CRIM's Speaker Diarization System for the DIHARD Diarization Challenge

Vishwa Gupta, Jahangir Alam

Centre de recherche informatique de Montréal (CRIM)

{Vishwa.Gupta, Jahangir.Alam}@crim.ca

Abstract

CRIM is taking part in both the diarization from scratch and from gold speech segments for the first DIHARD speech diarization challenge. For diarization, we used our diarization system developed internally. We did not have time to use diarization systems readily available over the internet.

For track2, our internal diarization system uses a deep neural net (DNN) for voice activity detection trained from MGB-3 challenge data. The output of voice activity detector is then used as ground truth to adapt the voice activity detector to the DIHARD challenge data. The adapted VAD is then used for final voice activity detection. The speech segments from the voice activity detector are then segmented into homogeneous segments. These segments are then clustered using BIC clustering, followed by GMM-based clustering as outlined in [1]. Preliminary results show that we get 33.8% DER on the eval set with the gold speech segments, and 52.4% with diarization from scratch.

Index Terms: Deep Neural Networks, DNN, voice activity detection, speaker diarization.

1. Data Resources

For the DIHARD challenge, we trained the voice activity detector (VAD) with acoustic training data from the MGB-3 challenge English data for 2017/2018 challenge [2]. This acoustic training data provided by MGB challenge committee contains lightly supervised alignments based on the transcripts from closed captioning. As a measure of confidence, they also computed phone matched error rates (PMER) and word matched error rates (WMER) [3]. The total acoustic data available is 500 hours of audio. These BBC broadcast audio contains shows from many different genres.

For training the male/female GMMs for diarization, we used the 114 files from 1997 Hub4 English broadcast news (LDC98S71) training data (97 hours in total). These audio files were chosen because they have been well segmented into speaker turns.

2. Detailed Description of the Algorithm

2.1. Overview

The diarization algorithm is similar to the one outlined in [1]. In that system, we divided the audio into speech, noise and music segments by training GMMs for the corresponding sounds. Here, we replaced it by a DNN-based voice activity detector (VAD) trained on roughly 140 hours of audio. The rest of the algorithm is similar to that described in [1]:

The audio segments labeled as speech (from VAD) are then divided into homogeneous segments using a change point detector. The acoustic change point detection step (CPD) uses a symmetric Kullback-Leibler (KL2) metric, and a 13-dimensional feature vector (12 MFCCs + energy) with diagonal covariance matrix [4]. This is followed by an iterative Viterbi re-segmentation stage that models each segment by its mean and variance and finds the optimal boundaries between segments. The resulting segments are clustered using BIC agglomerative clustering that uses a 13-dimensional feature vector (12 MFCCs + energy) with full covariance matrix [5]. In this step, the clustering threshold is set so as to under-cluster the segments. The Viterbi re-segmentation and BIC-clustering steps are iterated twice. The next stage is gender determination, which labels each cluster from the previous step as male or female using male/female GMMs trained using the Hub4 acoustic data. The next step is separate male/female speaker identification-style (SID) clustering that uses more complex models of the clusters for final clustering. For this step, we use Gaussianized MFCCs with cepstral mean subtraction, and separate male/female GMMs generated from the Hub4 data. We did not adapt the GMMs to either the DIHARD dev set or the eval set for this GMM based agglomerative clustering.

2.2. Voice Activity Detection

Cambridge University had successfully used DNN-based VAD for MGB-1 challenge [6]. To reduce VAD errors (false alarms + missed speech), we tried two different architectures for neural net based VAD: DNN architecture similar to that used in [6] with varying number of input frames, and a bidirectional LSTM with 1 to 3 levels. We also tried two different feature parameters: 40-dim MFCCs, and 40-dim MFCCs with senone posteriors added to them. The senone posteriors were generated from a bidirectional LSTM with 178 senones as outputs. To train these VAD DNN models, we tried different training sets. In the first training set, we aligned all the speech segments with zero PMER. The segments aligned to words were labeled as speech and the rest as non-speech. This resulted in 20 million speech frames and only 2 million non-speech frames. The resulting 3-level LSTM gave poor results on MGB-3 dev set due to many music and noise segments being recognized as speech. So we added many more non-speech frames for training in order to balance the speech/non-speech discrimination.

We noticed in the MGB-3 training data (with lightly aligned supervision) that intervals between speech segments with closed captioning were mostly silence or music. So we added all such segments as non-speech. Including all these frames increased the non-speech frames to 31 million frames, 1.5 times the number of speech frames. DNN trained from 20 million speech, 31 million non-speech frames gave good results on the MGB-3 dev set. Also, DNNs gave lower VAD error than LSTMs. The best DNN has 81-frames of MFCC features as input, 5 hidden layers, with 2000, 500, 500, 500, and 200 output nodes respectively. The softmax layer has 2 outputs (speech/non-speech).

Speech/non-speech detection using this DNN is as follows: We first label each frame as speech or non-speech based on DNN posterior likelihoods. Consecutive speech frames are merged into one segment. Segments with less than 0.3 sec silence in between are merged. Isolated segments less than 0.2 secs are discarded. Also, DNNs with MFCCs + senone posteriors as input gave lower VAD error for MGB-3 dev set than DNNs with MFCCs only as input. However, for the DIHARD dev set, DNNs without senone posteriors as input gave lower VAD error than DNNs with MFCCs + senone posteriors as input. So we report results with DNNs trained from MFCCs only.

