The SRI International STAR-LAB DiHard Challenge System Description

Mitchell McLaren[1], *Diego Castan*[1], *Martin Graciarena*[1], *Luciana Ferrer*[2]

[1]Speech Technology and Research Laboratory, SRI International, California, USA
[2]Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Argentina
{mitch,dcastan,martin}@speech.sri.com, lferrer@dc.uba.ar

Abstract

This document describes the submissions of STAR-LAB (the Speech Technology and Research Laboratory at SRI International) to both track 1 and track 2 of the DiHard 2018 challenge. The core components of the submissions included noise-robust speech activity detection including domain detection, speaker embeddings for initializing diarization as well as post-diarization linking, and variational Bayes (VB) diarization using two DNN bottleneck MFCC i-vector subspaces.

1. Introduction

SRI International has long focused on the task of speaker recognition, but has only recently branched into the field of speaker diarization. Our submissions attempt to leverage recent work in speaker embeddings for speaker recognition [1, 2, 3] the well-known variational bayes (VB) approach to diarization [4], and the fusion of two DNN bottleneck based i-vector subspaces internal to the VB process. We describe three systems being our baseline VB approach, the hybrid embeddings-VB approach, and finally the use of an additional system via fusion in the VB clustering process.

2. System Training and Development Data

Table 1 shows the databases used to train the SAD model. The clean speech data from RATS is composed of several languages: English, Urdu, Pashto, Farsi and Levantine Arabic; source audio has been drawn from LDCs Fisher English and Fisher Levantine Arabic corpora, plus new conversational telephone speech (CTS) data collected specifically for RATS. A total of 402.8 hours of source audio was processed for this task; the 8 transceiver channels yielded over 3222 hours of retransmitted audio. Overall, roughly 45% of the audio content is speech. The music used to pollute the clean speech is non-vocal and it is mainly composed of jazz and classical tracks. finally, we excluded the cafe noise from QUT noises to pollute the clean speech. Also the databases used to train the UBM and the total variability subspaces as well as the speaker embeddings system are shown in the same table. Details of each system are given in the following section.

3. The STAR-LAB System Submissions

We start with a general overview of each submission prior to breaking down into details of each module used in the submissions. Figure 1 shows a block diagram of the different parts of our submissions.

1. **Baseline**: Use of the BNiv$_1$ system in traditional VB diarization
2. **Hyb-BN-Emb**: Embeddings soft clustering as seed to VB diarization prior to embeddings-based speaker linking.
3. **Hyb-Multi**: Same as Hyb-BN-Emb with the use of both BNiv$_1$ and BNiv$_2$ in VB diarization fused at the speaker posterior level after each iteration.

Table 1: *Databases used for SAD training models*

System	Databases
SAD	LDC HAVIC database Clean speech from RATS SAD with music Clean speech from RATS SAD with QUT noises Noisy channel from RATS SAD
Embeddings	PRISM (NIST SRE'04-08)
BN Extractor UBM-IV	Fisher, Switchboard, AMI PRISM (NIST SRE'04-08)

3.1. Acoustic Features and Bottleneck i-vector extraction

We have used the Mel Frequency Cepstral Coefficients for our submissions. We extracted 80-dimensional bottleneck (BN) features from two different DNNs; only the the Hyb-Multi system leveraged both BN extractors.

For BNiv$_1$, a DNN was trained to predict 1933 English tied tri-phone states (senones). MFCCs were used for input to the DNN after transforming them with a pcaDCT transform [5] trained on Fisher data, and restricting the output dimension to 90. The DNN consisted of 5 hidden layers of 600 nodes, except the last hidden layer which was 80 nodes and formed the bottleneck layer from with activations were extracted as features.

The BNiv$_2$ features were similarly extracted from a DNN trained to predict 3k English senones, used a larger hidden layer size of 1200 nodes, but had more traditional input features being Log Mel Spectra of 40 dimensions, stacked across 15 frames.

Each set of BN features the above DNNs were used to train an i-vector extractor [6] consisting of a 2048-component universal background model (UBM) with diagonal covariance and a subspace of rank 400.

3.2. Domain Dependent Speech Activity Detection

We leverage a deep neural network-based SAD system in our submissions. These systems use a DNN trained to predict the posterior of the speech and non-speech classes at the output layer. The posteriors are converted into log-likelihood ratios (LLRs) by using Bayes rule, assuming equal priors for both classes. In a final step, these LLRs are smoothed by averaging their values over a rolling window (31 frames long in our case). The final SAD decisions are made by thresholding these LLRs. The DNN has three hidden layers with five hundred neurons each.

