted, including C0 (C0-C19), over a 25 ms hamming window every 10 ms (15 ms overlap). No derivatives are considered. The obtained features are normalized according to a Cepstral Mean and Variance Normalization (CMVN). The mean and variance are estimated taking into account the whole episode to diarize.

3.2. Voice Activity Detection

Voice Activity Detection (VAD) is performed by means of a 2-layer BLSTM [21] network of 128 neurons, trained on the 3/24 database. This network estimates the speech detection in 3-second duration sequences, providing one label each 10 ms of audio. The input parametrization is an 80-element Mel Filter bank and the log energy, computed in 25 ms windows with a 10 ms advance shift. The obtained features are normalized in mean and variance according to the whole audio.

3.3. Segment Representation

The speaker change detection is carried out making use of a ΔBIC [4] analysis, modeling the hypotheses with Full Covariance Gaussian distributions. A sliding window strategy has been considered for this purpose. Each one of the obtained acoustic segments is represented by an i-vector [7], with a 256-Gaussian 100-dimension i-vector extractor exclusively trained with 3/24 TV dataset. The obtained i-vectors have centering, whitening and length normalization [22] applied.

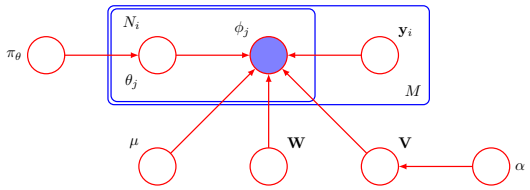

Figure 2: *Fully Bayesian PLDA*

3.4. Clustering Method

The i-vector clustering is performed by a 100-dimension Fully Bayesian PLDA [9] [10] trained with 3/24 and CARTV datasets. Its Bayesian Network is shown in Fig 2.

The high complexity of the model makes the Maximum Likelihood solution difficult to obtain. Therefore, the original paper suggests performing a Variational Bayes decomposition instead. This decomposition approximates the likelihood of the model by a product of factors q, each one depending on fewer hidden variables than the original distribution. In our case, the proposed decomposition is:

$$P(\mathbf{\Phi}, \mathbf{Y}, \theta, \pi_\theta, \boldsymbol{\mu}, \mathbf{V}, \mathbf{W}, \boldsymbol{\alpha}) \approx q(\mathbf{Y})\, q(\theta)\, q(\pi_\theta)\, q(\boldsymbol{\mu})\, q(\mathbf{V})\, q(\mathbf{W})\, q(\boldsymbol{\alpha}) \quad (1)$$

From all these factors, there is a set ($q(\boldsymbol{\mu})$, $q(\mathbf{V})$, $q(\mathbf{W})$, $q(\boldsymbol{\alpha})$) related to the Bayesian interpretation of the PLDA model parameters, while another set ($q(\theta)$,$q(\pi_\theta)$,$q(\mathbf{y})$), is related to the assignment of i-vectors in terms of soft speaker labels. The latter group is the base of our diarization system.

The whole clustering step consists of two differentiated steps. First we generate an initialization assignment for the i-vectors, becoming a seed for the corresponding hidden variable θ. Then all the related factors ($q(\theta)$,$q(\pi_\theta)$,$q(\mathbf{y})$) are iteratively reevaluated, reassigning the speaker labels in the process.

3.5. Unsupervised Adaptation

Due to the fact that known mismatches are present in the evaluation (Language regarding 3/24 dataset and media according to CARTV dataset), some loses in performance could be expected. We include an unsupervised adaptation [14] to the system. Prior to any evaluation, the out-of-domain PLDA is adapted to the evaluation domain considering the episode to diarize itself. Performed in the PLDA model, the adaptation requires speaker labels, which are obtained by unsupervised clustering. In this evaluation, the unsupervised labels are generated considering the diarization labels without adaptation.

3.6. Speaker number estimation

The considered PLDA and its Variational Bayes solution have the ability to recombine and eliminate speakers at will, only (and strongly) depending on the initial clustering. This initial clustering determines the maximum possible clusters to create as well as the first assignment of the i-vectors to these clusters. Our solution to this problem is the consideration of multiple initializations, with a different number of speakers. In order to

best compare the results from each hypothesis, we make use of a penalized version of the Evidence Lower Bound (ELBO) given by the PLDA model [13]. This metric, related to the likelihood, indicates how well the speaker labels represent the given data, but penalized in terms of the hypothesed number of speakers to avoid unnecessary subclustering, i.e., estimating more clusters than the real number of speakers.

4. Results

ViVoLab submission to the RTVE 2018 Diarization Challenge consists of two systems, a primary and a constrastive. Both follow the pipeline described previously, with the only difference that our primary system performs unsupervised adaptation while the contrastive does not. All the models were trained with 3/24 and CARTV data and no extra knowledge was considered, not even those provided by RTVE data.

The results obtained by the previously described configurations are exhibited in Table 1.

