Intelligent Voice Speaker Recognition and Diarization System for IberSpeech 2022 Albayzin Evaluations Speaker Diarization and Identity Assignment Challenge

Roman Shrestha[1], *Cornelius Glackin*[1] *Julie Wall*[2],*Mansour Moniri*[2], *Nigel Cannings*[1]

[1]Intelligent Voice Limited, London, UK
[2]University of East London, London, UK

{roman.shrestha, neil.glackin, nigel.cannings}@intelligentvoice.com, {j.wall, m.moniri}@uel.ac.uk

Abstract

This paper describes the system developed by Intelligent Voice for the IberSpeech 2022 Albayzin Evaluations Speaker Diarization and Identity Assignment Challenge. The presented Variational Bayes x-vector Voice Print Extraction system is capable of capturing vocal variations using multiple x-vector representations with two-stage clustering and outlier detection refinement and implements the Deep-Encoder Convolutional Autoencoder Denoiser network for denoising segments with noise and music on files identified by a signal to noise ratio classifier for robust speaker recognition and diarization. When evaluated against the Radiotelevision Espanola 2022 evaluation dataset, the system was able to obtain a diarization error rate of 35.59% for the Speaker Diarization task and assignment error rate of 28.88% for the Identity Assignment task.

Index Terms: Speaker Recognition, Diarization, Identity Assignment, Speech Enhancement

1. Introduction

The human voice constitutes of a multitude of acoustic features that can provide vital cues on a person's identity. The significance of systems that can recognise speakers from intrinsic audio recordings extends beyond the commercial importance of diarization for speech technology and downstream NLP tasks [1]. Hence, text-independent speaker diarization, recognition and verification research is gaining a lot of interest lately from active community of researchers and academics globally, which has nurtured several ground-breaking architectures to effectively address the "Who spoke when?" problem. However, recognizing speakers solely based on their acoustic features is still considered as an esoteric challenge due to the inability of the existing systems to cope with noise, overlapping speech and the acoustic variations in speech which can be easily influenced by environmental, emotional and linguistic factors.

Speaker Diarization and Identity Assignment is the task of segmenting audio segments within a conversation based on their utterances and associating those segments with their respective identities [2]. This process typically involves several stages, namely Voice Activity Detection (VAD), segmentation of the identified speech segments into shorter segments, extraction of the speaker's acoustic features using either i-vectors [3], d-vectors [4], or x-vectors [5], and clustering the segments using techniques such as k-Means [6] or Agglomerative Hierarchial Clustering (AHC) [7] to obtain accurate speaker separation from a multi-speaker recording. Research employing speech enhancement to cancel out noise, reverberation and normalize distortion from the noisy audio signals have also shown improvement in this domain [8]. The accessibility to real-world evaluation corpora such as Radiotelevision Espanola (RTVE) 2022 [9], DIHARD-2[10], DIHARD-3 [11], CALLHOME [12], AMI [13], VoxCeleb [14], MultiSV [15], HI-MIA [16] and CHiME-6 [17] have exposed the complexity of the task for real-world conversational scenarios.

Early speaker recognition systems were based on the Gaussian Mixture Model (GMM)-Universal Background Model (UBM) approach, where the GMM of individual speakers were adapted from the UBM trained on a large amount of unlabelled data to represent the acoustic feature distribution of speech, and the likelihood ratio of the test features was computed to identify the speakers present in a recording [18]. A few years later, Kenny et. al [19] proposed the Joint Factor Analysis (JFA) approach to improve GMM estimation by allowing the modelling of interspeaker variability and compensation for channel/session variability in the context of high-dimensional GMM supervectors [19].

With the advent of i-vectors [3], unique fixed length embeddings extracted from the recordings could be directly used for identifying speakers based on their voices using cosine similarity scoring [20]. Linear Discriminant Analysis (LDA) and Nuisance Attribute Projection (NAP) techniques were introduced to cope with unwanted variations that affected i-vectors due to a mismatch of linguistic content and recording channel information between segments of speech spoken by the same speaker, demonstrating an improved performance [20]. Probabilistic LDA (PLDA), originally introduced by Price and Gee for facial recognition [21], has emerged as a powerful tool for speaker verification capable of generating well-calibrated likelihood ratios between the vectors [22]. Kenny [23] was amongst the pioneers for implementing PLDA in the i-vector space for modelling channel variability [23].