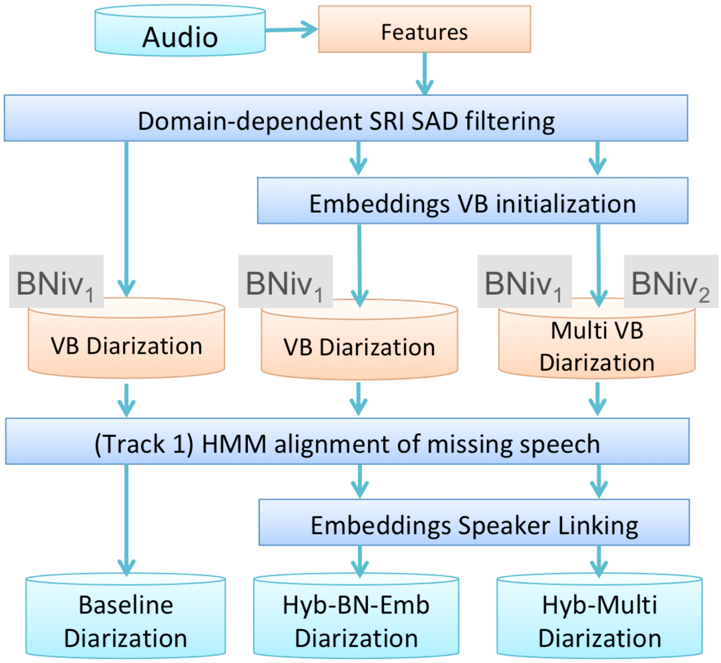

Figure 1: *Flow diagram of components used in the STAR-LAB team submissions to the DiHard 2018 challenge. The paths of data flow for the three different submissions are given.*

Regarding the threshold, development experiments on track 2 highlighted that the best SAD threshold was highly domain dependent, being -1.5 and 0.5 split equally between the develop set domains. This was an important finding since SAD errors are folded into DER for track 2 submissions. We therefore implemented a simple domain detection approach in which development i-vectors from the BNiv$_1$ system were used to train a Gaussian backend (as used in language identification) for each domain. The development and evaluation i-vectors were then used to predict the domain from which they originated and the appropriate SAD threshold used. The model detects the set correctly 92% of the time in development data.

In the case of track 1 submissions, we determine the output of our SRI SAD system, and remove any frames detected as speech that are marked as non-speech in the provided speech annotations. These frames are used for diarization on high quality speech. The speech not detected by the SRI SAD system that is labeled as speech in the annotations is later labeled using a HMM and Viterbi alignment after VB diarization has been performed.

3.3. Embeddings VB Initialization

Recent work in [1, 2] has shown significant advances in the related field of *speaker recognition* by replacing the i-vector extraction process with speaker embeddings extracted from a DNN trained to directly discriminate speakers. We decided to apply our findings on what makes a good speaker embeddings extractor [3] to the task of speaker clustering.

The model involves five frame-level hidden layers of 512 or 1500 nodes, a statistics pooling layer and two segment-level hidden layers of 512 nodes. Apart from the statistics pooling layer, hidden layers were based on a rectified linear unit (ReLU) activation and batch normalization, and the first three layers incremented time context in the network. In our submissions, we have used embeddings to seed the VB and to link speakers for the final alignment.

More specifics on the embeddings system can be found in [3] where the system used in this evaluation is referred to as raw+CNLRMx4. This system used PLDA classification for clustering and speaker linking.

The embeddings VB initialization process was performed as follows. The audio was first segmented into 1.5 second segments with 0.2 second shift. SAD was applied and segments with less than 0.15 seconds of speech were discarded. Following a similar strategy to VB diarization, we initialized a speaker cluster posterior matrix, q, to for 6 speakers. We calculated for each speaker cluster, a weighted-average embedding based on q and the 1.5s embeddings segments. These per-cluster embeddings were compared using PLDA against each individual embedding segment. We scaled the likelihood ratios (LLRs) that resulted from PLDA by 0.05 and performed Viterbi decoding of the LLRs to result in a new q and speaker priors. This process was iterated 8 times before using the result q and speaker priors in the subsequent VB diarization based on BN+MFCC features.