Table 1: *DER (%) Results for Primary and constrastive 1 systems in the development set. Results presented with Ground truth VAD and our BLSTM VAD for comparison reasons. Results include the DER term as well as its contributions: Miss speech (MISS), False Alarm speech (F.A.) and Speaker Error(SPK). Overlap is considered for evaluation purposes*

SYSTEM	MISS (%)	F.A. (%)	SPK(%)	DER(%)
ORACLE				
Primary	1.78	0.00	9.30	11.08
Contrastive	1.78	0.00	8.05	**9.83**
BLSTM				
Primary	2.51	1.74	6.69	**10.94**
Contrastive	2.51	1.74	9.99	14.24

An analysis of the obtained results in terms of the estimation of the number of speakers can be done. A simple analysis is available in Table2.

5. Conclusions

The ViVoLab submission to the RTVE 2018 Challenge includes two systems that have satisfactory results when considering the development set.

According to the development set, both systems are robust enough to deal with the subset mismatches, specially those known during the training phase: Language and Media. The trained models perform well while evaluating unknown audios from a new domain. Besides, when moving from ground truth VAD to a noisy one, a new sort of audio mismatch is introduced: i-vectors with non-speech. In this situation the unsupervised adaptation contributes learning from the evaluation data and adapting the model to the new conditions. This adaptation makes the system not to be degraded by the new scenario. In real conditions with noisy VAD the adaptive solution obtains a 24% relative improvement respect to the non-adaptive contrastive system.

Regarding the estimation of the number of speakers, the performed analysis indicates that our systems are likely to subcluster, i.e., hypothesize a larger number of clusters than the real value. If some of these extra clusters are dedicated to collect strange segments, the main clusters keep pure and clean and compensate any loss in performance. Therefore some subclustering could help avoiding relevant mistakes (primary system vs contrastive system using ground truth VAD). However, our results also indicate than an excesive subclustering causes a strong degradation of the performance (contrastive system with BLSTM VAD). Further research should be done to provide a deeper understanding.

Table 2: *Absolute and relative error in the estimation of the number of speakers for both the primary and contrastive systems. Error defined as $\#_{diar} - \#_{oracle}$. Ground truth (ORACLE) and BLSTM VAD conditions are studied. Analysis carried out on the development set.*

System	Absolute Error	Relative Error (%)
ORACLE		
Primary	12.91	97.87
Contrastive	1.83	16.40
BLSTM		
Primary	18.16	132.83
Contrastive	7.5	56.12

6. References

[1] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker Diarization: A Review of Recent Research," *IEEE Transactions On Audio Speech And Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[2] S. E. Tranter and D. Reynolds, "An Overview of Automatic Speaker Diarization Systems," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[3] D. Reynolds and P. Torres-Carrasquillo, "Approaches and Applications of Audio Diarization," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. V, pp. 953–956, 2005.

[4] S. Chen and P. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering Via the Bayesian Information Criterion," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, vol. 6, pp. 127–132, 1998.

[5] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," *Keynote presentation, Odyssey Speaker and Language Recognition Workshop*, 2010.

[6] C. Vaquero, *Robust diarization for speaker characterization*. PhD thesis, 2011.

[7] N. Dehak, P. 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] S. J. D. Prince and J. H. Elder, "Probabilistic Linear Discriminant Analysis for Inferences About Identity," in *Proceedings of the IEEE International Conference on Computer Vision*, 2007.

[9] J. Villalba and E. Lleida, "Unsupervised Adaptation of PLDA By Using Variational Bayes Methods," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 744–748, 2014.

[10] J. Villalba, A. Ortega, A. Miguel, and E. Lleida, "Variational Bayesian PLDA for Speaker Diarization in the MGB Challenge," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 667–674, 2015.

[11] Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang, and K. Yu, "Deep feature for text-dependent speaker verification," *Speech Communication*, vol. 73, pp. 1–13, 2015.

[12] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep Neural Network-based Speaker Embeddings for End-to-end Speaker Verification," *iEEE*

Spoken Language Technology Workshop (SLT), pp. 165–170, 2016.

[13] I. Viñals, P. Gimeno, A. Ortega, A. Miguel, and E. Lleida, "Estimation of the Number of Speakers with Variational Bayesian PLDA in the DIHARD Diarization Challenge," *Interspeech*, no. September, pp. 2803–2807, 2018.

[14] I. Viñals, A. Ortega, J. Villalba, A. Miguel, and E. Lleida, "Domain Adaptation of PLDA models in Broadcast Diarization by means of Unsupervised Speaker Clustering," *Interspeech*, pp. 2829–2833, 2017.

[15] A. Ortega, D. Castan, A. Miguel, and E. Lleida, "The Albayzin 2012 Audio Segmentation Evaluation," 2012.

[16] J. Tejedor and D. T. Toledano, "The ALBAYZIN 2014 Search on Speech Evaluation," no. November, 2014.

[17] A. Ortega, D. Castan, A. Miguel, and E. Lleida, "The Albayzin 2014 Audio Segmentation Evaluation," 2014.

[18] J. Tejedor and D. T. Toledano, "The ALBAYZIN 2016 Search on Speech Evaluation," 2016.

[19] A. Ortega, I. Viñals, A. Miguel, and E. Lleida, "The Albayzin 2016 Speaker Diarization Evaluation," 2016.

[20] M. Zelenák, H. Schulz, and J. Hernando, "Speaker diarization of broadcast news in Albayzin 2010 evaluation campaign," *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 2012, no. 1, pp. 1–9, 2012.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1–32, 1997.

[22] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 249–252, 2011.