Deep Neural Network (DNN) based feature vector extractors have shown an improved performance for speaker recognition and diarization tasks compared to the earlier systems [24, 25, 26, 27, 28]. For many years, DNN-based i-vector systems implementing PLDA scoring were regarded as the state of the art in the speaker verification domain. Recently, x-vector based systems which operate by extracting x-vectors from speech segments, performing LDA and using PLDA classifiers to perform a likelihood ratio test between the speakers have observed superior performance on speaker recognition and diarization across different acoustic channels [5, 29, 30, 22, 31].

Also, systems implementing the Weighted Prediction Error (WPE) speech dereverberation algorithm to cancel out reverberation and background noise[8] for generating clean audio signals and better speaker embeddings have demonstrated im-

pressive performance in the speaker verification domain where the waveform amplitude distribution analysis method was employed to estimate the Signal to Noise Ratio (SNR) of the real speech recordings, whereby degraded and noisy audio signals were processed by the Virtual Acoustic Channel Expansion (VACE)-WPE and speaker embeddings were extracted using a pre-trained Resnet-34 Deep Speaker Embedding (DSE) model employing dereverberation without Task specific Optimization (TSO), characterized by prefix Drv [8].

This paper extends our Variational Bayes x-vector Voice Print Extraction (VBxVPE) [1] research by implementing the Deep-Encoder Convolutional Autoencoder Denoiser (DE-CADE) speech enhancement model [32] to cancel out noise, music and reverberation from the files identified by a SNR classifier for effective speaker diarization and identity assignment. The VBxVPE system is capable of capturing an individual speaker's speech variability resulting from different speaking styles and varying vocal effort using multiple x-vector representations associated with a speaker. The novelty of our work lies in the core-extraction procedure where we refine the x-vectors by implementing a robust outlier detector followed by re-clustering of the vectors using the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm [33], to obtain refined clusters from which the centre of the refined clusters are extracted as the cores representing the different acoustic variations in speech of the speaker of interest. The core representations are then stored in a vector database, which supports semantic vector search using cosine similarity to identify the closest match between the enrolled and test speaker cores.

The rest of the paper is organised into three distinct sections. Section 2 provides the methodology and description of the implemented system; information on the benchmark dataset, reported evaluation metrics and discussions are provided in Section 3, followed by the conclusions in Section 4.

2. System Description

2.1. Data Preprocessing

The audio files were provided in the Advanced Audio Coding (AAC) file format sampled at 44100 Hz with two channels, by default for both enrolment and inference. These were downsampled to 8 KHz, a standard sample rate for recording the human voice, and converted to a mono channel WAV file format using the Fast Forward Motion Picture Experts Group (FFMPEG) Python library [34]. The WAV files were then used for either enrolment (Section 2.7) or evaluation (Section 2.8) as per the RTVE 2022 [9] evaluation set specifications [2].

2.2. SNR classification

A Spleeter [35] based SNR classifier was implemented to detect audio files containing excessive noise and music. Spleeter [35] is a powerful tool that can separate vocal and music accompaniment in music audio. Using functionalities from the numpy [36] Python library, simple smoothing was applied to the accompaniment audio signals after normalization in the range [0,1]. Then the average of the energies was calculated from normalized accompaniment audio signals to obtain a SNR between 0 and 1. Based on our experimental observations through trial and error, audio files containing a SNR value greater than 0.2 obtained from the accompaniment audio signals benefitted from speech enhancement performed by DE-CADE [32].

2.3. Speech Enhancement

DE-CADE is a two-stage DNN architecture for speech enhancement that outperforms the current state of the art systems in this domain [32], was used to generate clean audio signals from the noisy and distorted audio recordings provided in the evaluation set. The implemented deep convolutional denoising autoencoder-based speech enhancement network [32] operates by performing denoising first in the frequency domain stage using the magnitude spectrum as a training target followed by denoising and speech reconstruction in the temporal domain in the second stage.

