Speaker Diarization with Embeddings from a VGGish Model

Bishal Lamichhane
lamichhane.bishal@gmail.com

***Abstract*—VGGish model based embeddings are commonly used for representing audio segments in several audio processing applications. In this work, the relevance of VGGish model embeddings for speaker diarization in the context of the DIHARD-III challenge was investigated. The pre-trained VGGish network (VGGish-Vanilla) is primarily developed as an audio classification model. Therefore, the VGGish-Vanilla model was adapted towards speaker recognition using the Voxceleb1 dataset to obtain VGGish-Adapted model. Then embeddings from the VGGish-Vanilla and the VGGish-Adapted were compared with the commonly used Xvectors for the diarization task. A simple diarization pipeline based on gaussian divergence for speaker segmentation and AHC clustering with cosine distance based scoring was used in this comparison. The adaptation of VGGish network towards speaker recognition improved diarization performance. However, the embeddings from VGGish models were still not competitive compared to the Xvectors. Given that the VGGish network is primarily trained for audio classification, the VGGish embeddings could rather be useful for the identification of audio context/domain and adaptation of the diarization pipeline based on the identified context/domain. The preliminary evaluations done in this work, with AHC threshold adapted to identified audio context, showed that diarization pipeline adaptation gives marginal gains in diarization performance. The adaptive threshold provided gains in the DIHARD-III baseline system also where DER improved from 20.31 (with global threshold) to 19.58 (with adaptive threshold) in the development set. Further investigation is required to better understand how and where pre-trained networks like VGGish model could be useful in speaker diarization tasks. Identification of audio context/domain based on VGGish embeddings for a context-dependent diarization model seems to be a promising direction for further explorations.**

I. Introduction

Speaker diarization of an audio recording is the process of identifying 'who spoke and when' in the given recording. Several datasets have been used for developing and validating speaker diarization methods. Good diarization performance has been obtained in many datasets such as AMI [1], CallHome [2], NIST SRE 2003 [2], etc. However, these datasets mostly represented audio from a single domain only like meetings, telephone calls, or broadcasts. To better estimate the speaker diarization performance across diverse domains and challenging conditions representative of real-world scenarios, the DIHARD diarization challenge dataset was released. The dataset consists of audio recordings from diverse domains such as restaurant conversations, clinical interactions, meetings, youtube recordings, etc.

The state-of-the-art in speaker diarization is obtained with Xvectors [3]. Xvectors are a fixed dimensional representation of a (variable length) audio recording obtained from a deep neural network model trained for speaker recognition. In general audio processing tasks, one of the most commonly used models for audio processing is the VGGish model [4]. A pre-trained VGGish model (referred to as VGGish-Vanilla here), trained on a large AudioSet dataset [5] for audio classification, is used to generate embeddings or features from audio for various downstream tasks. This work investigates if and how the VGGish model could be helpful for the speaker diarization task. Towards that end, the embeddings generated from different layers in the VGGish model were directly used in the speaker diarization pipeline for discriminating speakers. Further, as the VGGish-Vanilla network has been trained on an audio classification task, the network was adapted for speaker recognition using the VoxCeleb1 dataset [6]. This was done by training the VGGish model for speaker recognition in multi-class classification setting followed by siamese training to make embeddings discriminative. The VGGish model trained and adapted for speaker recognition is referred to as the VGGish-Adapted model in this work. The embeddings from the VGGish models were compared to the Xvectors for diarization in the DIHARD-III dataset.

As for the speaker diarization method, a simple diarization pipeline consisting of speaker segmentation, embedding generation, and Agglomerative Hierarchical clustering (AHC) with cosine distance scoring was used. With this diarization pipeline, the diarization performance obtained from Xvectors and VGGish model based embeddings (extracted from different layers within the model) were evaluated.

II. Data resources

The VGGish-Vanilla model has been trained on the Audioset dataset and the trained model available at [7] was used in this work. For the adaptation of the VGGish model towards speaker recognition, the Voxceleb1 dataset [6] was used. Voxceleb1 dataset consists of more than 100,000 utterances from 1251 speakers obtained from youtube recordings. For the Xvectors, the Kaldi SRE16 model available at [8] was used. This model has been trained using Switchboard [9], Mixer 6 [10], and NIST SREs dataset [11]. The DIHARD-III dataset used for development and evaluation consists of 254 audio recordings in the development set and 259 audio recordings in the evaluation set (full set). A subset of recordings from the full set, referred to as the core set, is also identified where the recordings from different domains are roughly balanced in their representation in the dataset. Further information about the DIHARD-III dataset is available from [12].

III. Detailed description of algorithm

The basic pipeline used for speaker diarization in this work was: (Oracle) SAD, gaussian divergence based speaker segmentation, embedding extraction, and AHC (Agglomerative hierarchical clustering) with cosine distance based scoring.

