Speaker Recognition using Multiple X-Vector Speaker Representations with Two-Stage Clustering and Outlier Detection Refinement*

Roman Shrestha *, Cornelius Glackin *, Julie Wall †, Nigel Cannings *, Marvin Rajwadi *,
Satya Kada *, James Laird *, Thea Laird *, Chris Woodruff *

*Intelligent Voice Ltd., London, UK
Email: {roman.shrestha, neil.glackin, nigel.cannings, marvin.rajwadi,
satya.kada, james.laird, thea.laird, chris.woodruff}@intelligentvoice.com

†University of East London, London, UK
Email: {j.wall}@uel.ac.uk

***Abstract*—This paper presents a novel Variational Bayes x-vector Voice Print Extraction (VBxVPE) system, capable of capturing vocal variations using multiple x-vector representations with two-stage clustering and outlier detection for robust speaker recognition and verification. The presented approach demonstrates beyond the state-of-the-art results when evaluated against the 'core-core' and 'core-multi' evaluation conditions of the Speakers In the Wild dataset, achieving an Equal Error Rate of 1.06%, Cost of Detection score of 0.052, minimum Cost of Detection score of 0.010, Speaker Identification Accuracy of 95.84% with Precision, Recall and F1 score values of 0.964, 0.958 and 0.961, respectively on the 'core-core' evaluation condition and Equal Error Rate of 1.07%, Cost of Detection score of 0.066, minimum Cost of Detection score of 0.010 with Precision, Recall and F1 score values of 0.967, 0.963 and 0.965, respectively on the 'core-multi' evaluation condition.**

***Index Terms*—Voice Biometrics, Speaker Recognition, Voice Print Extraction, X-Vectors, Speakers in the Wild.**

I. Introduction

Speaker diarization and verification research is gaining traction in recent years with regard to the advances in the state-of-the-art. This increasing interest is due to a number of factors, such as the commercial importance of diarization and biometrics for speech technology and downstream NLP tasks. The accessibility to realistic real-world evaluation datasets, such as DIHARD-2 [1] & DIHARD-3 [2], CALLHOME [3], AMI [4], VoxCeleb [5], MultiSV [6], HI-MIA [6] and CHiME-6 [7]), and advances made with deep learning have nurtured several ground-breaking architectures, fostering an active community of researchers working to try and solve this long-standing complex problem. Diarization is the task of determining the boundaries of speakers in utterances within a conversation, i.e. who is speaking when. 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 features using either i-vectors [8], d-vectors [9], or x-vectors [10]), and clustering the segments using techniques such as k-Means [11] or, Agglomerative Hierarchial Clustering (AHC) [12] to obtain accurate speaker separation from a multi-speaker recording. The real-world evaluation corpora have exposed the complexity of the task for real-world conversational scenarios complicated by noise, reverberation and overlapping speech.

During the early 2000s, the trailblazing speaker recognition systems were based on the Gaussian Mixture Model (GMM)-Universal Background Model (UBM) approach, where the GMMs of individual speakers were adapted from 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 [13]. Kenny et. al [14] proposed the Joint Factor Analysis (JFA) approach which improved GMM estimation by allowing the modelling of interspeaker variability and compensation for channel/session variability in the context of high-dimensional GMM supervectors.

With the advent of i-vectors [8], unique fixed length embeddings extracted from the recordings could be directly used for performing verification using cosine similarity scoring [15]. Since i-vectors were highly susceptible to unwanted variations due to a mismatch of linguistic content and recording channel information between segments of speech spoken by the same speaker, Linear Discriminant Analysis (LDA) and Nuisance Attribute Projection (NAP) were proposed as a solution with improved performance [15]. Probabilistic LDA (PLDA), originally introduced by Price and Gee for facial recognition [16], has emerged as a powerful tool for speaker verification capable of generating well-calibrated likelihood ratios between the vectors [17]. Kenny [18] was amongst the pioneers for implementing PLDA in the i-vector space for modelling channel variability.

