Odyssey 2020 The Speaker and Language Recognition Workshop
1-5 November 2020, Tokyo, Japan

Speaker detection in the wild: lessons learned from JSALT 2019

Paola Garcia[1], *Jesús Villalba*[1], *Hervé Bredin*[2], *Jun Du*[3], *Diego Castan*[4], *Alejandrina Cristia*[5], *Latané Bullock*[9], *Ling Guo*[8], *Koji Okabe*[8], *Phani Sankar Nidadavolu*[1], *Saurabh Kataria*[1], *Sizhu Chen*[10], *Leo Galmant*[2], *Marvin Lavechin*[5], *Lei Sun*[3], *Marie-Philippe Gill* [7], *Bar Ben-Yair*[1], *Sajjad Abdoli* [7], *Xin Wang*[3], *Wassim Bouaziz*[5], *Hadrien Titeux*[5], *Emmanuel Dupoux*[6], *Kong Aik Lee*[8], *Najim Dehak*[1]

[1] Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, USA,
[2] LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, Orsay, France
[3] University of Science and Technology of China, Hefei, China
[4] Speech Technology and Research Laboratory, SRI International, California, USA
[5] Cognitive Machine Learning, ENS/INRIA/PSL, PSL Research University, Paris, France
[6] Cognitive Machine Learning, ENS/CNRS/EHESS/INRIA/PSL, PSL Research University, Paris, France
[7] École de Technologie Supérieure, Université du Québec, Montreal, Canada
[8] NEC Corporation, Japan, [9] Rice University, Houston, USA, [10] University of California, San Diego
leibny@gmail.com

Abstract

This paper presents the problems and solutions addressed at the JSALT workshop when using a single microphone for speaker detection in adverse scenarios. The main focus was to tackle a wide range of conditions that go from meetings to wild speech. We describe the research threads we explored and a set of modules that was successful for these scenarios. The ultimate goal was to explore speaker detection; but our first finding was that an effective diarization improves detection, and not having a diarization stage impoverishes the performance. All the different configurations of our research agree on this fact and follow a main backbone that includes diarization as a previous stage. With this backbone, we analyzed the following problems: voice activity detection, how to deal with noisy signals, domain mismatch, how to improve the clustering; and the overall impact of previous stages in the final speaker detection. In this paper, we show partial results for speaker diarizarion to have a better understanding of the problem and we present the final results for speaker detection.

1. Introduction

For the past years, speaker recognition research has mainly focused on telephone and close-talk microphone applications with high speech quality levels. However, far-field recordings have recently emerged as an area of interest. This is due to new applications like video annotation; home assistant devices which need to distinguish between family members; and wearables that document our everyday life. These applications provide a massive amount of data which requires automatic means of analysis. Furthermore, such devices are often used in very challenging environments with multiple speakers, and where the audio is affected by noise and reverberation. Most devices use a single microphone and, therefore, multichannel signal processing techniques (e.g., beamforming) cannot be applied to alleviate the impact of the real-life conditions. As reference, speaker detection error rates of reverberant speech are 2 times worse than close-talk speech in voices dataset [1]; internet videos with noise and multi-speaker error multiplied by 6 in recent NIST SRE18 w.r.t. clean videos [2]; and systematic evaluation of diarization (Dihard I and Dihard II [3, 4]) in real-life domain shows diarization error rates above 60% when all the stages are automatized [5].

Knowing that speaker diarization and detection require further research, the aim of our workshop was to investigate, develop, and benchmark speaker diarization and speaker recognition systems on far-field speech using single microphones in realistic scenarios that include background noises such as a television audio, music, or other people talking. The key aspects that we found were fundamental to investigate are: voice activity detection, speech enhancement, domain adaptation and improving clustering. Each part will be explained briefly in the following paragraphs.

2. Research threads

2.1. Initial system: baseline

Our baseline follows a traditional diarization workflow connected to a speaker detection branch (see Figure 1, solid line rectangles). For diarization (who spoke when in a recording), the first stage is a voice activity detection (VAD), followed by an acoustic feature extractor and an embedding extractor. The embedding extractor uses a sliding window to produce a sequence of speaker embeddings. The clustering block intends to group the sequence of embeddings into single speaker clusters. A speaker label is created for each cluster. Thus, the diarization output is a sequence of speaker labels with its corresponding time marks. The speaker detection stage performs a verification task, *i.e.*, decide whether a known speaker is in the recording or not. We are interested in a scenario where multiple speakers can be present in the test recording; therefore, we apply diarization as a first step. The diarization output is used to compute a speaker embedding for each of the speakers identified in the diarization stage. Then each test speaker embedding is compared against the enrollment embedding. To compare two embeddings we usually employ probabilistic linear discriminant analysis (PLDA), that outputs scores. When we encounter a mismatch condition those scores should need calibration. The

