IberSPEECH 2024
11-13 November 2024, Aveiro, Portugal

AUDIAS-UAM System Description for the Albayzin-RTVE 2024 Speaker Diarization Challenge

Alicia Lozano-Diez[1], *Juan Ignacio Alvarez-Trejos*[1], *Laura Herrera*[1], *Beltran Labrador*[1], *Jeremie Touati*[2], *Sara Barahona*[1]

[1] AUDIAS Universidad Autónoma de Madrid, Spain
[2] École Polytechnique, France

alicia.lozano@uam.es, juani.alvarez@uam.es, laura.herrera@estudiante.uam.es

Abstract

In this paper, we describe the speaker diarization system submitted by the AUDIAS-UAM team for the Albayzin-RTVE 2024 Speaker Diarization Challenge. Our primary submission consists of the combination via DOVER-Lap of three speaker diarization systems within the state-of-the-art: Pyannote, VBx and DiaPer. Both Pyannote and DiaPer systems are based on neural networks for diarization of shorter segments, followed by a matching algorithm to assigned predicted speaker labels for the whole recording. VBx is used to obtained speaker diarization labels over the whole recordings. The combination of these individual systems yields a 9.26% DER on our development set, with respect to a 12.26% DER of Pyannote, 15.30% DER of VBx and 23.25% DER of DiaPer, showing the potential of a fusion of three quite distinct diarization systems.

Index Terms: Speaker Diarization, DiaPer, Pyannote, VBx, DOVER-Lap

1. Introduction

In this paper, we describe the systems developed by the AUDIAS-UAM team for the speaker diarization task within the Albayzin-RTVE 2024 challenge.

The speaker diarization task consists of automatically segmenting an audio with respect to the speaker turns, and grouping segments of the same speaker, without specifying speaker identities [1].

Traditionally, speaker diarization systems have been built with a modular approach, in which a voice activity detector (VAD) is applied to select the speech frames. After that, speaker embeddings are extracted and clustered to obtained speaker diarization outputs. Although each of these modules are optimized separately, these approaches keep showing state-of-the-art results, especially in wide-band data. A successful example is the system known as VBx [1], based on variational Bayes training of HMMs to cluster speaker embeddings (x-vectors). This system is one of the systems that compose our primary submission for the challenge.

More recently, the neural network based models known as End-to-End Neural Diarization (EEND) have shown their potential and outperformed modular systems, especially when dealing with telephone speech and two speaker conversations [2]. These models are typically based on Transformers [3] and perform speaker diarization as a per-frame multi-label classification problem. One of the most successful approaches has been the EEND with Encoder-Decoder Attractors (EEND-EDA) [4] and its variant based on powerset loss [5]. However, these approaches based on self-attention require computational resources that make them impractical for long audio recordings such as the ones in RTVE datasets. The Pyannote [6] implementation applies EEND models on short segments and clusters labels afterwards, which is the approach of another system used for our submission to the challenge.

Some alternatives to EEND-EDA have been explored in order to overcome this limitation of long audio recordings, as well as to deal with larger number of speakers. In this line, the DiaPer model [7] is based on cross-attention to convert variable-sized input into a fixed-size representation, reducing complexity to linear. This model has shown as well good performance in wide-band domains with larger number of speakers, and therefore, we have explored it as part of our system for the challenge [8]. We found that it was still unable to perform well and in a reasonable time for segments longer than 5 minutes, so we have explored a matching algorithm to combine the outputs obtained for these shorter segments [9]. This algorithm could be used with any diarization model, and we have applied it to the DiaPer model.

Finally, our submission is a combination of three speaker diarization models via DOVER-Lap [10], as detailed in the rest of the paper and shown in Figure 1.

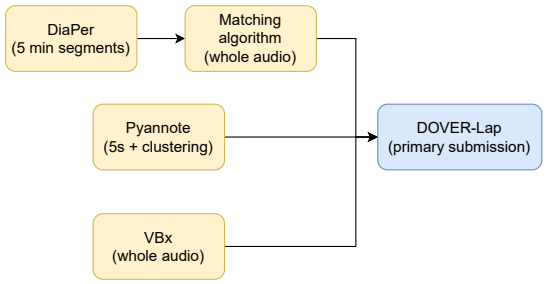

Figure 1: *Diagram of the primary submission of the AUDIAS-UAM team for the Albayzin-RTVE 2024 speaker diarization challenge.*

2. Individual and Combined Systems

In this section, we describe the individual systems developed for the AUDIAS-UAM team for the participation on the Albayzin-RTVE 2024 Speaker Diarization Challenge, as well as the combination of these systems, which constitutes our primary submission.

[1]Note that we do not tackle the speaker attribution part of the evaluation, in which some target speakers are identified within that segmentation.