Since the DNN for VAD was not trained on DIHARD data, we also adapted this DNN separately on DIHARD dev and eval sets using a small learning rate of 0.000006 and 1 epoch of training with this DIHARD data. The resulting adapted DNNs gave us a small reduction in the VAD error rate.

3. RESULTS

We participated in both track1 (diarization from gold segmentation) and track2 (diarization from scratch) in order to compare our algorithms with those of other research labs. In both diarization from gold segments and diarization from scratch, the diarization is sensitive to the threshold δ [5] used for stopping the clustering process in SID (speaker identification-style) agglomerative clustering using GMMs. So we varied this threshold in order to get an optimal value.

3.1. Diarization from Gold Segmentation

The results for the dev set for different values of δ are shown in Table 1. From the table we notice that even with gold segmentation, we get significant voice activity detection errors (11.5%). This is probably due to the overlapped speech. So looks like diarization systems should also include speech overlap detection.

Table 1: *DER for dev set using gold segmentation with varying thresholds for δ.*

δ	False Alarm	missed speech	DER
-0.25	0.4	11.1	28.7
-0.20	0.4	11.1	28.2
-0.15	0.4	11.1	28.2
0.0	0.4	11.1	29.3

Table 2 shows DER for the eval set using gold segmentation as scored by the DIHARD diarization committee. The DER for the eval set is around 5% worse than that for the dev set. The lowest DER for the eval set is 23.7% by JHU.

Table 2: *DER for eval set using gold segmentation with varying values for δ.*

δ	DER
-0.15	33.8
-0.10	33.8

3.2. Diarization from Scratch

Diarization from scratch uses a voice activity detector (VAD-DNN) to remove non-speech segments before diarization. Table 3 shows DER for the dev set with/without adaptation of the VAD DNN to the dev set. The VAD error before adaptation is 30.2%, while after adaptation it is 28.3%. So we gain 1.9% with VAD adaptation. However, after diarization, we loose all these gains because of the differences in the speaker error. Maybe we should have adapted VAD with thresholds that lead to lower FA rates.

Table 3: *DER for diarization from scratch for dev set using varying thresholds for δ.*

δ	VAD	False Alarm	missed speech	DER
-0.5	no adapt	10.7	19.5	47.1
-0.25	no adapt	10.7	19.5	45.7
0.0	no adapt	10.7	19.5	48.0
0.25	no adapt	10.7	19.5	59.8
-0.30	adapt	14.5	13.8	45.7
-0.25	adapt	14.5	13.8	45.9
-0.20	adapt	14.5	13.8	46.2

Table 4 shows the DER for eval set for diarization from scratch. Both the CRIM submissions used VAD adapted to the eval set. Compared to the dev set, the DER for the eval set is 6.7% higher. The lowest DER on eval set is 35.5% by BUT.

Table 4: *DER for eval set using diarization from scratch and VAD adapted to eval set with varying thresholds for δ.*

δ	DER
0.25	68.3
-0.10	52.4

4. Hardware Requirements

CRIM has 8 compute servers with 2 GPU's each. The operating system is linux centos 7, with cpu Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz. Each compute server has 16 CPU's, 256 Gbytes of RAM, and 2 NVIDIA TITAN X (Pascal) GPU's. The GPU's were only used for adapting the VAD.

Adaptation of the VAD DNN to dev or eval set takes approximately 31 minutes with 1 GPU. However, this adaptation is done only once for all the Dev set audio files or all the eval set audio files. The adaptation of VAD DNN was done using the Kaldi toolkit.

Voice activity detection with VAD-DNN is done once for dev or eval set. The primary computing is calculation of VAD-DNN posteriors. For the dev set, it took 2 minutes and 47 seconds (on 1 CPU) to compute these posteriors. Since there are 164 dev audio files, it takes roughly 0.65 secs per file to compute the VAD posteriors.

We only used 1 CPU for diarizing each file. Diarization of each 5 minute file took less than a minute. The software for diarization was developed at CRIM.

5. References

[1] V. Gupta, G. Boulianne, P. Kenny, P. Ouellet, and P. Dumouchel, "Speaker Diarization of French Broadcast News", Proc. ICASSP-2008, pp. 4365–4368.

[2] MGB challenge website: http://www.mgb-challenge.org/

[3] P. Bell et. al., "The MGB challenge: Evaluating multi-genre broadcast media transcription", in Proc. ASRU 2015.

[4] M. Siegler, B. Jain and R. Stern, "Automatic segmentation and clustering of broadcast news audio", Proc. DARPA Speech Recognition Workshop, Feb. 1997, pp. 97–99.

[5] C. Barras, X. Zhu, S. Meignier and J. Gauvain, “Multistage Speaker Diarization of Broadcast News”, IEEE Trans. ASLP, vol. 14, no. 5, 1505–1512, 2006.

[6] P. Woodland, X. Liu, Y. Qian, C. Zhang, M. Gales, P. Karanasou, P. Lanchantin, L. Wang, “Cambridge University Transcription Systems for the Multi-Genre Broadcast Challenge”, in Proc. ASRU 2015, pp. 639–646, Dec. 2015.