Speech enhancement was performed for all 38 files from the evaluation set and VAD was always performed on the evaluation files processed with DE-CADE. However, the SNR classifier described in Section 2.2 was used to determine whether to use files processed by the DE-CADE speech enhancement algorithm or not as input audio for the rest of the inferencing pipeline.

2.4. VAD and X-Vector Extraction

An energy based VAD system operates on the audio files processed by DE-CADE to get rid of non-speech segments within the audio that might lead to noisy x-vectors. 256 dimensional x-vectors were extracted from the segments specified by VAD using a pre-trained ResNet-101 8 KHz network [31]. The extracted x-vectors were reduced to 128 dimensions using LDA dimensionality reduction for further processing.

2.5. Speaker Diarization

VBx diarization [31] was chosen as the reference architecture for speaker diarization due to its superior performance on three of the most popular datasets for evaluating diarization: the CALLHOME [12], AMI [13] and DIHARD-2 datasets [31]. The AHC algorithm [7] used by the VBx diarization system [31] was replaced by a greedy clustering algorithm that operates by calculating the cosine similarity between a vector and every other x-vector that appears on the sequence after the reference x-vector. The algorithm scans for the drop in similarity below the threshold of 60% which was defined based on our experimental observations between the vectors and forms a mini cluster and then starts clustering again with the next x-vector in the sequence as a reference vector. Once all the x-vector clusters are obtained, similar clusters are merged based on the similarity between the reference x-vectors. The implemented greedy algorithm runs 1.8 times faster than AHC and improves the Diarization Error Rate (DER) by 0.91% [31] when evaluated against the evaluation set of the DIHARD-3 Challenge [11]. Then, a PLDA model pre-trained on a large number of speaker-labeled x-vectors [31] scores the obtained clusters to verify the likelihood ratio between them [22], thereby preparing the final diarization output detailing who spoke when in the audio file.

2.6. Core Extraction

Core Extraction also known as Voice Print Extraction can be regarded as the process of generating a distinct vocal signature from the acoustic features present in a person's speech. For every speaker recognized, the core extraction is performed in two stages, Outlier Detection and then HDBSCAN Clustering[33].

Initially, all the x-vectors representing a speaker are grouped together and investigated for outlier detection where the system calculates a cosine similarity matrix between all the

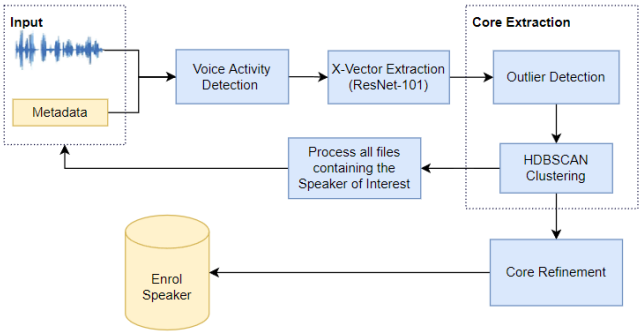

Figure 1: *Enrolment Pipeline*

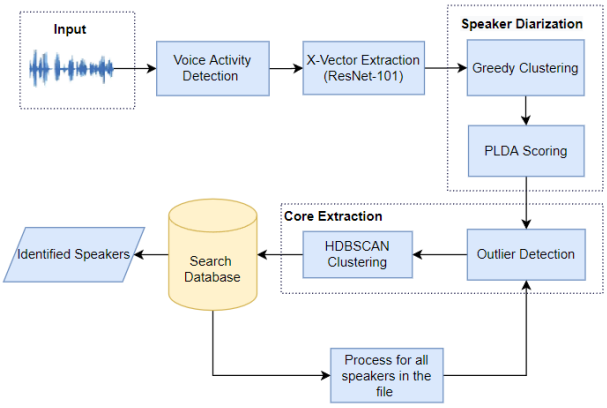

Figure 2: *Evaluation Pipeline*

x-vectors and eliminates any noisy x-vectors. Noisy x-vectors are identified based on the cosine similarity measure and the vectors that cannot demonstrate a strong association with any of the major clusters are discarded. The remaining vectors are then processed with HDBSCAN clustering with an aggressive setting by enabling the 'allow_single_cluster' parameter [33] i.e. the x-vectors are re-clustered. This will yield a minimum of one cluster. The number of clusters indicates the distinct speaking styles captured from a speaker's vocal features, enabling the system to capture and identify the speaker of interest across a variety of domains. Finally, the centres of the obtained clusters (simple centroid calculation) are extracted and stored as the voice print of the speaker.