10.21437/Odyssey.2020-59

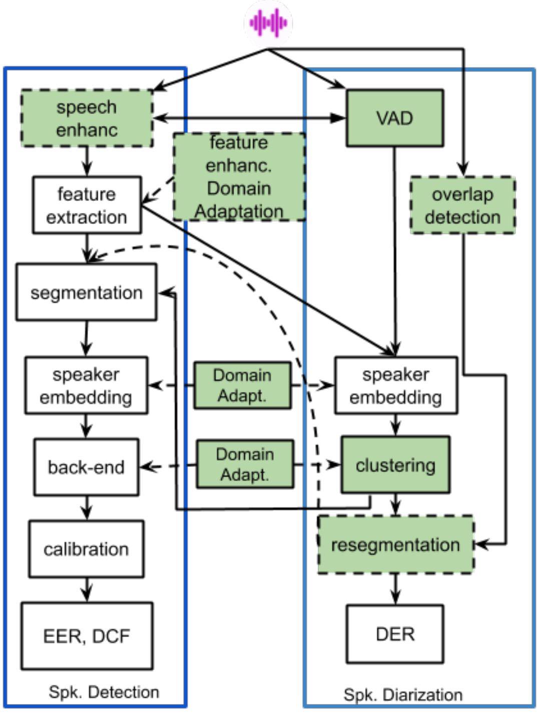

Figure 1: Speaker detection pipeline. □ is the baseline system, ⬚ are the new parts that are included in the pipeline, shaded-□ are the modules that were investigated.

metric employed for diarization is diarization error rate (DER), which takes into account the false alarms, the misses (classifying speech as non-speech) and speaker confusion (speaker mismatch between hypothesis and reference). The speaker detection metrics are the EER, minimum and actual DCF.

To explore in detail the challenging conditions in far-field, we analyzed the shaded rectangles blocks in Figure 1.

2.2. Voice Activity Detection and Speaker Separation

Improving speaker separation and voice activity detection addresses some of the key issues we had identified as being particularly problematic for our task. Notably the fact that these audio recordings contain noise, overlapping speech, and more diverse speakers than in more run-of-the-mill datasets makes the task more challenging. In this area, we explored three main routes.

Knowing that one portion of our dataset includes children. The first route focuses on separating child and adult speech in realistic conditions [6] involving background noises, reverberations and overlapping speech. First, by measuring the speech dissimilarities between children-and-babies and adults using i-vectors, we demonstrate that distances between child-and-baby and adult speech are large enough to warrant a possible separation through establishing child-and-baby and adult speech groups. Motivated by this, we conduct a joint framework of speech enhancement and speech separation for child speech signal processing. Our model is based on progressive learning to extract child speech from simulated mixtures during training. The data from toddlers between 2 and 5 years is limited. Based on this constraint, a progressive learning framework generates intermediate target outputs and stacks them together with the original limited-sized mixed input feature vectors to increase the training samples. To boost the speech separation we employed Multiple-target learning.

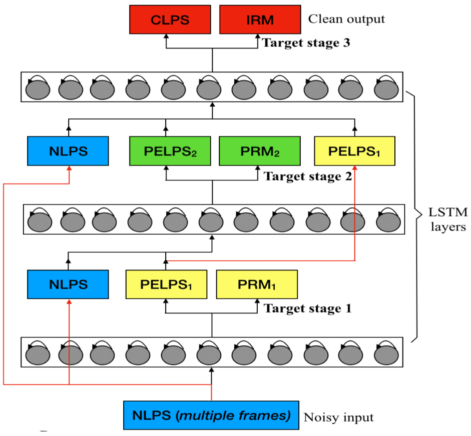

Figure 2: Multi-Target Learning based Speech Enhancement Architecture

The second study addresses the *overlap detection* [7] as a sequence labeling task. We hypothesized that detecting regions of overlapped speech (two or more speakers at the same time) is most effectively solved from a temporal perspective. Hence, our system relies on two stacked bi-LSTM layers, two feed-forward layers, and a final classification layer, fed into binary cross-entropy loss. We generate shorter sub-sequences of speech from a training set to form mini-batches. Half of the training sub-sequences were computed by a weighted sum of two random sub-sequences, simulating the overlap. The input can be MFCCs or Sincnet features following the configuration described in [8]. Both features showed competitive results, but in cases where the overlap is higher the waveform end-to-end overlap detector (*i.e.* detector that includes the Sincnet) shows better performance (see Figure 3 for details on the architecture).

The third approach combines the VAD and speaker type classification in a multi-task learning set-up. We trained a system to diarize five classes: key child (wearing a recorder), other child, male adult, and female adult and speech, all of which could be active at the same time. Silence is marked by the absence of activation in the classes. We explored diverse features (MFCC + pitch, spectrogram, or raw waveform) and architectures. The best architecture consisted of learning from the raw waveform with the SincNet model followed by a LSTM and a fully connected layer.