10.21437/IberSPEECH.2024-69

2.1. VBx

One of the individual system used for this submission is the VBx (Variational Bayes HMM Clustering of x-vector Sequences) model [1], based on the cascaded speaker diarization pipeline. This system clusters the speaker embeddings (x-vectors) using variational Bayes HMMs and Probabilistic Linear Discriminant Analysis (PLDA) for distance estimation between embeddings. In particular, we used the pretrained model from the authors' repository [2], which shows state-of-the-art performance in several datasets.

2.2. Pyannote

As another individual system for our submission, we explored the speaker diarization system released within the open-source Python toolbox Pyannote [6]. In particular, we used Pyannote.audio 3.1 [5] from the authors' repository [3], which performs speaker diarization as follows. First, local speaker segmentation is performed over short segments of 5 seconds, with overlapping windows, using neural speaker embedding based on EEND state-of-the-art techniques. Then, a global agglomerative clustering is performed to match speaker label outputs from the EEND model [11].

2.3. DiaPer

Finally, another individual systems of our primary submission is based on the DiaPer [4] (Perceiver-based Attractor for End-to-End Neural Diarization) model [7]. This model builds on the idea of End-to-End Neural Diarization with Encoder-Decoder Attractors (EEND-EDA). Therefore, it takes as input acoustic features extracted from the audio, which are then embedded through a frame-encoder based on self-attention layers. However, instead of using the EDA based on LSTMs, DiaPer replaces it with a Perceiver-based attractor utilizing iterative cross-attention layers. The attractors are initialized randomly, and are latent representations of the speakers. As output of the system, the per-speaker per-frame activities are determined calculating a cross-product between the attractors and the frame embeddings obtained in the encoder part. This system is explored instead of the well-known EEND-EDA due to its low computational cost, which allows us to use it for longer input sequences. Further exploration of this model for the Albayzin/RTVE2022 database can be found in [8].

2.4. Matching Algorithm

As mentioned before, DiaPer system can deal with longer sequences than other EEND approaches with a reasonable computational cost. However, there are still constraints when dealing with long audio recordings as the ones of the Albayzin-RTVE challenges. Therefore, we apply a matching algorithm to stick together local speaker diarization outputs of few minutes segments to obtain diarization labels for the full recording.

The matching algorithm used on top of DiaPer outputs for the primary submission (and on top of VBx for a contrastive submission) starts by segmenting the audio into equal-sized chunks, which are then processed individually using DiaPer model. This approach generates a prediction matrix per chunk, with the per frame per speaker activations.

With this information, speaker embeddings are computed for each chunk by concatenating the predicted frames associated with the same speaker within the chunk and then extracting the embedding using the ECAPA-TDNN model. We incorporate Voice Activity Detection (VAD) to filter out irrelevant noise and ensuring that only valid speech segments are considered for embedding extraction. The original labels obtained from the audio chunks are not altered during the matching process.

Using the embeddings extracted, a distance matrix is created by performing pairwise cosine similarity, and including a constrained linkage to maximize the distance between embeddings corresponding to different speakers within the same chunk, ensuring that the embeddings of distinct speakers in the same segment are well-separated in the full permuted output.

Finally, agglomerative hierarchical clustering is performed using the affinity matrix, and a post-processing step is performed to permute the predictions into a consolidated output providing a coherent diarization result for the whole audio.

Further exploration of this model for the Albayzin/RTVE2022 database can be found in [9].

2.5. Fusion with DOVER-Lap

Our primary submission consists of the fusion of the three speaker diarization systems mentioned before: Pyannote, VBx and DiaPer+Matching.

The combination of these systems has been performed using the well-known DOVER-Lap [10] algorithm to fuse outputs of different diarization systems taking into account overlapping segments. The implementation of the algorithm used was the one from the authors' repository [5]. Equal weights were applied to the three systems due to exploration of results on our development dataset.

3. Datasets and Metrics

3.1. Pretrained Models

The three main models used to develop our submitted system are based on pretrained models on different datasets and then adjusted to the RTVE setup.

The DiaPer model was initially trained on simulated conversations from LibriSpeech [12]. The simulated conversations were created following the procedure described in [13], with the statistics of our development set created from RTVE data describe in Section 3.2. This model was trained from scratch on simulated conversations of 2 speakers and fine-tuned on simulated conversations of up to 10 speakers, with approximately 10k hours of speech. Finally, we adapted it to real conversations from RTVE datasets, with training segments of 5 minutes.

The VBx model used was the one available in the repository mentioned in Section 2.1. This model uses an x-vector extractor and a PLDA model trained on VoxCeleb [14, 15]. This model was not adapted to any RTVE data.

The Pyannote model was pretrained on a wide range of datasets, including Albayzin/RTVE2022 [16], AliMeeting channel 1 [17], AMI array 1 channel 1 [18], CALLHOME part 2 [19], Ego4d v1 validation [20], and This American Life [21]. The model used was the one available in the repository mentioned in Section 2.2.

