Speaker Diarization On Telephony Audio Using Deep Learning Neural Networks

Machine Learning Engineer Nanodegree Program– Capstone Project

Bernie Armour

June 4 2018

Contents

Background

Speaker diarization is the process by which an audio recording is divided into segments and each segment is assigned a label according to the speaker's identity. Therefore speaker identification answers the question of "who spoken when" in an audio recording. The actual identity of the speakers is not determined. Instead speaker diarization seeks to distinguish between each unknown speaker in the audio recording.

Applications of speaker diarization include:

- Analysis of phone conversations, meeting, legal proceedings and transcripts
- Speaker identification
- speech to text with speaker annotations

Speaker diarization is an active are of research over the last 20 years. In particular the NIST Rich Transcription Evaluations in 2003 and 2004 yielded a variety of approaches for speech detection and segmentation (see Rich Transcription Evaluation). These approaches were based on conventional algorithms based on spectral feature extraction, change detection, and feature clustering.

A significant advancement was achieved based on the application of joint factor analysis to generate an i-vector feature which compactly encompasses the differentiating information for speaker identification and speaker change detection [Kenny, 2010]. This work focussed on diarization of telephony audio.

More recently researchers have been experimenting with neural networks to achieve lower error rates in diarization. In [Cyrta, 2017], the authors developed a recurrent convolutional neural network (R-CNN) that they trained to classify short audio segments in a large training set of annotated speaker segments. The resulting classifier has activation layers which are said to contain *speaker embeddings*. These speaker embeddings are said to be sufficiently generalized such that they can be applied to other sources of audio recordings and speaker who are not part of the training set. It is similar to the networks developed to achieve high accuracy with ImageNet, but then are also useful to transfer learning to other related image classification tasks.

Speaker diarization appears to be a very difficult task that continues to attract new research work. Unfortunately, even the best solutions are still achieving 10% or higher error rates.

Problem Statement

In this project, the goal is follow the approach used in [Cyrta, 2017] to design, implement, and test a speaker diarization solution based on a deep neural network solution. Instead of working with high quality broadcast audio, regular land-line and mobile (cellular) telephone audio will be the audio source. To measure the results in experiments a Diarization Error Rate is to be defined and applied to the diarization tests.

Datasets

- Switchboard-2 Phase III Audio https://catalog.ldc.upenn.edu/LDC2002S06

  "During the collection period, the LDC collected a total of 2,728 calls, or 5,456 sides, from 640 participants (292 Male, 348 Female), under varied environmental conditions."

- Switchboard Cellular Part 2 Audio https://catalog.ldc.upenn.edu/LDC2004S07

  "During the study period, LDC collected a total of 2,020 calls, or 4,040 sides (2,950 cellular) from 419 participants (2,405 female speakers, 1,635 male speakers) under varied environmental conditions."

Preprocessing the Data

The two datasets listed above contain 2-channel audio recordings of telephone conversations. In each recording there are two persons speaking: one speaker per channel. A voice activity detector (VAD) will be used to mark the speech segments in each channel. These segments will then be used to create speaker labels: speaker A, speaker B, speaker A+B (double talk). In addition, speakers A and B are assigned PIN numbers uniquely identifying the person speaking.

The original 2-channel audio files with PIN identifiers can be directly used for training a speaker classifier. We simply use one channel at a time and use a VAD to only process the speech in the recordings.

For speaker diarization testing, we of course require that the source audio consists of only one channel containing multiple speakers. To achieve this, we will simply sum the two channels in the test audio set. The segments for speaker A, B and A+B that were generated using the 2-channel recordings is used as

the truth for the diarization outcome. We use this info to calculate the Diarization Error Rate (DER) for the test dataset.

Solution

Use the approach described in [Cyrta, 2017] to implement a R-CNN based diarizer.

The solution is outlined in detail in the Design section.

Benchmark Model

The author has previously implemented a variation of the algorithm [Barras, 2006]. This was implemented in C++. The results of this conventional algorithm will be compared against the CNN and R-CNN designs.

Evaluation Metrics

The Diarization Error Rate (DER) is defined as possible.

$DER = E_{Spk} + E_{FA} + E_{Miss}$

E_{Spk} = time assigned to incorrect speakers divided by total time

E_{FA} = amount of time incorrectly detected as speech divided by total time

E_{Miss} = amount of speech time that has not been detected as speech divided by total time

Finally, a tolerance of 250ms is allowed on speech or speaker start/stop boundaries.