2.3. Dealing with noisy signals

We explored two main ideas to overcome the noisy signals in our different scenarios.

In the first approach, we built a SNR-progressive multi-target learning based *speech enhancement* model for adverse acoustic environments. The progressive multi-targets (PMT) network is divided into successively stacking blocks with one LSTM layer and one fully connected layer via multi-targets learning per block (see Figure 2). The fully connected layer in every block (target layer) is designed to learn intermediate speech targets with a higher SNR than the targets of previous target layers. A series of progressive ratio masks (PRM) are concatenated with the progressively enhanced log-power spectra (PELPS) features together as the learning targets. For our

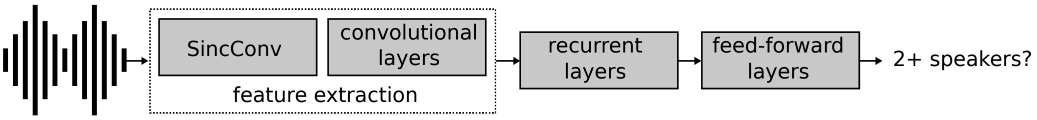

Figure 3: Architecture of the neural network used for end-to-end overlapped speech detection.

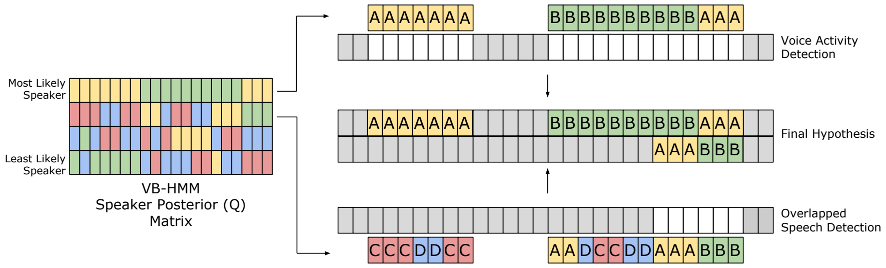

Figure 4: Speaker assignation process given matrix **Q** and the overlap detector output.

experiments and to enhance the generalization ability of speech enhancement model, we built a 1000-hour training set. At test time, we directly feed the enhanced audios processed by our enhancement model to the back-end systems, including speaker diarization and speaker detection. More details about this approach are described in [9].

The second approach we explored is *feature enhancement* with deep feature losses [10]. The main idea is to train a feature-domain enhancer which can serve as a pre-processing module to the x-vector system during inference. We develop on the ideas of perceptual loss [11] and speech denoising work [12]. This approach requires a pre-trained auxiliary network for loss estimation between enhanced and clean samples. For the auxiliary network, we chose an x-vector network based on ResNet-34 with LDE (Learnable dictionary encoding) pooling and trained for speaker classification using an Angular Softmax loss objective [13]. For the enhancement network, we design networks based on [14] (Encoder-Decoder residual network) and [15] (Context Aggregation Network). We obtained significant improvement on real datasets using auxiliary x-vector network trained on clean speech. Using the augmented x-vector auxiliary network, we observed slight improvements on simulated noisy sets.

2.4. Domain Mismatch

We explored two ways to deal with domain mismatch between training and test data: domain adaptation in acoustic feature space and domain adversarial training for VAD training.

One way to mitigate the effect of the noisy signals is to perform domain adaptation. In our first case, we examined how to train an unsupervised speech enhancement system, which can be used as a front-end pre-processing module to improve the quality of the features before they are forward passed through the x-vector network. The details of the procedure can be found in [16]. Simply put, the unsupervised adaptation system is based on CycleGAN [17, 14]. We trained a CycleGAN network using log-Mel filterbank features as input to each of the generator network. During testing, we process the far-field test data through the *reverb to clean generator* of CycleGAN. These enhanced acoustic features are then used to extract x-vectors; which, in turn, are used for PLDA scoring. Though CycleGAN network was trained for doing de-reverberation task, we also tested it on noisy datasets to investigate its generalization abiliy to unseen test conditions. We observed improvements on both reverberant and noisy test datasets.

To mitigate the domain variability, we also proposed a domain-adversarial training for robust end-to-end VAD [18]. Two-seconds audio chunks are passed to a Sincnet to compute meaningful filters that are the input for a VAD and a domain classifier. On one branch we stack two bi-LSTM layers followed by three feed-forward building an end-to-end VAD. On a second branch we conduct an domain-adversarial multitask training to distinguish among domains. The Sincnet filters are forced to be domain independent using one uni-directional LSTM, followed by max-pooling along the time axis and one feed-forward layer. The output of this branch is the probability distribution over the domains. It is possible to use both branches separately, or connect both parts using gradient reversal. The latter will provide feedback to the sincnet and force the network to compute independent features.