[2] https://github.com/BUTSpeechFIT/VBx

[3] https://github.com/inferless/pyannote-speaker-diarization-3.1

[4] https://github.com/BUTSpeechFIT/DiaPer

[5] https://github.com/desh2608/dover-lap

3.2. RTVE Datasets

In order to adapt the DiaPer model, as well as obtaining the statistics to generate simulated conversations and tuning the matching algorithm and hyperparameters, we use the dataset from the Albayzin-RTVE2022 diarization challenge. We consider the training set for our fine-tuning and the development set for showing the development results. We also remove from our test set a few recordings that were present in both sets when released for previous evaluations. Therefore, our setup contains approximately 42 hours of audio for training and 33 hours for testing.

Regarding the labels, we cleaned the data by removing music or screams labels, and corrected labels that referred to the same speaker with different notation (example: #1 and 1 or #5 and # 5 in different segments of the same audio).

3.3. Error Metrics

In order to compare our different systems developed, we evaluate the Diarization Error Rate (DER), since this is the metric used as well officially in the challenge. We use a collar of 0.25 seconds for all the results, and the tool provided by the organizers of the challenge, as mentioned in the evaluation plan [22].

4. Results

4.1. Development Results

Results obtained for individual systems as well as the fusion corresponding to our primary submission on our development set are summarized in Table 1.

Table 1: *Results of final systems over our development set corresponding to the test set of Albayzin-RTVE2022 challenge. Results are shown in %. Our primary submission corresponds to the system refer to as DOVER-Lap.*

System	Miss	FA	SpkErr	DER
Pyannote	4.3	3.0	5.0	**12.26**
VBx	10.1	0.0	5.2	15.30
DiaPer+Matching	7.2	2.9	13.2	23.25
DOVER-Lap	3.2	0.9	5.2	**9.26**

As seen in Table 1, the Pyannote system obtains the best results in terms of DER. VBx is not able to outperform Pyannote regarding DER. Although VBx reduces FA errors to 0, there is an increase in speaker errors and miss speech. DiaPer + Matching Algorithm is the system that obtains the worst results in terms of DER, nevertheless, the outcomes are quite satisfactory for an EEND model in this setup. Despite this, the fusion of systems allows obtaining more satisfactory result, obtaining an almost 25% relative improvement in DER with respect to Pyannote.

4.2. Official Test Results

The official test results for our primary submission are presented in Table 2. These results correspond to the test set of the Albayzin-RTVE2024 diarization challenge.

As seen in Table 2, our primary system submission achieved a DER of 24.64% on the official test set. While the miss rate and speaker error are notably higher than in the development results, the false alarm rate remains low, indicating robustness in speech detection accuracy. The overall DER reflects a strong performance under the challenge's real-world conditions, though further reduction in miss rate and speaker error could enhance system reliability. This result underscores the potential of our fusion approach, particularly for its balance of false alarms and speaker errors.

Table 2: *Results of the official test set of the Albayzin-RTVE2024 diarization challenge for our primary system submission. Results are shown in %.*

System	Miss	FA	SpkErr	DER
Primary	16.75	1.43	6.44	24.64

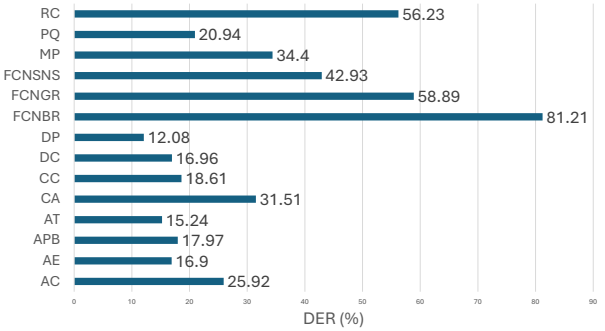

Figure 2: *DER performance across different program types in the RTVE 2024 test set, illustrating variability in system performance based on program characteristics.*

Figure 2 presents the DER across different program types within the RTVE 2024 test set. Notably, there are substantial variations in DER depending on the program type, which suggests that the system's performance is highly dependent on specific program characteristics. Such differences highlight the potential impact of factors like background noise, speaker overlap, and recording conditions, which vary significantly across program genres and contribute to the variability in system accuracy.

5. Computation Resources

For the individual modules Pyannote, Diaper, and the matching algorithm, it is convenient to use a GPU for processing. We utilized a GeForce RTX 2080 Ti graphics card. DOVER-Lap run on CPU since it only works on the output RTTMs of individual systems. The execution times for each individual system and the fusion via DOVER-Lap are shown in Table 3.