2.7. The Enrolment Pipeline

Based on the audio files provided for enrolment by the RTVE2022 dataset [9], 74 speakers were enrolled from audio files containing speech from the speaker of interest (∼30 seconds per speaker). The total enrolment time was reported as 3 minutes and 13 seconds.

Fig. 1 shows the enrolment pipeline for enrolling the speakers for performing speaker recognition with the proposed VBxVPE system. The enrolment procedure commences by accepting the audio and label for the speaker of interest as metadata.

Since the enrolment files only contained clear speech segments from the speaker of interest, diarization and speech enhancement was not performed in the enrolment pipeline, only in the evaluation pipeline. After VAD and x-vector extraction is performed as described in Section 2.4, the voice print for the speaker of interest is extracted from the file containing the speaker based on the metadata provided as described in Section 2.6.

After extracting the cores as described in section 2.6, core refinement is performed by comparing the obtained voice prints against each other using cosine similarity and discarding the cores with similarity greater than 85%, which is the ideal threshold determined by trial and error to prevent duplication. All unique voice prints derived from the acoustic features constituting the speaker's voice are then enrolled in the vector search database along with a unique speaker id.

2.8. The Evaluation Pipeline

For evaluation of the speaker identity recognition system, the evaluation set of the RTVE 2022 dataset [9] contained 25 hours of audio with overlapping speech segments, background music and frequent speaker changes. A SNR classifier described in Section 2.2, was implemented to determine the levels of noise and music contained in the audio files. Based on a SNR threshold of 0.2 decided by trial-and-error, the audio files with a SNR value of greater than 0.2 were processed by the DE-CADE speech enhancement model [32] to further clean the noisy audio signals.

Fig. 2 presents the evaluation pipeline for the VBxVPE system. The developed speaker recognition system operates on segments of audio recordings containing speech identified by VAD. After extracting the x-vectors from the audio segments containing speech, speaker diarization facilitates the grouping of x-vectors associated with the speakers identified.

For every speaker identified by diarization, the core/s is extracted from the pool of x-vectors associated with the speaker as explained in Section 2.6. The cores are then searched across the vector search database which is capable of performing semantic vector search based on cosine similarity. Speakers are identified if there is a match greater than the identification threshold of 80%, determined through trial-and-error experiments, with any enrolled core representing the speaker in the vector search database. The final output was provided in the Rich Transcription Time Marked (RTTM) file format.

3. Results and Discussions

The total processing time for all the experiments were identical and took approximately 5 hours to process the 38 files containing 25 hours of speech comprising the RTVE 2022 evaluation set [9]. X-vector extraction was performed on an NVIDIA GeForce GTX 1080 Ti GPU whereas the rest of the processes were executed on a single core 64-Bit CPU with Intel Xeon 2.20GHz processor and 128GB RAM.

The Albayzin Evaluation Speaker Diarization and Identity Assignment Challenge (SDIAC) is comprised of two sub tasks: Speaker Diarization and Identity Assignment. The Speaker Diarization task was compulsory and required the participants to separate and group the recordings based on an unknown identity whereas the optional Identity Assignment task required the participants to retrieve the speaker's identity and assign names to the diarization labels [2]. The evaluation metrics were reported in terms of DER for the Speaker Diarization task and Assignment Error Rate (AER) which is a slightly modified version of DER for the Identity Assignment task [2].

A total of 11 systems were submitted for the challenge

Table 1: *System Results*

System	threshold	DER (Speaker Diarization)	AER (Identity Assignment)
SDIAC_TEAMIV_p–c1	75%	45.92%	1191.64%
SDIAC_TEAMIV–l1	75%	38.91%	153.36%
SDIAC_TEAMIV–l3	75%	40.74%	185.42%
SDIAC_TEAMIV–l4	80%	35.82%	36.07%
SDIAC_TEAMIV–l8	80%	37.20%	44.34%
SDIAC_TEAMIV–l9	**80%**	**35.59%**	**28.88%**

for experimenting with various conditions and thresholds of the system out of which 6 major systems that improved the performance overtime are discussed as shown in Table 1. The primary system submitted to the challenge identified as 'SDIAC_TEAMIV_p–c1' operated on audio without speech enhancement with speaker identification threshold of 75%, where the DER was reported as 45.92% and AER was reported as 1191.64%. The system observed a very high false alarm rate due to the inaccuracy of the implemented VAD algorithm which identified segments with only music as speech. Hence, speech enhancement using DE-CADE was implemented to filter out 7,697 seconds of noise and music from the recordings leading to an improved performance as evidenced by the decrease in DER and AER in the subsequent systems.