2.5. Improving clustering

Knowing that *clustering* is a crucial stage in diarization, we examined a set of ways to address the problem.

A first solution was to turn the problem into a supervised learning task so that it considers the temporal information and optimization of the DER as the main goal. The model input is the set of embeddings from a recording with multiple speakers. This is an encoder-decoder model, where the encoder converts the sequence of embeddings into a context vector **c**. This vector contains information about the dialogue, like the number of clusters (speakers), cluster centroids, etc. The decoder compares embeddings from the input with a context vector and assigns a cluster. Both encoder and decoder used a bi-directional multi-layer RNN to deal with sequentiality.

A second approach was to refine the clustering by combining the *resegmentation with overlap detection* (see Section 2.2) [7]. Once we get the speaker segments from the clustering, it is recommended to perform a resegmentation stage that provides a refinement of the speaker boundaries. The state-of-the-art method for the refinement is variational Bayes (VB) HMM (Hidden Markov Model) resegmentation, studied in [19] and applied

	AMI				BabyTrain				CHiME5			
	DER	FA	Miss	Conf	DER	FA	Miss	Conf	DER	FA	Miss	Conf
Baseline	49.25	4.46	35.41	9.38	85.38	46.67	**8.03**	30.68	69.22	38.36	11.45	**19.41**
+ E2E VAD	36.41	2.98	21.04	12.39	50.63	**8.88**	14.82	26.93	70.68	32.02	11.22	27.44
+ Enhanc	36.17	2.98	21.04	12.15	49.67	8.95	14.82	25.9	66.13	32.02	11.22	22.89
+ VB reseg	34.86	**2.97**	21.06	**10.83**	48.49	8.9	14.88	24.71	63.03	32.02	**11.21**	19.8
+ Overlap Assign	**30.76**	3.73	**13.64**	13.39	**47.49**	8.9	14.88	**23.71**	**58.59**	**23.08**	13.15	22.36

Table 1: DER % for different subdatasets and modules.

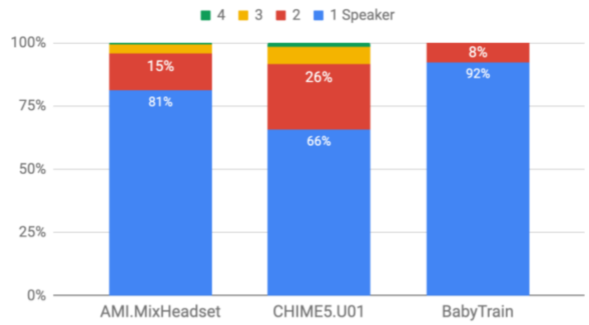

Figure 5: Percentage of number of speakers in voiced regions.

to different scenarios [5]. VB-HMM resegmentation computes a per-frame speaker posterior matrix **Q**. Meanwhile, overlap detection provides the regions with two or more speakers talking. We hypothesized two speaker labels in those regions, *i.e.*, take the two speaker with higher posterior in matrix **Q** and continue the diarization process. The DER was reduced considerably for all the cases by including the overlap assignment. The speaker assignation process is described in [7] and depicted in Figure 4.

2.6. Speaker detection

We explore how to overcome the main sources of confusion in speaker detection: overlap speech and the multiple speakers in the same audio stream. Firstly, we compare two different techniques to create homogeneous segments, computing the diarization as input to speaker detection and the sliding-window without knowledge of the speaker labels. Secondly, we used the overlap regions (explained in Section 2.2) to remove them from the speaker detection pipeline. The main conclusion of our experiments is that diarization is still necessary for these type of environments, where the speech of the speaker-of-interest is highly degraded by other sources of audio. While there is a slight improvement in matching scenarios, the speaker detector degrades dramatically when the overlap detector is not trained within the same domain showing a clear problem with the purification in mismatch data.

3. Experimental setup

3.1. Datasets

To study the main issues of the far-field scenario we built a dataset that is summarized by the following four corpora:

- **Meeting** (***AMI*** [20]): with a setting of 3 different meeting rooms with 4 individual headset Microphones, 8 Multiple Distant Microphones forming a microphone array; 180 speakers x 3.5 sessions per speaker (sps); suitable for diarization and detection. Since we are exploring single microphones, we focused only on the mix Headset scenario; 98 hours
- **Indoor controlled** (***SRI data [21]*** [1]): with a setting of 23 different microphones placed throughout 4 different rooms; controlled backgrounds, 30 speakers x 2 sessions and 40 hours, live speech along with background noises (TV, radio); suitable for detection (only reliable labeling of target speaker was provided).
- **Indoor not controlled** (***CHiME5*** *[22]*): with a setting of kitchen, dining, living room, 80 speakers, 50 hours; 4 speakers in two-hour recordings; 32 microphones per session; suitable for diarization only (there are not enough impostor speakers in different sessions within the corpus)
- **Wild** (***BabyTrain***[2]): with an uncontrolled setting, 450 recurrent speakers, up to 40 sps (longitudinal), 225 hours; suitable for diarization and detection.