Deep Neural Networks (DNN) for extracting feature vectors were found to be effective for achieving better speaker recognition performance [19]–[23]. For many years, DNN-based i-

978-1-6654-6297-6/22/$31.00 ©2022 IEEE

vector systems implementing PLDA scoring were considered the gold standard in speaker verification domain. Recently, x-vectors have emerged as an alternative form of speaker embedding that are better suited for speaker recognition purposes and are 10-25% better than acoustic i-vectors, and slightly better than i-vectors implementing phonetic bottleneck features (BNF) at all operating points [10].

The x-vector based systems were able to demonstrate an impressive performance on speaker recognition and diarization across different acoustic channels [10], [17], [24]–[26]. These systems operated by extracting x-vectors from speech segments, performing LDA and using PLDA classifiers to perform a likelihood ratio test between the enrolled and the test speakers in a verification task. Research employing speech enhancement to cancel out noise, reverberation and normalize distortion from the noisy audio signals have also shown improvement in this domain [27].

In this paper, we propose a robust x-vector based voice print extraction system (VBxVPE) for speaker verification and recognition 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) clustering algorithm, 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 four sections. Section II provides information on the benchmark dataset, the reported evaluation metrics and a review of the literature relating to systems evaluated on the benchmark dataset. System components and the methodology are described in Section. III, results obtained and the discussions are provided in Section IV, finally followed by Conclusions in Section V.

II. Related Work

The proposed speaker verification system was evaluated on the Speakers in the Wild (SITW) corpus developed by the Stanford Research Institute (SRI) International [28] comprising of speech utterances recorded from diverse "in-the-wild" conditions. The corpus was initially released for the SITW Speech Recognition Challenge 2016 with the aim of benchmarking and supporting the development of robust speaker recognition technologies in both single and multi-speaker audios and has nurtured a substantial amount of cutting-edge speaker recognition systems to date. Since the VBxVPE system relies on a PLDA model, pre-trained on a large number of speaker-labeled x-vectors [26], the SITW development set [28] was not required at any stage. The SITW evaluation set is composed of a total of 180 different speakers across 2,883 audio files naturally containing overlapping utterances, noise, reverberation, and compression artifacts, making the dataset challenging from a speaker recognition perspective [28], [29].

The evaluation metrics for evaluating the performance of the proposed speaker verification system are reported in terms of Cost of Detection (C_{Det}) or Detection Cost Function (DCF), Minimum Cost of Detection ($minC_{Det}$), Equal Error Rate (EER), Speaker Identification Accuracy (SIA) along with standard Precision, Recall and F1 scores.

The EER evaluates the operating point at which the missed and false alarm rates are identical. C_{Det} was the primary metric used for the SITW Speech Recognition Challenge 2016 [30] and is usually used to assess performance by computing the weighted sum of cost for miss and false alarm error probabilities [31].

$$C_{Det} = C_{miss} \times P_{miss} \times P_{tar} + C_{fa} \times P_{fa} \times (1 - P_{tar}) \quad (1)$$

From Equation 1 [31], C_{Det} was calculated by setting the prior probability of target speaker occurrence with the recommended thresholds [30]; P_{tar} was set as 0.01 and the costs for both missed C_{miss} and false alarms C_{fa} was set as 1. The optimal value obtained for C_{Det} is regarded as $min\,(C_{Det})$ [30]. The script for calculation of C_{Det}, $min\,(C_{Det})$ and EER was taken from the speechbrain library [32].