A very good diarizer typically has an error range of 10 to 15%.

Design

The proposed solution is to closely follow the work done in [Cyrta, 2017], but adapted to the reduced bandwidth of telephony audio (8000 samples/s instead of 16000 samples/s). Though some experimentation may be done with the options for spectral features (Mel-freq., SFTF, etc), the Mel-frequency coefficients will be the first choice for features.

The reduced bandwidth of the telephone audio as compared with the broadcast news audio used in in [Cyrta, 2017], is that there is less speaker differentiating feature information in the telephony audio. The feature vector is at least half the size. It is likely that the network designs used in the reference will not be directly suitable for this data. Nevertheless, this is the starting point.

The majority of the effort is to be focussed in the design and refinement of the neural network architecture. In the reference [Cyrta, 2017], they experimented and produced results using CNN and -R-CNN designs.

For the R-CNN, they used Keras and the following for training the speaker classifier:

1. Input
2. 4 convolutions layers.
    a. Layer 1: 3X3 convolution, batch-normalization layer, max pooling layer 2X2 non-overlapping stride, ELU
    b. Layer 2: 3X3 convolution, batch-normalization layer, max pooling layer 3X3 non-overlapping stride, ELU
    c. Layer 3: 3X3 convolution, batch-normalization layer, max pooling layer 4X4 non-overlapping stride, ELU
    d. Layer 4: 3X3 convolution, batch-normalization layer, max pooling layer 4X4 non-overlapping stride, ELU
3. 2 recurrent layers with GRU gating
4. 1 fully connected layer with presumably softmax and C outputs (C = number of speaker classes).

The batch normalization is claimed to address an "internal covariate shift". It will be interesting to observe if this occurs in this project.

ELU (exponential linear units) is said to improve slightly over ReLU for classification accuracy. This will be tested.

The following tasks will be attempted:

1. Data preparation as described in the Datasets section.
2. Feature calculation and labeling (only speech parts of recordings used, labels are PIN value identifying speakers)
3. Network design and experimentation:
    a. Experiment with CNN designs
    b. Experiment with R-CNN (recurrent designs were not covered by the ML course so this will mainly rely on what the Keras GRU implementation offers).
    c. Attempt to determine if covariate shift is occurring and evaluate if a batch-normalization layer reduces the effect.
4. Design and test the diarization function:
    a. Freeze the neural net designed in the previous task including the final fully connected layer. This network is used to extract time-dependent speaker characteristics called "speaker embeddings" in the literature. The trained network is used to generate classifications on speakers that were *not* part of the training dataset. Variations in the classification scores are measured and thresholded to determine positions of speaker change.
    b. An important part of the diarization step is the use of an accurate voice activity detector. Only the speech parts of the recording should be fed into the diarizer. It proposed that a basic variance type VAD be used. The author has implemented VAD algorithms in the past and intends to use a conventional approach.
    c. Tuning the detector threshold is critical to the success for determining the speaker change points. It is important that normalization is used to make the threshold independent of speech segment duration, loudness, bandwidth, and background noise.
5. Produce tables evaluating error rates for various configurations.

References

[Barras, 2006]

Claude Barras, Xuan Zhu, Sylvain Meignier, Jean-Luc Gauvain. *Multi-stage speaker diarization of broadcast news*. **IEEE Transactions on Audio, Speech and Language Processing**, Institute of Electrical and Electronics Engineers, 2006, 14 (5), <10.1109/TASL.2006.878261>. <hal-01434241>

[Kenny, 2010]

P. Kenny, D. Reynolds and F. Castaldo, "Diarization of Telephone Conversations Using Factor Analysis," in **IEEE Journal of Selected Topics in Signal Processing**, vol. 4, no. 6, pp. 1059-1070, Dec. 2010.

doi: 10.1109/JSTSP.2010.2081790

[Cyrta, 2017]

Pawel Cyrta, Tomasz Trzciński, Wojciech Stokowiec. *Speaker Diarization using Deep Recurrent Convolutional Neural Networks for Speaker Embeddings*. **arXiv:1708.02840v2 [cs.SD]** 10.1007/978-3-319-67220-5_10, (https://arxiv.org/abs/1708.02840)