To have a further analysis on the overlap Figure 5 shows the percentage of number of speakers in voiced regions. Note that CHiME5 data contains around 35% of overlap, AMI 29% and BabyTrain 8%. SRI was not included in this analysis since the annotations for the non-target speakers were not available.

For *speaker detection* the enrollments were generated by accumulating non-overlapping speech (5, 15 and 30s duration) of every target speaker along one or multiple utterances. For the test, we cut the audio into 60 second chunks. We do a Cartesian product between the enrollments and the test segments to generate all possible trials. Then based on conditions, some trials are filtered out. For example, same session and same microphones are not allowed to produce a target-trial pair.

3.2. System configuration

Our experimental setup is depicted in Figure 1. The dotted blocks are the new approaches added to the pipeline. The underlined methods along the text show the final combination of modules that obtained effective results and that are part of the final code contribution in [24]. The definite pipeline follows:

1. **Speech enhancement(Diar/SpkD)**: we used a 1000-hour training set. The noisy mixtures are made at three SNR levels (-5dB, 0dB and 5dB), and the progressive increasing SNR between two adjacent targets is set to 10 dB. The audios are sampled at 16 kHz rate and the frame length is 256 samples (PELPS and the PRM are 257 dimension). During the testing stage, we directly feed the enhanced audios processed by our enhancement model to back-end systems.
2. **Acoustic features(Diar/SpkD)**: 23 dimension MFCC for x-vector systems based on Kaldi TDNN x-vectors;

[1] This data was recorded by SRI international and was submitted to LDC for publication

[2] This data uses daylong recordings from Homebank [23], expected to be public

and 23 log-Mel filter bank features for ResNet based x-vectors. Features were short-time centered before silence removal with a 3 seconds sliding window.

3. **Feature enhancement(Diar/SpkD)**: We use a an SNR estimation algorithm (WADASNR [25]) to select the 50% highest SNR signals of VoxCeleb [26] as clean data. Noisy data is formed by contamination with samples from MUSAN [27], CHiME-3 [28], and DEMAND [29] noises. For the auxiliary network, we choose ResNet-34 x-vector network to filter VoxCeleb set. For the enhancement network, we use a ConvGenNet and a network we design based on CAN.
4. **Voice activity detection(Diar/SPkD)**: Trained on 2s audio chunks for each corpora, with trainable SincNet features (using the configuration [14]). Features are fed into BiLSTM network with binary output (speech/non-speech).
5. **Overlap detection(Diar)**: Follows the same architecture as the VAD where output classes are single/multiple speakers.
6. **Embedding extraction(Diar/SpkD)**: We used an extended TDNN architecture (E-TDNN) presented in [30]. This network interleaves fully connected layers in between the convolutional layers. It comprises 4 blocks of 1D dilated convolutions plus affine layers. The first layer uses a kernel of size 5 and the other ones use a kernel of size 3. The dilation factors are 1, 2,3 and 4 respectively. The embedding is extracted from the output of an affine layer that follows the statistics (mean + standard deviation) pooling layer.
7. **Clustering(Diar)**: the system employs an Agglomerative Hierarchical Clustering (AHC) to compute the speaker labels in a recording. PLDA models were trained on VoxCeleb [26] and adapted to each domain using a small amount of in-domain training data.
8. **Resegmentation and overlap assignment (Diar)**:We first perform resegmentation using HMM-VB resegmentation module. We used the labels from the clustering and the 400 dimensional i-vectors. Feature vectors for the module are length-60 MFCCs with deltas and double deltas, extracted in 10ms steps with a25ms window. We use a single VB inference iteration in accordance with [17]. The most likely speaker is assigned to frames detected as speech by the voice activity detector. It can therefore be used as-is by taking the most likely speaker for each voiced frame and aligning them according to a VAD binary vector. A second most likely speaker is only assigned for frames detected as overlapped speech.
9. **Traditional speaker detection pipeline**: Includes the speech/feature enhancement, the automatic segmentation obtained from the diarization stage, the shared embedding extractor and PLDA (both with augmentation), and a calibration stage. The Data augmentation was performed by adding noise and reverberation.

We used the same pipeline for our complete dataset without distinction; meaning that all four corpora were treated the same and employed the same models (except for the PLDA). We employed Kaldi [31] as our primary toolkit to develop our pipeline. For specific parts, such as VAD and overlap detection we used pyannote [32]. For speech and feature enhancement we implemented the algorithms in a combination of PyTorch and Kaldi.