Two additional systems implementing DE-CADE for speech enhancement with a speaker identification threshold of 75% were submitted for evaluation. SDIAC_TEAMIV–l1 implemented DE-CADE for VAD only and inferencing was performed on audio recordings without speech enhancement based on the speech segments identified by VAD. This system observed a DER of 38.91% and AER of 153.36% . The system SDIAC_TEAMIV–l3 implemented DE-CADE for inferencing and exhibited 40.74% DER and 185.42% AER.

SDIAC_TEAMIV–l1 and SDIAC_TEAMIV–l3 demonstrated better performance compared to the primary system, the false alarm rate was still relatively high and hence the system was experimented with several speaker identification thresholds to tune the system with an ideal threshold for the task.

Out of the various speaker identification thresholds, the systems SDIAC_TEAMIV–l4 and SDIAC_TEAMIV–l8 used speaker identification threshold of 80% and demonstrated superior performance compared to the systems executing alternative thresholds. The DER and AER were reported as 35.82% and 36.07% respectively for system Speaker DiarizationIAC_TEAMIV–l4, which only used speech enhancement for VAD. Whereas the system SDIAC_TEAMIV–l8 processed files produced by DE-CADE speech enhancement for inferencing and observed DER of 37.20% and AER of 44.34%.

Upon detailed analysis of the results, it was observed that the DE-CADE speech enhancement algorithm only impacted the performance positively when there was excessive noise or music in the evaluation audio. Hence, the system SDIAC_TEAMIV–l9 implemented a SNR classifier to decide whether to employ DE-CADE speech enhancement algorithm or not on the input audio signals for inferencing based on the SNR threshold of 0.2 and hence the optimal results were obtained as 35.59% DER and 28.88% AER.

4. Conclusions

The extraction of multiple x-vectors to capture individual speaker speech variability resulting from different speaking styles and varying vocal effort, followed by the use of outlier detection and two-stage clustering for obtaining distilled voice prints of the speakers of interest, underpins the novelty of the paper. The system also implements the DE-CADE speech enhancement algorithm to clean degraded audio signals identified by the SNR classifier for effective speaker diarization and recognition. The results obtained on the RTVE 2022 dataset with the implemented system show promise. In future work, we aim to evaluate the system using larger and more challenging datasets such as VoxCeleb1 & VoxCeleb2 [14], MultiSV [15] and HI-MIA [16].

5. Acknowledgment

Special thanks to Soha A. Nossier for providing DE-CADE inferencing scripts and helpful discussions. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823907 (MENHIR project https://menhir-project.eu)

6. References

[1] R. Shrestha, J. Wall, C. Glackin, N. Cannings, M. Rajwadi, S. Kada, J. Laird, T. Laird, and C. Woodruff, "Speaker recognition using multiple x-vector speaker representations with two-stage clustering and outlier detection refinement," in *CyberSciTech 2022: IEEE Cyber Science and Technology Congress*. IEEE, 2022.

[2] A. Ortega, A. Miguel, E. Lleida, V. Bazan-Gil, C. Perez, and A. d. Prada, "Albayzin Evaluation IberSPEECH-RTVE 2022 Speaker Diarization and Identity Assignment." catedrartve.unizar.es, 2022. [Online]. Available: http://catedrartve.unizar.es/reto2022/SDIAC2022_Evalplan.pdf

[3] 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, 2010.

[4] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[5] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[6] D. Dimitriadis and P. Fousek, "Developing on-line speaker diarization system." in *INTERSPEECH*, 2017, pp. 2739–2743.

[7] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4930–4934.