Table 3: *Execution times over the entire RTVE2024 test set for each individual system. Results are shown in seconds.*

System	Execution Time (s)
Pyannote	10786.5
VBx	10773
Diaper	110.5
+Matching	575.64
DOVER-lap	7.87

6. Conclusions

In this paper, we described the AUDIAS-UAM submission for the speaker diarization task of the Albayzin-RTVE 2024 challenge. Our final system is the fusion via DOVER-Lap of three individual speaker diarization systems, which have shown state-of-the-art performance in several well-known frameworks. The individual systems are based on different models: Pyannote, based on EEND on very short segments followed by clustering; VBx, which builds on Bayesian HMMs on top of x-vectors; and DiaPer, which uses cross-attention and an end-to-end approach applied on 5-minute segments followed by a matching algorithm. As individual systems, they showed quite different performance in terms of DER over our development set (test set of RTVE 2022), being the Pyannote model the best performing system, closely followed by VBx. However, the complementary information provided by the different approaches is reflected on the approximately 25% relative improvement obtained by simply using the DOVER-Lap as combination method.

7. Acknowledgments

This work was supported by PID2021-125943OB-I00, MCIN/AEI/10.13039/501100011033/FEDER, UE from the Spanish Ministerio de Ciencia e Innovacion, Agencia y del Fondo Europeo de Desarrollo Regional.

8. References

[1] F. Landini, J. Profant, M. Diez, and L. Burget, "Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: Theory, implementation and analysis on standard tasks," *Computer Speech Language*, vol. 71, p. 101254, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230821000619

[2] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Computer Speech Language*, vol. 72, p. 101317, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230821001121

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[4] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and K. Nagamatsu, "End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors," in *Interspeech*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:218719457

[5] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in *Proc. INTERSPEECH 2023*, 2023.

[6] H. Bredin, "Pyannote.audio 2.1 speaker diarization pipeline: Principle, benchmark, and recipe," in *Interspeech*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260906702

[7] F. Landini, M. Diez, T. Stafylakis, and L. Burget, "Diaper: End-to-end neural diarization with perceiver-based attractors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–16, 2024.

[8] J. Touati, J. I. Alvarez-Trejos, B. Labrador, and A. Lozano-Diez, "Efficient transformers for end-to-end neural speaker diarization," in *Proc. Iberspeech (to appear)*, 11 2024.

[9] J. I. Alvarez-Trejos, L. Herrera, J. Touati, and A. Lozano-Diez, "Analysis of speaker label matching for diarization of long audios on rtve2022 dataset," in *Proc. Iberspeech (to appear)*, 11 2024.

[10] D. Raj, L. Paola Garcia-Perera, Z. Huang, S. Watanabe, D. Povey, A. Stolcke, and S. Khudanpur, "Dover-lap: A method for combining overlap-aware diarization outputs," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 881–888.

[11] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Interspeech*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233204561

[12] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[13] F. Landini, A. Lozano-Diez, M. Diez, and L. Burget, "From Simulated Mixtures to Simulated Conversations as Training Data for End-to-End Neural Diarization," in *Proc. Interspeech 2022*, 2022, pp. 5095–5099.

[14] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, 2017, pp. 2616–2620.

[15] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018.*, 2018, pp. 1086–1090.

[16] A. Ortega, A. Miguel, E. Lleida, V. Bazán, C. Pérez, and A. de Prada, "Albayzin Evaluation IberSPEECH-RTVE 2022 Speaker Diarization and Identity Assignment," http://catedrartve.unizar.es/reto2022/SDIAC2022_Evalplan.pdf, 2022.

[17] F. Yu, S. Zhang, Y. Fu, L. Xie, S. Zheng, Z. Du, W. Huang, P. Guo, Z. Yan, B. Ma, X. Xu, and H. Bu, "M2MeT: The ICASSP 2022 Multi-Channel Multi-Party Meeting Transcription Challenge," in *Proc. ICASSP 2022*, 2022.

[18] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, and M. Kronenthal, "The AMI Meetings Corpus," in *Proc. Symposium on Annotating and Measuring Meeting Behavior*, 2005.

[19] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream-Based Speaker Segmentation using Speaker Factors and Eigenvoices," in *Proc. ICASSP 2008*, 2008.

[20] K. Grauman, A. Westbury, and E. B. et al., "Ego4D: Around the World in 3,000 Hours of Egocentric Video," in *Proc. CVPR 2022*, 2022.

[21] H. H. Mao, S. Li, J. J. McAuley, and G. W. Cottrell, "Speech Recognition and Multi-Speaker Diarization of Long Conversations," in *Proc. Interspeech 2020*, 2020.

[22] A. Ortega, A. Miguel, E. Lleida, V. Bazán, C. Pérez, and P. Vila, "Albayzin evaluation iberspeech-rtve 2024 speaker diarization and identity assignment," in *Online*. Online, 2024. [Online]. Available: https://catedrartve.unizar.es/reto2024/SDIAC2024_Evalplan.pdf