Diarization for Track 1 only was pursued in this work and thus the oracle SAD segmentation available in the dataset was used in the diarization pipeline.

Within each speech segment of the oracle SAD, the possible speaker change locations were obtained from the local maximums of gaussian divergence based on the MFCC features in the two halves of a moving window. Successive segments from the same speaker were fused according to Bayesian Information Criteria (BIC). The implementation in sidekit toolkit [13] was used for this speaker change detection, with a window size set to 1 second.

For each segment identified as being speaker-homogeneous, its embedding was extracted using VGGish model and Xvector model.

A. VGGish model

VGGish model is a deep neural network model consisting of several layers of convolution and max-pooling, followed by three fully connected layers of size 4096, 4096, and 128 respectively.

1) Input features: The input for the VGGish model is 64 dimensional log mel spectogram features computed from a frame of 25 ms with 10 ms overlap. Features from a 960 ms window are accumulated and provided as input to the model for embedding extraction. All the embeddings from a speaker-homogeneous segment were averaged to represent the segment.

2) Embeddings: The embeddings generated from the three top fully connected layers of the VGGish model, each with dimensions of 128 (layer E1), 4096 (layer E2), and 4096 (layer E3) respectively were extracted. As described earlier, two VGGish models were evaluated for embedding extraction: VGGish-Vanilla (pre-trained VGGish model) and VGGish-Adapted (VGGish-Vanilla adapted for speaker recognition using Voxceleb1 dataset). Thus, the following embeddings were evaluated for representing speaker segments in diarization:

- VGGish-Vanilla-E1,-E2, -E3: E1, E2, and E3 embedding respectively from the VGGish-Vanilla model
- VGGish-Adapted-E1, -E2, -E3: E1, E2, and E3 embedding from the VGGish-Adapted model

B. Xvector model

We used the Xvector extraction model trained on the Switchboard, Mixer6, and NIST SRE dataset as available in [8].

The embeddings from different models were scored using cosine distance metric and clustered using Agglomerative Hierarchical clustering (AHC). The threshold for stopping the clustering is obtained based on a parameter sweep in the DIHARD-III development set (Global Threshold Approach). An adaptive threshold approach was also evaluated where the threshold for a recording is obtained based on its similarity to other recordings for which optimal threshold has been identified based on a threshold sweep (Adaptive Threshold Approach). The adaptive threshold approach would allow setting a per-recording threshold, the threshold being adapted to the particular domain/context of the audio recording.

IV. Results on the development set

Speaker diarization performance in the DIHARD-III development set, with different embeddings, was evaluated using the Diarization Error Rate (DER) and Jaccard Error Rate (JER) metric. The performance in the full set and the core set was assessed and the obtained results are presented in .

TABLE I
Diarization Error Rate (DER) and Jaccard Error Rate (JER) obtained with speaker diarization pipeline using different embeddings for Dihard-III development set. The evaluation results for both the full set and the core set are provided.

Embedding	Full		Core	
	DER	JER	DER	JER
Dummy baseline (all same speakers)	39.32	68.90	38.14	69.28
Vggish-Vanilla-E1	37.71	64.31	35.87	63.02
Vggish-Vanilla-E2	38.13	63.99	36.16	62.71
Vggish-Vanilla-E3	37.41	63.53	35.45	62.04
Vggish-Adapted-E1	35.06	59.31	33.50	59.04
Vggish-Adapted-E2	36.95	59.59	35.37	59.16
Vggish-Adapted-E3	38.16	63.92	36.73	63.82
Xvector (SRE16)	34.13	55.32	33.46	59.77

The VGGish model based embeddings, though provided improved diarization performance after adapting the model towards speaker recognition, still under-performed the Xvector representation.

The ability of VGGish embeddings to represent the context/domain of an audio recording was also investigated. Identification of context/domain would be helpful to tune a diarization pipeline specifically towards a recording's context/domain. This context adaptation was done in this work by identifying the best threshold for AHC per recording, based on the similarities of audio recordings (obtained with cosine distance between mean VGGish-Vanilla E1 embeddings obtained for the whole recording). The results of the adaptive threshold approach in comparison to the fixed AHC threshold (Global Threshold Approach) with VGGish embeddings used for assessing the similarity of recordings is given in Table IV.

TABLE II
Diarization Error Rate (DER) and Jaccard Error Rate (JER) obtained with Xvectors based speaker diarization pipeline when using adaptive threshold per recording, with mean embedding per recording used to identify similarity of recordings.

