The BGU 2018 NIST Speaker Recognition Evaluation System

Nave Algarici , Haim Permuter

Department of Electrical Engineering, Ben-Gurion University, Beer-Sheva, Israel

ABSTRACT

- In our submitted speaker recognition system, we exploited the advances of Deep Neural Networks in the field of speaker recognition. We use the framework of the X-vector extractor for speaker embedding, and for diarization. In addition, we adapt the PLDA using in-domain data to better fit it to the task. We have improved the EERs of the Call My Net 2 (CMN2) and the Video Annotation for speech Technology (VAST) data sets by 16% and 28% respectively, in comparison to the best results published in the 2018 baseline systems.

1. SYSTEM DESCRIPTION

1.1. Acoustic features

- The features are 23 MFCCs with a frame length of 25ms every 10ms using a 23 channel mel-scale filterbank spanning the frequency range 20Hz-3700Hz.
- Feature vectors are mean-normalized over a sliding window of up to 3 seconds.
- Delta and acceleration are appended to create 60 dimension feature vectors.
- Energy based speech activity detection (SAD) is applied to select only features that correspond to speech frames.

1.2. Speaker diarization

- In the case of the VAST data set, we apply speaker diarization on the test segments.
- Each utterance is segmented using the same SAD mentioned in 1.1.
- For each speech segment in the utterance a speaker embedding is extracted, and a PLDA score is given to each pair of segments.
- Segments are clustered using agglomerative hierarchical clustering (AHC).
- The most dominant cluster in the recording is chosen to replace the original recording in the evaluation, by concatenating the segments belonging to it.

1.3. Speaker embedding

- Speaker embeddings are extracted using a pre-trained X-vector extractor model.
- Unlike in the original algorithm, do not split the input of the extractor in to chunks and average the X-vectors, instead use all the feature vectors (after VAD or diarization) to extract a single X-vector.
- The 512-dimentional speaker embeddings are centered, whitened, and unit-length normalized.

1.4. LDA

- Dimensionality reduction from 512 to 150 is performed using linear discriminant analysis.

1.5. PLDA

- For scoring, a Gaussian PLDA model with a full-rank Eigen-voice subspace is used.

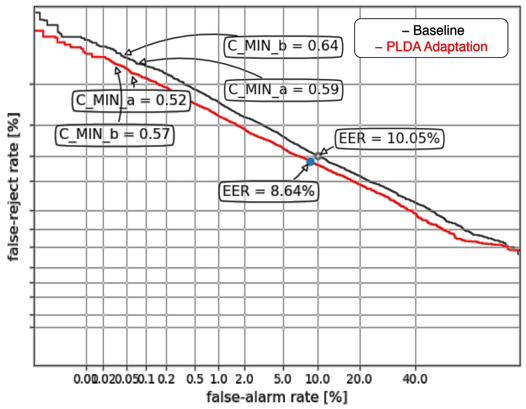
1.6 PLDA Adaptation

- In the case of the CMN2 data set, an adapted version of the PLDA scorer is trained using in-domain unlabeled data.
- Use out-of-domain PLDA to cluster the in-domain dataset.
- Clusters are treated as speakers and are subsequently used to adapt the parameters of the PLDA system to the new domain.

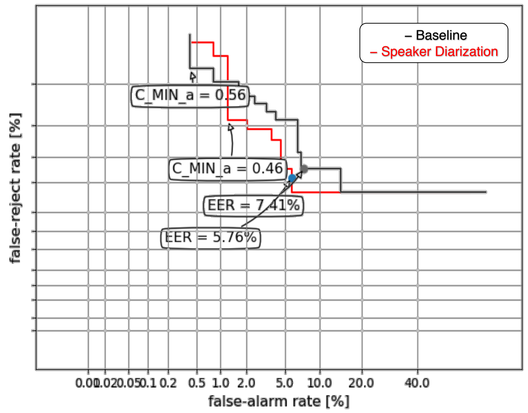
2. DATA DESCRIPTION

- The x-vector extractor is trained using conversational telephone and microphone speech data extracted from the NIST 2004-2010 SRE datasets, as well as MIXER 6, Switchboard Cellular (SWB-CELL) Parts I and II, and Switchboard (SWB) Phases I, II, and III corpora
- The Gaussian PLDA is trained using the x-vectors extracted from all speech segments from the SRE and MIXER 6 sets.
- The PLDA adaptation is performed using the unlabeled portion of the development set.

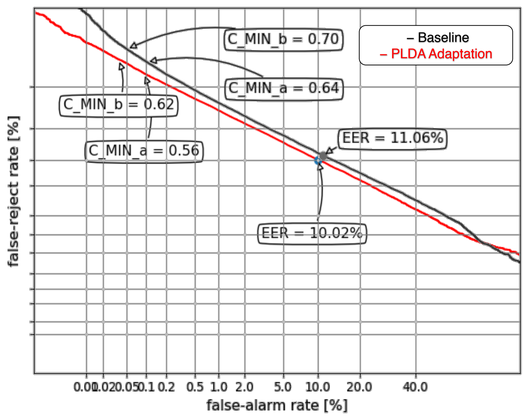
3. RESULTS



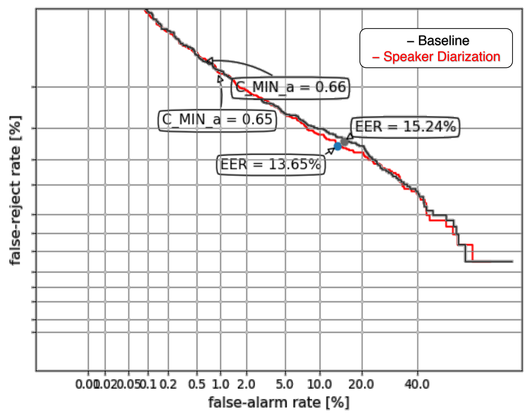
Performance on CMN2 development dataset



Performance on VAST development dataset



Performance on CMN2 evaluation dataset



Performance on VAST evaluation dataset

Dataset	System	CMN2			VAST			
		EER(%)	min DCF	actual DCF	EER(%)	min DCF	actual DCF	Total DCF
Development	I-Vector (baseline)	10.38	0.64	0.892	9.05	0.630	0.778	0.835
	X-Vector (baseline)	9.96	0.647	0.844	7.41	0.572	0.704	0.774
	X-Vector + PLDA adaptation	**8.64**	**0.544**	**0.557**	-	-	-	**0.507**
	X-Vector + diarization	-	-	-	**5.35**	**0.457**	**0.457**	
Evaluation	X-Vector (baseline)	11.06	0.671	0.958	15.24	0.656	0.677	0.818
	X-Vector + PLDA adaptation	**10.02**	**0.593**	**0.602**	-	-	-	**0.633**
	X-Vector + diarization	-	-	-	**13.65**	**0.654**	**0.663**	

4. CONCLUSIONS

- Utilizing the unlabelled data to train the PLDA adaptation has improved the performance significantly.
- Applying diarization on the VAST test set helped remove other speakers and background audio, so the speaker embedding is more relevant and helped the overall performance of the system.