Speaker recognition using deep neural networks

Bajibabu Bollepalli

Department of Signal Processing and Acoustics
Aalto University

June 14, 2016

Outline

- Renaissance of neural networks
- DNNs role in speaker recognition (SR)
- Extraction of i-vectors in standard SR framework
- i-vector extraction using DNNs
- Results
- Pros and cons using DNNs in SR
- Other strategies
- References

Renaissance of neural networks – Deep Neural Networks (DNNs)

- Deep learning's successes can be explained by three factors:
  1. Advances in computer hardware and software (e.g. GPUs)
  2. Abundance of data
  3. Complex models with massive number of parameters, even if they are unidentifiable and uninterpretable
- DNN's success in ASR [1] draws attention in speech research community
  - Outperformed the Gaussian Mixture Models in acoustic modelling
  - Appearance of "deep" keyword in the proceedings of **ICASSP**– 2011: 13, 2012: 18, 2013*: 61, 2014: 102, 2015: 111, 2016: 149

DNNs role in speaker recognition (SR)

DNNs are employed in two ways in SR

1. **Direct** way [2]
   - A DNN is trained as a classifier for the intended recognition task directly to discriminate between speakers for SR
2. **Indirect** way [3]
   - A DNN is possibly trained for a different purpose to extract data that is then used to train a secondary classifier for the intended recognition task
   - Extract frame-level features or accumulate multi-modal statistics

Most of the exist studies applied DNNs in indirect way. Thus the focus of this presentation is on the same.

DNNs in indirect way

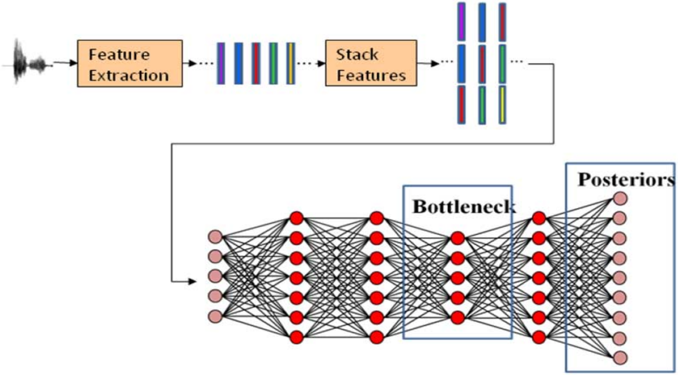

Figure : Example of DNN architecture [4].

DNNs are used to extract frame-level bottleneck and/or posterior features which further processed in later steps.

A typical speaker recognition system

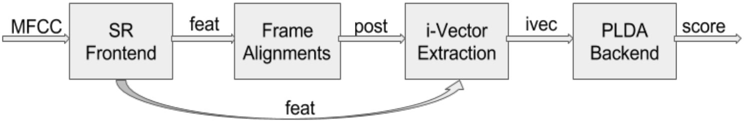

Figure : Block Diagram of a typical speaker recognition system.

- In frontend:
  - VAD; Mean-variance normalization
  - Append delta and delta-delta features
- In frame alignments:
  - Estimate posteriors for each frame based on an assumed model
  - Normally a multi-variate Gaussian Model is used
- In i-vector extraction:
  - Estimate total variability matrix with posteriors and SR features
- In backend:
  - Mean and length normalization on i-vectors
  - Compute similarity score between i-vectors by Probabilistic Linear Discriminant Analysis (PLDA)

i-vector extraction

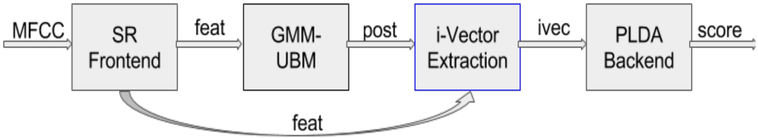

Figure : Block Diagram of GMM-UBM speaker recognition system.

- Represent a full utterance with a low dimensional vector
  - E.g., 300x60 $\rightarrow$ 200 dimensional vector
- Retain both speaker- and channel-dependent information
- Suppose $\mathbf{x}$ is a given speech utterance and it contains T feature vectors $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T)$.
  - $\mathbf{x}_t \sim \mathrm{p}(\mathbf{x},\, \theta)$
  - Following stats are needed to extract i-vector

$$\begin{aligned} \gamma_t^{(\theta)} &= p(\theta|\mathbf{x}_t) \quad \text{(frame posterior)} \\ N_x^{(\theta)} &= \sum_t \gamma_t^{(\theta)} \quad \text{(zero order statistics)} \\ F_x^{(\theta)} &= \sum_t \gamma_t^{(\theta)} \mathbf{x}_t \quad \text{(first order statistics)} \end{aligned} \tag{1}$$

Frame alignments with GMM-UBM

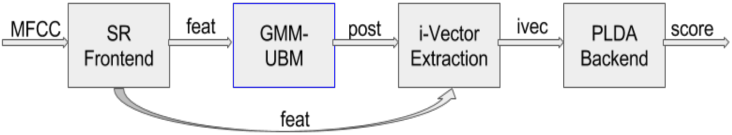

Figure : Block Diagram of GMM-UBM speaker recognition system.