SIA can be expressed as the percentage of the Genuine Number of Speakers Identified by the system out of the total Number of speakers as shown in Equation. 2 [33].

$$SIA = \frac{Genuine\ Speakers\ Identified}{Number\ of\ Speakers} \times 100\% \quad (2)$$

The systems developed as part of the SITW Speech Recognition Challenge in 2016 were amongst the first systems to be evaluated with the SITW corpus. The top two systems [19], [20] had both implemented i-vector based speaker recognition systems with a PLDA classifier. The winning system computed the i-vector for each speaker detected by the diarization, and then scored each such i-vector against the i-vector representing the enrollment speaker and the maximum of all scores (log-likelihood ratios) was selected as the final score. The PLDA classifier used by the system was trained with crowd noise at various levels of Signal-to-Noise Ratio (SNR) and was able to achieve an EER of 5.85%, C_{Det} score of 0.506 and $min\,(C_{Det})$ score of 0.5032 on the 'core-core' evaluation set and EER of 7.34%, C_{Det} score of 0.5834 and $min\,(C_{Det})$ score of 0.5650 on the 'core-multi' evaluation set [19].

The system that was ranked second in the SITW Speech Recognition Challenge 2016 [20], was based on a DNN i-vector PLDA system with inter-dataset variability compensation used to improve cross-domain evaluations and achieved 0.6477 C_{Det}, 0.6038 $minC_{Det}$ and 9.69% EER on the 'core-core' evaluation set.

Snyder et. al. [10] proposed a system implementing an augmented x-vector extractor and PLDA obtained minimal

error rates on the evaluation with the SITW core evaluation set with an EER of 4.16% and C_{Det} score of 0.393 which showed an improvement compared to the previous systems [10]. This research was extended to cope with speaker verification and diarization on multi-speaker conversations by removing the fixed AHC threshold and proposing a method to tune the thresholding value for more robust x-vector based diarization with PLDA backend [24]. LDA dimensionality reduction was performed to reduce the dimensions of the x-vectors to 225 and one x-vector per speaker was chosen [24]. All the speaker x-vectors from the test set were compared against the enrollment x-vectors using PLDA, and the ones with the highest PLDA log-likelihood ratio score were considered as the result, and others were discarded [24]. With these improvements, the system was able to achieve an EER of 1.7% and 0.20 C_{Det} on the 'core-core' evaluation set without diarization whereas an EER of 2% and C_{Det} score of 0.22 was obtained for the 'core-multi' evaluation set for diarization without AHC threshold which was regarded as the best benchmark results at that time [24].

Villalba et. al. [25] improved the best performing x-vector based speaker recognition system [24] by employing a JHU-MIT primary fusion system with Factorized Time Delay Neural Network (FTDNN) encoder network for x-vector speaker recognition [25]. During the evaluation phase, the x-vector embeddings were extracted from the first affine transform after the pooling layer and the rest of the layers after the embedding layer were discarded. LDA dimensionality reduction, centering, whitening and length normalization was applied to the x-vectors followed by the PLDA log-likelihood ratio evaluation, achieving an EER of 1.53% on the core evaluation set and 1.82% EER on the core-multi evaluation set [25], [34].

Recently, systems using the Weighted Prediction Error (WPE) speech dereverberation algorithm for cancelling out reverberation and background noise [27] and generating clean audio signals for extracting speaker embeddings have shown improved performance for speaker verification. The waveform amplitude distribution analysis method was employed to estimate the 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) [27]. The Drv-VACE-WPE system was able to obtain an EER of 1.46% and minC_{Det} of 0.143 on the 'core-core' evaluation condition of the SITW corpus [28] which surpassed the existing state of the art results.

In contrast to the deep learning approaches, recent research [33] demonstrated a SIA of 85.83% on the 120 speakers from the single and unbalanced multi-speaker recordings belonging to the SITW corpus by implementing a combined i-vector and classification approach using an Extreme Learning Machine (ELM) for speaker recognition which was much faster to train compared to DNNs, and was capable of employing a universal approximator property to support the predictions [33].

To the best of our knowledge, the proposed VBxVPE speaker verification system outperforms all the existing state-of-the-art systems evaluated against the SITW 'core-core' and 'core-multi' conditions [28]. 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.