Threshold	Full		Core	
	DER	JER	DER	JER
Adaptive Threshold (with VGGish-Vanilla embedding)	32.26	57.21	31.37	58.55
Global Threshold	34.13	55.32	33.46	59.77

As the AHC threshold adaptation based on recording similarity assessed with VGGish-Vanilla E1 embeddings gave

improved diarizaiton performance, the adaptive threshold approach was tested for the DIHARD-III baseline system in [14] also. The baseline system without the resegmentation step was used for the evaluation. The results obtained for the development set are presented in Table IV. Due to time constraints, evaluation set submission for this Dihard-3 baseline with adaptive threshold could not be made.

TABLE III
DIARIZATION ERROR RATE (DER) AND JACCARD ERROR RATE (JER) OBTAINED WITH DIHARD-III BASELINE SPEAKER DIARIZATION PIPELINE WHEN USING ADAPTIVE THRESHOLD PER RECORDING, WITH MEAN EMBEDDING PER RECORDING USED TO IDENTIFY SIMILARITY OF RECORDINGS.

Threshold	**Full**		**Core**	
	DER	**JER**	**DER**	**JER**
Adaptive Threshold (with VGGish-Vanilla embedding)	19.58	31.26	19.74	33.76
Global Threshold	20.31	41.86	20.63	45.69

From the results in Table IV, it can be seen that VGGish model could be adapted such that the embeddings generated from the network are more suitable for speaker recognition and diarization. However, the embeddings generated from the VGGish-Adapted model still under-performed the Xvectors. This might just be reflecting the effort put into the adaptation of the VGGish model. For example, the much larger Voxceleb2 dataset was not utilized and only limited model adaptation techniques were investigated. Further, due to the time constraints, only a simple backend scoring based on cosine distance was evaluated. The standard backend scoring with PLDA and resegmentation techniques need to be included in the pipeline for further investigation.

The AHC threshold adaptation were found to be helpful for diarization performance. Since VGGish-Vanilla network are trained for audio classification, they could be well suited for the task of identifying audio context of a recording and adapting diarization pipeline accordingly. This was tested and validated in the diarization pipeline implemented in this work. Mean VGGish-Vanilla embeddings of each recording were used to automatically select the best AHC threshold for a given audio recording based on the identified thresholds from other similar audio recordings.

For a mandatory submission on the evaluation set in the DIHARD-III challenge, the output based on Xvectors and adaptive threshold was submitted. The results obtained for the test set are given in Table IV.

TABLE IV
DIARIZATION ERROR RATE (DER) AND JACCARD ERROR RATE (JER) OBTAINED WITH XVECTORS BASED SPEAKER DIARIZATION PIPELINE WHEN USING ADAPTIVE THRESHOLD PER RECORDING FOR THE EVALUATION SET OF DIHARD-III.

Diarization Pipeline	**Full**		**Core**	
	DER	**JER**	**DER**	**JER**
Xvectors with adaptive threshold	31.96	63.61	33.06	67.67

V. HARDWARE REQUIREMENTS

The model training for the adaptation of the VGGish-Vanilla model using the Voxceleb1 dataset was done on a server with 125 GB of RAM and two NVIDIA GeForce GTX 1080 Ti GPUs. All other inferences and diarization evaluations were done on a PC with 24 GB RAM and an NVIDIA Quadro P520 GPU. As a representative estimate of run time for evaluations, 'wall clock time' for diarization of a typical sequence (10 min audio recording) with speaker change detection, embedding generation, AHC clustering, and evaluation was about 30 seconds in the PC environment.

REFERENCES

[1] 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.

[2] A. Canavan, D. Graff, and G. Zipperlen, ""callhome american english speech ldc97s42," 1997, lDC Catalog. Philadelphia: Linguistic Data Consortium.

[3] 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)*, 2018, pp. 5329–5333.

[4] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. [Online]. Available: https://arxiv.org/abs/1609.09430

[5] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[6] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[7] "Vggish model," https://github.com/tensorflow/models/tree/master/research/audioset/vggish.

[8] "Kaldi sre16 xvector model," http://www.kaldi-asr.org/models/m3.

[9] J. J. Godfrey and E. Holliman, "Switchboard-1 release 2," 1993, lDC Catalog No.: LDC97S62.

[10] L. Brandschain, D. Graff, K. Walker, and C. Cieri, "Mixer 6 speech," 1993, lDC Catalog No.: LDC2013S03.

[11] C. Greenberg, O. Sadjadi, T. Kheyrkhah, K. Jones, K. Walker, S. Strassel, and D. Graff, "2016 nist speaker recognition evaluation test set," 2019, lDC Catalog No.: LDC2019S20.

[12] N. Ryant, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "Third dihard challenge evaluation plan," 2020.

[13] P.-A. Broux, F. Desnous, A. Larcher, S. Petitrenaud, J. Carrive, and S. Meignier, "S4D: Speaker Diarization Toolkit in Python," in *Interspeech*, Hyderabad, India, Sep. 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02280162

[14] "Dihard-iii baseline model," https://github.com/dihardchallenge/dihard3_baseline.