GMM-UBM

- Gaussian Mixture Model - Universal Background Model
- Typically contain 1024 or 2048 Gaussians
- Trained on tens or hundreds hours of speech from a large number of speakers
- Gender dependent or independent
- EM algorithm is used to estimate the GMM parameters
- Generally defines the speaker manifold

GMM-UBM based i-vector extraction

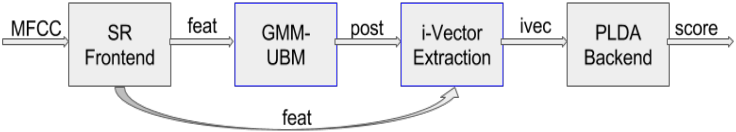

Figure : Block Diagram of GMM-UBM speaker recognition system.

- In this model

$$\mathbf{x}_t \sim \sum_{k=1}^{K} \gamma_t^{(k)} N(\boldsymbol{\mu}_k + \mathbf{T}_k \boldsymbol{\omega}, \boldsymbol{\Sigma}_k) \tag{2}$$

  - K is the number of Gaussians
  - $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are mean and covariance of k-th Gaussian in UBM
  - $\mathbf{T}_k$ is a low-rank rectangular matrix also called as total variability matrix
  - $\boldsymbol{\omega}$ is a latent-variable drawn from a standard normal distribution $N(0, I)$
- The i-vector $\phi_{\mathbf{x}}$ of the utterance $\mathbf{x}$ is the maximum a posterior (MAP) point estimate of the latent vector $\boldsymbol{\omega}$.

DNN based i-vector extraction

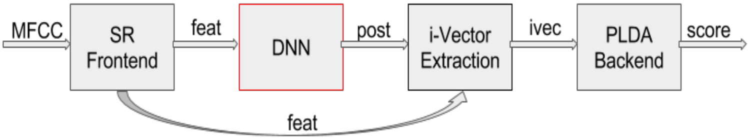

Figure : Block Diagram of DNN based speaker recognition system.

DNN

- Typically the DNN is trained for ASR task
- Inputs are ASR specific features and outputs are senones (tied triphone states)
- Need a transcribed data for training (supervision)
- No standard architecture settings

Bottleneck (BN) features based i-vector extraction

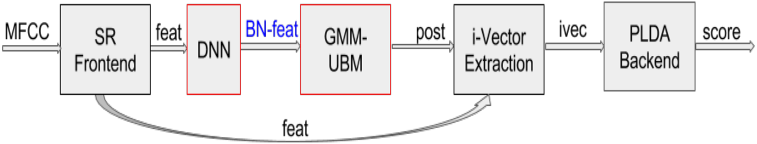

Figure : Block Diagram of BN-UBM based speaker recognition system.

- Typically BN layer has fewer nodes then input layer
- BN layer with linear activation is very much like a PCA or LDA
- Extract using the same DNN trained for ASR task
- A GMM-UBM is trained with BN features to get frame alignments
- Have the tendency to suppress the "unimportant" speaker related information
- The BN features trained with GMM can depict the phonetic space more accurately

Results

System	Database	EER (%)	Rel. Imp (%)
UBM-EM(4096)	NIST SRE'12 C2	1.81	- [3]
DNN(3450)	"	1.39	23
UBM-EM(4096)	NIST SRE'12 C5	2.55	- [3]
DNN(3450)	"	1.92	25
UBM-EM(2048)	NIST SRE'10 C5	2.91	- [5]
DNN(2227)	"	2.58	11
BN-UBM(2227)	"	2.28	22

Table : Equal error-rate (EER) comparison of gender dependent models on diffent datasets.

For my course project, I am replicating the demo shared in Kaldi "egs/sre10/v2/"

Pros and cons

Pros

- The UBM-defined classes and posteriors have no inherent meaning
- Each Gaussian in UBM may cover more than one phoneme or part of phoneme
- DNNs success in ASR i.e improvements in word error rate compared to GMMs
- DNNs trained with senones capture the speaker specific pronunciations

Cons

- Senones are dependent on language
- Need huge amount of computational resources
- The recognition performance greatly depends on the data used for training the DNN

Other strategies

- Convolutional neural networks (CNN) [6]
  - Showed better performance in noisy conditions in both ASR and SR
- Time delay neural networks (TDNN) [7]
  - It is an extended MLP architecture
  - Uses sequential information in speech
- May be recurrent neural newtorks (RNNs)?
  - No papers yet
- Can we use autoencoders to extract BN features?

References I

Geoffrey Hinton, et al.
Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups.
IEEE Signal Processing Magazine, vol. 29.6, pp. 82-97, 2012.

Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez.
Deep neural networks for small footprint text-dependent speaker verification.
ICASSP, pp. 4052-4056, 2014.

Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren.
A novel scheme for speaker recognition using a phonetically-aware deep neural network.
ICASSP, pp. 1695-1699, 2014.

References II

Fred Richardson, Douglas Reynolds, and Najim Dehak.
Deep neural network approaches to speaker and language recognition.
IEEE Signal Processing Letters, vol. 22.10, pp. 1671-1675, 2015.

Yao Tian, Meng Cai, Liang He, Jia Liu.
Investigation of bottleneck features and multilingual deep neural networks for speaker verification.
Interspeech, Dresden, Germany, pp. 1151-1155, 2015.

Mitchell McLaren, Yun Lei, Nicolas Scheffer, Luciana Ferrer
Application of convolutional neural networks to speaker recognition in noisy conditions.
Interspeech, Singapore, pp. 686-690 2014.

References III

David Snyder, Daniel Garcia-Romero and Daniel Povey.
Time delay deep neural network-based universal background models for speaker recognition.
IEEE ASRU, pp. 92-97, 2015.