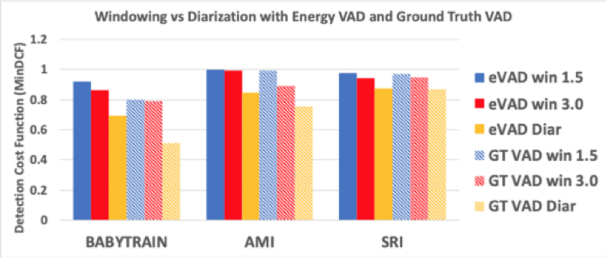

Figure 6: Diarization VS Windowing results for Babytrain, AMI and SRI databases using Minimum Detection Cost Function.

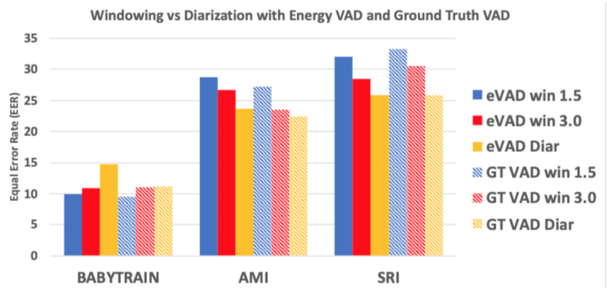

Figure 7: Diarization VS Windowing results for Babytrain, AMI and SRI databases using EER.

4. Results

We built a robust system that can be useful for the research community[3]. We followed the same pipeline for the corpora, meaning that we achieved some degree of generalization in the models for the VAD and overlap detection, speech and feature enhancement, and the embedding training. The PLDAs are treated separately and are corpus dependant. Diarization provides the first evidence for an effective speaker detection. Table 1 shows the DER by adding a module starting from the single baseline. We can observe improvements for the three sub-datasets. We emphasize the DER relative improvement of 37%, 44% and 15% for AMI, BabyTrain and CHiME5 respectively. One of our main findings was that an enhancement phase, either speech or feature, is necessary when dealing with adverse scenarios. The VB-HMM resegmentation gives some improvement, but the improvement is increased when combined with the overlap detector.

One of the first tasks was to compare the diarization with a simple sliding-window process that chunks the signal in segments of 1.5s or 3.0s that can be considered homogeneous for the speaker detection process. Figure 6 and 7 show the results of the diarization versus the sliding-window strategy on three different databases: Babytrain, AMI and SRI applied over the groundtruth VAD and a VAD based on energy. As we can see from the figures, the results are consistent over all the corpus: while increasing the size of the window improves the performance of the speaker detection system, the diarization is still better delimiting and clustering the segments of the same speaker.

Figure 8, 9 and 10 present the results for speaker detection considering diarization. For the purpose of this paper, we selected a combination of 30s enrollments and above 5s tri-

[3] Some of the research threads are on-going but with good perspectives of further improvements.