3.4. Variational Bayes diarization

We have implemented a frame-level diarization in a i-vector subspace [4, 7] where the statistics have been computed using a space given by concatenated BN+MFCC features [8]. With VB diarization, we have used a left-to-right HMM structure of three states per speaker in order to smooth the transitions between speakers that was proposed in [9].

With the exception of the Baseline system, the initialization of the VB diarization approach is done with the speaker posteriors estimated from the speaker embeddings initialization. We performed a maximum of 20 iterations of VB diarization.

In the case of our Hyb-Multi submission, we developed a parallel VB diarization scheme in which two i-vector subspaces were used and the speaker posterior matrix for each was averaged after each iteration. This allowed each subspace to start from the same point after each iteration and maintain soft-fusion of information within the VB process.

Table 2: *Evaluation results for each system submission*

System Name	DER	MI
Baseline - Track1	30.56%	8.27
Hy-BN-Emb - Track1	27.98%	8.33
Hyb-Multi - Track1	27.61%	8.33
Baseline - Track2	45.20%	7.79
Hy-BN-Emb - Track2	41.83%	7.87
Hyb-Multi - Track2	41.56%	7.87

Table 3: *Computational requirements of STAR-LAB submissions from based on RT factor (higher than 1.0 is slower than real time), and maximum resident memory needed to diarize the 10m 53s development file DH_0083.flac.*

System	x RT	Max. Res. RAM
Baseline	0.59	3.45G
Hybrid	0.96	4.25G
Multi-Hybrid	1.27	5.06G

3.5. HMM Alignement of Missing Speech

As mentioned previously, the speech not detected by the SRI SAD system that is labeled as speech in the annotations was not use in the VB diarization process. Instead, it was labeled using a HMM and Viterbi alignment after VB diarization had been performed to ensure SAD errors were not counted in our track 1 submission.

3.6. Embeddings Speaker Linking

We noted during development that several single-speaker files were split into several clusters. This was due to an over clustering problem, which tended to only occur in the one or two speaker files. To aid in linking these same-speaker clusters together, we employed a straight-forward embeddings PLDA speaker recognition system. The following process was iterated: Audio from each cluster was used to extract a corresponding embeddings; an exhaustive comparison of these embeddings was performed; if the maximum score was above a threshold of 20.0, the two clusters were merged, otherwise the process stopped.

4. Results

This section compares development and evaluation performance of the STAR-LAB submissions.

Table 2 shows the system performances for both track 1 and track 2 for the eval set. Note how the use of the BN features into the VB diarization reduce the DER in a 10% relative respect the baseline system for both tracks. However, the system Hyb-Multi is the best system in terms of DER also for both tracks.

5. Computation

We benchmarked the computation requirements of each of the STAR-LAB submissions on a single core. The machine was an Intel Xeon E5-2760 Processor operating at 2.6GHz. The approximate processing speed and resource requirements are listed in Table 3.

These calculations are based on total CPU time divided by the total duration of of the audio.

6. Acknowledgments

We'd like to thank Lukas Burget and the BUT team for their python implementation of Variational Bayes diarization which was leveraged in this work [9].

7. References

[1] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and Sa Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 165–170.

[2] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proc. Interspeech*, 2017, pp. 999–1003.

[3] M. McLaren, D. Castan, M. Nandwana, L. Ferrer, and E. Yilmaz, "How to train your speaker embeddings extractor," in *Submitted to Speaker Odyssey*, 2018.

[4] Patrick. Kenny, "Baysedian analyssi of speaker diarization with eigenvoice priors," in *Tech. Rep. CRIM*, 2008.

[5] Mitchell McLaren and Yun Lei, "Improved speaker recognition using dct coefficients as features," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4430–4434.

[6] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Speech and Audio Processing*, vol. 19, pp. 788–798, 2011.

[7] Gregory Sell and Daniel Garcia-Romero, "Diarization resegmentation in the factor analysis subspace," in *ICASSP*, 2015.

[8] M. Mclaren, L. Ferrer, and A. Lawson, "Exploring the role of phonetic bottleneck features for speaker and language recognition," in *Proc. ICASSP*, 2016.

[9] *VB diarization with eigenvoice and HMM priors*, 2013, `http://speech.fit.vutbr.cz/software/vb-diarization-eigenvoice-and-hmm-priors`.