III. Experimental Setup

A. Data Preprocessing

The audio files were provided in the Free Lossless Audio Codec (FLAC) file format sampled at 16 KHz by default, along with the metadata and instructions for enrollment and evaluation purposes [28]. These were down-sampled to 8 KHz, a standard sample rate for recording the human voice, and converted to the WAV file format using the Fast Forward Motion Picture Experts Group (FFMPEG) python library [35]. The WAV files were then used for either enrolment (Section III-E) or evaluation (Section III-F) as per the SITW evaluation set specifications [28].

B. VAD and X-Vector Extraction

An energy based Gaussian Voice Activity Detection (VAD) system operates on the audio files 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 (8Khz) network [26]. The extracted x-vectors were reduced to 128 dimensions using LDA dimensionality reduction for further processing.

C. Speaker Diarization

VBx diarization [26] was chosen as the reference architecture for speaker diarization due to its superior performance on three of the most popular datasets for evaluating diarization: CALLHOME [3], AMI [4] and DIHARDII datasets [26]. The Agglomerative Hierarchial Clustering (AHC) algorithm [12] used by the VBx diarization system [26] was replaced by a greedy clustering algorithm which 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 DER by 0.91% [26] when evaluated against the evaluation set of the third DIHARD Challenge [2]. Then, a PLDA model pre-trained on a large number of speaker-labeled x-vectors [26] scores the obtained clusters to verify the likelihood ratio between them [17], thereby preparing the final diarization output detailing who spoke when in the audio file.

D. 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 i.e. Outlier Detection and then HDBSCAN Clustering [36].

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 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 [36] 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.

E. The Enrolment Pipeline

Based on the "assist" enrolment conditions specified on the dataset [28], 742 wav files containing speech from 180 different speakers (∼4 files per speaker) were further segmented into small chunks as per the annotations provided, typically around 5 seconds per recording, which were known to contain the speaker of interest. Since all the speakers in the evaluation set of the corpus could be enrolled using the "assist-enrol" set, other enrolment sets were not required.

Fig. 1 shows the enrolment pipeline for enrolling the speakers from the SITW corpus enrolment set [28] for performing speaker verification with the proposed VBxVPE system. The enrolment procedure commences by accepting the audio and label for the speaker of interest as metadata.

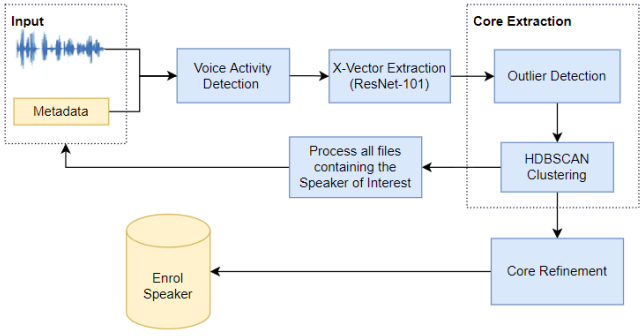

Fig. 1. SITW Enrolment Pipeline

Since the segments specified for enrolment in the "assist" condition only contains speech from the speaker of interest, diarization is not performed in the enrolment pipeline, only in the evaluation pipeline. After VAD and x-vector extraction is performed as described in Section III-B, the voice print for the speaker of interest is extracted across all the files containing the speaker based on the metadata provided as described in Section III-D.

After extracting the cores across all the files containing the speaker of interest, 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 phonetic features constituting the speaker's voice are then enrolled in the vector search database along with a unique speaker id.

F. The Evaluation Pipeline

For evaluation of the speaker verification system, the SITW database supports "core-core" evaluation track composed of 1202 recordings without any overlapping speech segments and "core-multi" evaluation track that requires the enrolled speakers to be identified from the provided 2275 audio files containing multiple speakers within a single recording [28], [30].