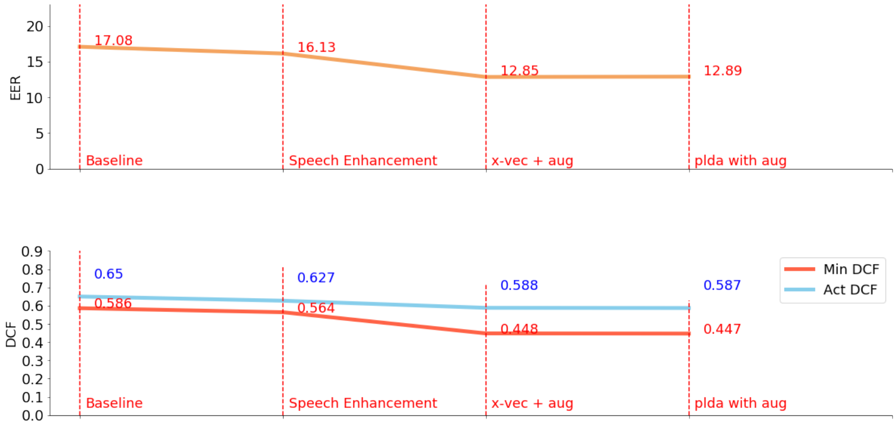

Figure 8: AMI results for Speaker Detection

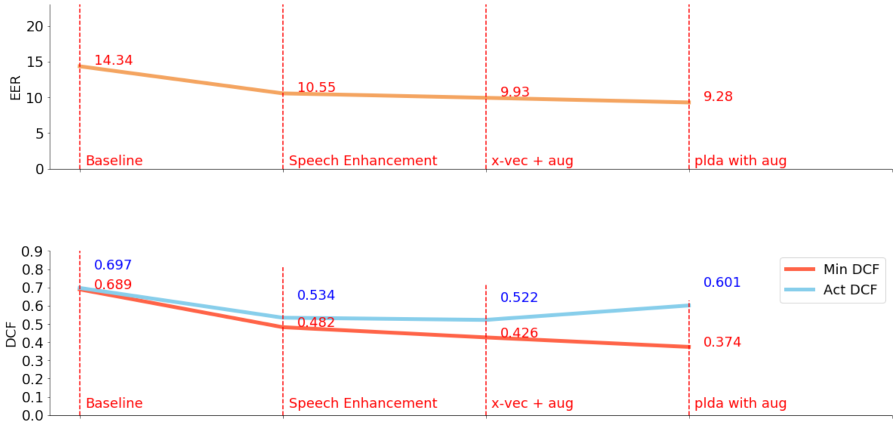

Figure 9: BabyTrain results for Speaker Detection

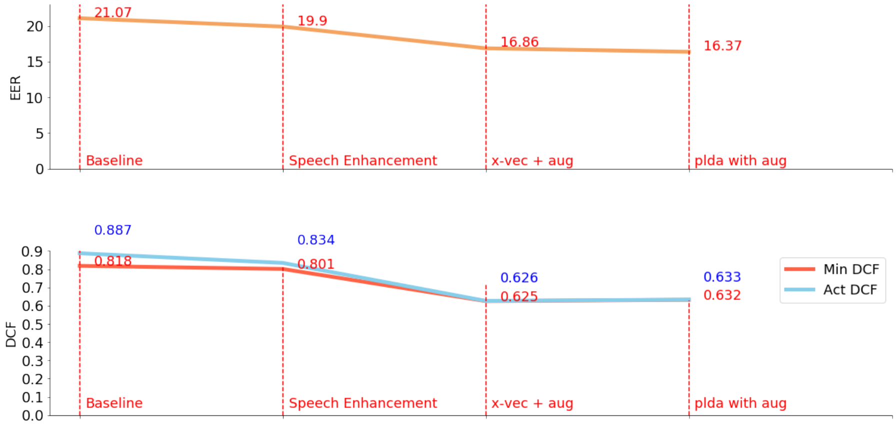

Figure 10: SRI results for Speaker Detection

als. Other combinations are also possible, but they follow the same trend. An automatic segmentation, *i.e.,* an automatic diarization, was performed previously to label the speakers in the recording. We observe the EER relative improvement of 24%, 35% and 22% for AMI, BabyTrain and SRI data. The enhancement provided improvements on the three scenarios. To improve the robustness of the system we used augmentation to train the embedding (x-vector) and for the back-end (PLDA) [4].

5. Conclusions

In this paper we presented our contribution to the JSALT Workshop 2019. We showed which aspects of the pipeline are most influential in challenging scenarios. The very first finding was that speaker detection depends on a previous diarization stage to obtain successful results. Hence, we showed the necessary elements that have to be competitive on their own in the diarization pipeline: VAD, overlap detection, speech/feature enhancement, embedding extractor, clustering with resegmentation, and overlap assignment based on resegmentation. Once we tune this diarization stage, the output can be connected to the detection phase, which is basically a speaker verification system. The speech enhancement, the embedding and PLDA augmentation gave the most improvements. There are open research threads that will require further study such as: customization of speech enhancement for a dataset, exploration of other architectures for feature enhancement, how to handle domain mismatch, unsupervised adaptation in the clustering, and how to highlight the speaker of interest in the detection, among others. As we have shown, the diarization/detection problem is far from being solved in adverse scenarios.

6. Acknowledgements

The research reported here was conducted at the 2019 Frederick Jelinek Memorial Summer Workshop on Speech and Language Technologies, hosted at L'École de Technologie Supérieure (Montreal, Canada) and sponsored by Johns Hopkins University with unrestricted gifts from Amazon, Facebook, Google, and Microsoft.

7. References

[1] Colleen Richey, Maria A Barrios, Zeb Armstrong, Chris Bartels, Horacio Franco, Martin Graciarena, Aaron Lawson, Mahesh Kumar Nandwana, Allen Stauffer, Julien van Hout, et al., "Voices obscured in complex environmental settings (voices) corpus," *arXiv preprint arXiv:1804.05053*, 2018.

[2] Seyed Omid Sadjadi, Craig S Greenberg, Elliot Singer, Douglas A Reynolds, Lisa P Mason, and Jaime Hernandez-Cordero, "The 2018 NIST speaker recognition evaluation.," in *Interspeech*, 2019.

[3] Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman, "Second DIHARD challenge evaluation plan," *Linguistic Data Consortium, Tech. Rep*, 2019.

[4] Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman, "First DIHARD challenge evaluation plan," 2018.