[8] J.-Y. Yang and J.-H. Chang, "Task-specific optimization of virtual channel linear prediction-based speech dereverberation front-end for far-field speaker verification," 2021. [Online]. Available: https://arxiv.org/abs/2112.13569

[9] E. Lleida, A. Ortega, A. Miguel, V. Bazan-Gil, C. Perez, and A. d. Prada, "RTVE 2018, 2020 and 2022 database description." catedrartve.unizar.es, 2022. [Online]. Available: http://catedrartve.unizar.es/reto2022/RTVE2022DB.pdf

[10] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "The second dihard diarization challenge: Dataset, task, and baselines," *arXiv preprint arXiv:1906.07839*, 2019.

[11] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "The third dihard diarization challenge," 2020. [Online]. Available: https://arxiv.org/abs/2012.01477

[12] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream-based speaker segmentation using speaker factors and eigenvoices," *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4133–4136, 2008.

[13] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The ami meeting corpus: A pre-announcement," in *Machine Learning for Multimodal Interaction*, S. Renals and S. Bengio, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 28–39.

[14] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[15] L. Mošner, O. Plchot, L. Burget, and J. Černocký, "Multisv: Dataset for far-field multi-channel speaker verification," 2021. [Online]. Available: https://arxiv.org/abs/2111.06458

[16] X. Qin, H. Bu, and M. Li, "Hi-mia : A far-field text-dependent speaker verification database and the baselines," 2019.

[17] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj *et al.*, "Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *arXiv preprint arXiv:2004.09249*, 2020.

[18] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1051200499903615

[19] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified [speaker verification applications]," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1, 2005, pp. I/637–I/640 Vol. 1.

[20] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Cosine similarity scoring without score normalization techniques."

[21] J. R. Price and T. F. Gee, "Face recognition using direct, weighted linear discriminant analysis and modular subspaces," *Pattern Recognition*, vol. 38, no. 2, pp. 209–219, 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320304002730

[22] B. J. Borgström, "Discriminative training of plda for speaker verification with x-vectors," 2020.

[23] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2010)*, 2010, p. paper 14.

[24] O. Novotný, P. Matějka, O. Plchot, O. Glembek, L. Burget, and J. Černocký, "Analysis of Speaker Recognition Systems in Realistic Scenarios of the SITW 2016 Challenge," in *Proc. Interspeech 2016*, 2016, pp. 828–832.

[25] H. Ghaemmaghami, M. H. Rahman, I. Himawan, D. Dean, A. Kanagasundaram, S. Sridharan, and C. Fookes, "Speakers in the wild (sitw): The qut speaker recognition system," 09 2016, pp. 838–842.

[26] W. B. Kheder, M. Ajili, P.-M. Bousquet, D. Matrouf, and J.-F. Bonastre, "Lia system for the sitw speaker recognition challenge," in *INTERSPEECH*, 2016.

[27] A. Kanagasundaram, D. Dean, S. Sridharan, J. Gonzalez-Dominguez, J. Gonzalez-Rodriguez, and D. Ramos, "Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques," *Speech Commun.*, vol. 59, p. 69–82, apr 2014. [Online]. Available: https://doi.org/10.1016/j.specom.2014.01.004

[28] M. McLaren, D. Castán, M. K. Nandwana, L. Ferrer, and E. Yilmaz, "How to train your speaker embeddings extractor," in *Odyssey*, 2018.

[29] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5796–5800.

[30] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, L. P. García-Perera, F. Richardson, R. Dehak, P. A. Torres-Carrasquillo, and N. Dehak, "State-of-the-art speaker recognition with neural network embeddings in nist sre18 and speakers in the wild evaluations," *Computer Speech & Language*, vol. 60, p. 101026, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230819302700

[31] 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.

[32] S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, "Two-stage deep learning approach for speech enhancement and reconstruction in the frequency and time domains," in *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–10.

[33] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *The Journal of Open Source Software*, vol. 2, no. 11, p. 205, 2017.

[34] S. Tomar, "Converting video formats with ffmpeg," *Linux Journal*, vol. 2006, no. 146, p. 10, 2006.

[35] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, vol. 5, no. 50, p. 2154, 2020. [Online]. Available: https://doi.org/10.21105/joss.02154

[36] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2