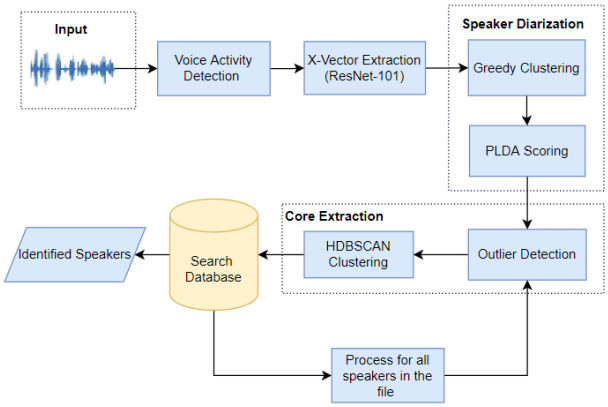

Fig. 2. SITW Evaluation Pipeline

Fig. 2 presents the evaluation pipeline for the VBxVPE system. The proposed speaker recognition system only requires audio recordings containing speech as the input for evaluation. Since the evaluation files may contain multiple utterances from different speakers, diarization is performed as discussed in Section III-C. 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 III-D. 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 75%, determined through trial and error experiments, with any enrolled core representing the speaker in the vector search database. The final output consists

of the identified speaker/s from the recordings, which are compared against the ground truth for the calculation of the evaluation metrics as described in Section II.

IV. Results and Discussion

The proposed speaker recognition system was benchmarked against the SITW corpus [28]. The x-vector extraction was performed on an NVIDIA GeForce GTX 1080 Ti GPU where as the rest of the processes were executed on a single core 64-Bit CPU with Intel Xeon 2.20GHz processor and 128GB RAM. The system was able to enrol 180 speakers across 742 files from the evaluation "assist-enrol" set in 21 minutes and 50 seconds whereas the total time elapsed for generating predictions on the 1202 recordings from the core set was 57 minutes and 5 seconds and 3 hours 37 mins and 3 seconds for the 2275 files from the "core-multi" set of the SITW corpus [28].

The results obtained on the 'core-core' evaluation condition were reported as C_{Det} score of 0.052, $min\,(C_{Det})$ score of 0.010, EER of 1.06%, SIA of 95.84% with 0.964 Precision, 0.958 Recall and 0.961 F1 score values. For the 'core-multi' evaluation condition the system was able to obtain C_{Det} score of 0.066, $min\,(C_{Det})$ score of 0.010, EER of 1.07%, SIA of 96.30% with 0.967 Precision, 0.963 Recall and 0.965 F1 score values. The achieved results show a major improvement in the speaker recognition domain compared to the current state of the art systems, as seen in Table I.

TABLE I
Model Performance 1

System	Core-Core (SITW)			Core-Multi (SITW)		
	C_{Det}	$minC_{Det}$	EER	C_{Det}	$minC_{Det}$	EER
BUT [19]	0.506	0.5032	5.85%	0.5834	0.5650	7.34%
QUT [20]	0.6477	0.6038	9.69%	N/A	N/A	N/A
X-Vec PLDA 2018 [10]	0.393	N/A	4.16%	N/A	N/A	N/A
X-Vec PLDA 2019 [24]	0.20	N/A	1.7%	0.22	N/A	2%
JHU-MIT [25]	N/A	N/A	1.53%	N/A	N/A	1.82%
DRv-VACE-WPE [27]	N/A	0.143	1.46%	N/A	N/A	N/A
VBxVPE	**0.052**	**0.010**	**1.06%**	**0.066**	**0.010**	**1.07%**

TABLE II
Model Performance 2

System	Number of speakers	SIA
i Vector with ELM [33]	120	85.83%
VBxVPE	**180**	**96.30%**