[5] Gregory Sell, David Snyder, Alan McCree, Daniel Garcia-Romero, Jesús Villalba, Matthew Maciejewski, Vimal Manohar, Najim Dehak, Daniel Povey, Shinji Watanabe, et al., "Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge.," in *Interspeech*, 2018, pp. 2808–2812.

[6] Xin Wang, Jun Du, Alejandrina Cristia, Lei Sun, and Chin-Hui Lee, "A study of child speech extraction using joint speech enhancement and separation in realistic conditions," in *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.

[7] Latané Bullock, Hervé Bredin, and Paola Garcia-Perera, "Overlap-aware resegmentation for speaker diarization," in *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.

[8] Mirco Ravanelli and Yoshua Bengio, "Speaker recognition from raw waveform with sincnet," in *Proc. SLT 2018*, 2018.

[9] Lei Sun, Jun Du, Xueyang Zhang, Tian Gao, Xin Fang, and Chin-Hui Lee, "Progressive multi-target network based speech enhancement with snr-preselection for robust speaker diarization," in *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.

[10] Saurabh Kataria, Phani Sankar Nidadavolu, Jesús Villalba, Nanxin Chen, Paola Garcia-Perera, and Najim Dehak, "Feature enhancement with deep feature losses for speaker verification," in *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[12] Francois G Germain, Qifeng Chen, and Vladlen Koltun, "Speech denoising with deep feature losses," *arXiv preprint arXiv:1806.10522*, 2018.

[13] Jesús Villalba, Nanxin Chen, David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Jonas Borgstrom, Fred Richardson, Suwon Shon, François Grondin, et al., "The jhu-mit system description for nist sre18," *Johns Hopkins University, Baltimore, MD, Tech. Rep*, 2018.

[14] Phani Sankar Nidadavolu, Jesús Villalba, and Najim Dehak, "Cycle-gans for domain adaptation of acoustic features for speaker recognition," in *ICASSP 2019, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6206–6210.

[15] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[16] Phani Sankar Nidadavolu, Saurabh Kataria, Jesús Villalba Villalba, Paola Garcia-Perera, and Najim Dehak, "Unsupervised feature enhancement for speaker verification," in *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.

[4] Note that we are not providing detection results on CHiME5 since the dataset is quite small (80 speakers).

[17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[18] Marvin Lavechin, Marie-Philippe Gill, Ruben Bousbib, Hervé Bredin, and Paola Garcia-Perera, “End-to-end domain-adversarial voice activity detection,” *arXiv preprint:arXiv:1910.10655*, 2019.

[19] Mireia Diez, Lukás Burget, and Pavel Matejka, “Speaker diarization based on bayesian hmm with eigenvoice priors.,” in *Odyssey*, 2018, pp. 147–154.

[20] Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, Vasilis Karaiskos, et al., “The ami meeting corpus,” in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, 2005, vol. 88, p. 100.

[21] Diego Castán et al., “Ldc2019e60, distant microphone conversational speech in noisy environments,” Private communication in support of the 2019 JHU/CLSP Summer Workshop, 2019.

[22] Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal, “The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines,” *Interspeech 2018 - 19th Annual Conference of the International Speech Communication Association*, 2018.

[23] Mark VanDam, Anne S Warlaumont, Elika Bergelson, Alejandrina Cristia, Melanie Soderstrom, Paul De Palma, and Brian MacWhinney, “Homebank: An online repository of daylong child-centered audio recordings,” in *Seminars in speech and language*. Thieme Medical Publishers, 2016, vol. 37, pp. 128–142.

[24] JSALT Speaker Detection in Adverse Scenarios with Single microphone contributors, “jsalt2019-diadet,” `https://github.com/jsalt2019-diadet/jsalt2019-diadet`, 2019, [Online].

[25] Chanwoo Kim and Richard M Stern, “Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis,” in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[26] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, “Voxceleb: a large-scale speaker identification dataset,” *arXiv preprint arXiv:1706.08612*, 2017.

[27] David Snyder, Guoguo Chen, and Daniel Povey, “Musan: A music, speech, and noise corpus,” *arXiv preprint arXiv:1510.08484*, 2015.

[28] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, “The third chime speech separation and recognition challenge: Dataset, task and baselines,” in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 504–511.

[29] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent, “The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings,” *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.

[30] David Snyder, Daniel Garcia-Romero, Gregory Sell, Alan McCree, Daniel Povey, and Sanjeev Khudanpur, “Speaker recognition for multi-speaker conversations using x-vectors,” in *ICASSP 2019, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5796–5800.

[31] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., “The kaldi speech recognition toolkit,” Tech. Rep., IEEE Signal Processing Society, 2011.

[32] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill, “pyannote.audio: neural building blocks for speaker diarization,” in *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.