Table I compares the proposed VBxVPE system against the top two systems [19], [20] in the SITW Speech Recognition Challenge 2016 [30] along with four state-of-the-art speaker verification and recognition systems [10], [24], [25], [27]. It can be observed that the VBxVPE speaker verification system demonstrates an improved performance on both the single and multi-speaker settings. The majority of existing speaker verification systems evaluated under the "core-multi" conditions have struggled in comparison to the "core-core" condition due to the presence of multi-speaker utterances within a single clip, in addition to noise, reverberation and compression artifacts [19], [24], [25]. However, The VBxVPE system employs a greedy clustering algorithm in diarization, outlier detection and HDBSCAN clustering implemented for core-extraction, that generates precise x-vector representations of the speaker even in multi-speaker scenarios which is the key to robust speaker recognition [24]. Due to highly accurate speaker separation and effective voice print extraction technology, the proposed system demonstrates optimal performance across the varying conditions across both the evaluation sets. As shown in Table II, compared to the approach combining i-vector with ELM [33], the proposed VBxVPE system was able to achieve 96.30% SIA across 180 speakers demonstrating an improved performance on the "core-multi" evaluation set, which also contained a mixture of single and multi-speaker recordings.

V. Conclusions

Text-independent speaker recognition and verification is still regarded as a challenging field of research owing to the complications surrounding the human voice, as it is sensitive to the speaker's environment, mood and language. A myriad of research and advancements in this domain have facilitated the development of ground-breaking systems and architectures that enable the systems to identify the speakers present in the recording with impressive accuracy. The presented x-vector based voice biometric system aims to refine the speaker verification approach further by proposing a novel voice print extraction algorithm able to capture numerous speech based phonetic variations associated with the speaker of interest using multiple x-vector representations through outlier detection refinement and two-stage clustering.

The experiments conducted on the VBxVPE system and the results obtained on the SITW benchmark [28] demonstrate a major improvement in the speaker recognition and verification domain. It was observed that by implementing a highly accurate diarization system, optimal speaker recognition performance can be obtained on multi-speaker recordings as well. Also, an effective outlier detection algorithm can eliminate noisy, distorted and overlapping speech vectors leading to the extraction of high-quality voice prints for improved speaker recognition. Our research demonstrates promising results on the SITW dataset [28] containing challenging recordings from 180 different speakers in an intrinsic setting without any domain specific tuning. In future work, we aim to enhance the VBxVPE speaker verification system by applying robust dereverberation and speech enhancement and evaluate the system using larger and more challenging datasets including multilingual speech such as VoxCeleb [5], MultiSV [6] and HI-MIA [37].

Acknowledgment

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)

REFERENCES

[1] 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.

[2] 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

[3] 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.

[4] 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.

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

[6] 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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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

[14] 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.

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

[16] 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

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

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

[19] 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.

[20] 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.

[21] 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.

[22] 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

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

[24] 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.

[25] 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

[26] 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.

[27] 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

[28] M. McLaren, L. Ferrer, D. Castán, and A. D. Lawson, "The speakers in the wild (sitw) speaker recognition database," in *INTERSPEECH*, 2016.

[29] 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

[30] M. McLaren, L. Ferrer, D. Castán, and A. D. Lawson, "The 2016 speakers in the wild speaker recognition evaluation," in *INTERSPEECH*, 2016.

[31] N. Scheffer, L. Ferrer, M. Graciarena, S. Kajarekar, E. Shriberg, and A. Stolcke, "The sri nist 2010 speaker recognition evaluation system," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5292–5295.

[32] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[33] M. T. S. Al-Kaltakchi, M. A. M. Abdullah, W. L. Woo, and S. S. Dlay, "Combined i-vector and extreme learning machine approach for robust speaker identification and evaluation with sitw 2016, nist 2008, timit databases," vol. 40, no. 10, 2021. [Online]. Available: https://doi.org/10.1007/s00034-021-01697-7

[34] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, F. Richardson, S. Shon, F. Grondin, R. Dehak, L. P. García-Perera, D. Povey, P. A. Torres-Carrasquillo, S. Khudanpur, and N. Dehak, "State-of-the-Art Speaker Recognition for Telephone and Video Speech: The JHU-MIT Submission for NIST SRE18," in *Proc. Interspeech 2019*, 2019, pp. 1488–1492.

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

[36] 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.

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