Speaker Recognition (SR)

Jagabandhu Mishra

EE Department, IIT Dharwad
183081002@iitdh.ac.in

Feb, 2022

Speaker Recognition

Speaker Recognition (SR)

- SR: Task is to identify/verify/mark the speaker's identity.
- Application: Bio-metric authentication, Forensic

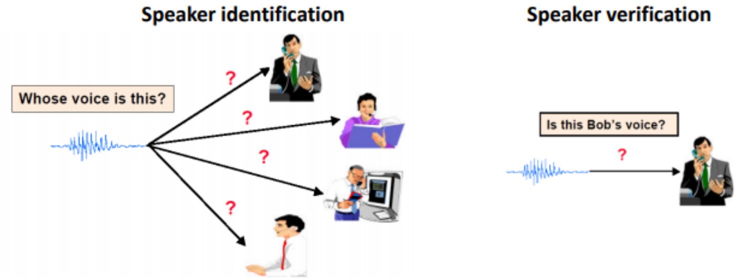

Figure: Speaker Recognition: Speaker Identification and verification, courtesy: Aalto University wiki

Types of Speaker recognition

- Speaker: Close set Vs. Open set
- Text: Text Independent Vs. Text Dependent
- Task:
  1. Speaker Identification
  2. Speaker Verification
  3. Speaker detection
  4. Speaker segmentation
  5. Speaker clustering
  6. Speaker diarization

Types of Speaker recognition

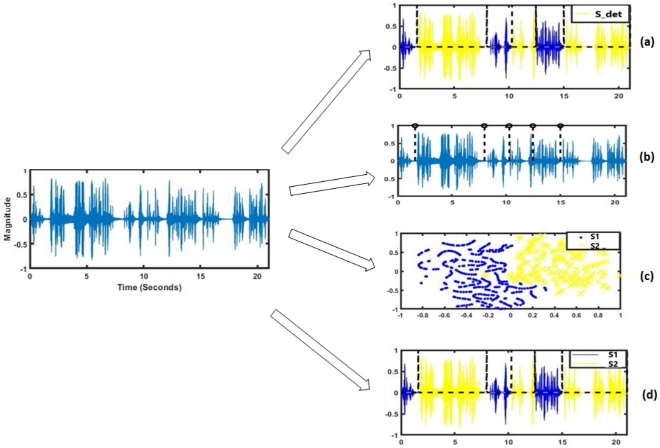

Figure: (a): Speaker detection, (b): Speaker segmentation, (c): Speaker clustering and (d) Speaker diarization.

Basic block diagram

- Text independent speaker identification/verification

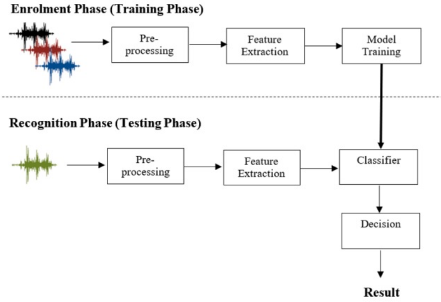

Figure: Basic block diagram of speaker recognition

Courtesy: Hanifa, R. M., Isa, K., Mohamad, S. (2021). A review on speaker recognition: Technology and challenges. Computers Electrical Engineering, 90, 107005.

Speaker Specific Features

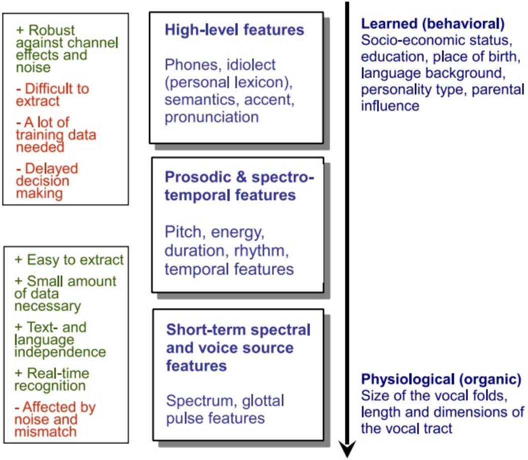

Figure: Summary of features

Courtesy: Kinnunen, T., Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. Speech communication, 52(1), 12-40.

Feature Extraction

- Short term spectral features
- Excitation source features
- Prosodic features
- High-level features

Short term spectral features

- Quasi periodic signal
- Stationary assumption: 10-30 msec
- Physiologocal structure of Speaker: pole location

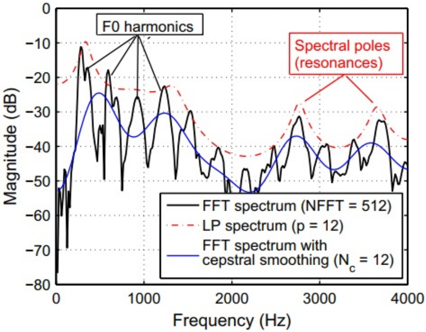

Figure: Spectral evidence

Courtesy: Kinnunen, T., Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. Speech communication, 52(1), 12-40.

Short term spectral features

- Features
  1. Mel-frequency cepstral coefficients (MFCC)
  2. Linear prediction cepstral coefficients (LPCC)
  3. Line spectral frequencies (LSF)
  4. Perceptual linear prediction (PLP)
- Most popular: MFCC

Reference: Kinnunen, T. (2003). Spectral features for automatic text-independent speaker recognition. Licentiate's thesis.

Mel-frequency cepstral coefficients

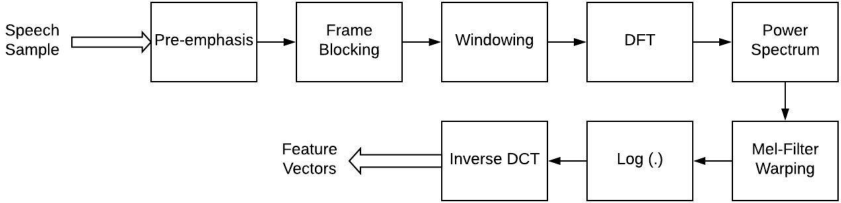

Figure: MFCC feature extraction

Delta and Delta-Delta cepstra

- Static MFCC feature vectors provides a good estimation of local spectra.
- Fails to capture the dynamics of human speech. (important for discrimination)

$$\Delta c_i(n) = \frac{\sum_{k=-N}^{k=N} k c_i(n+k)}{\sum_{k=-N}^{k=N} k^2} \tag{1}$$

- N defines the no of frames to be used for the computation of delta coefficients. (N=2 to 4)
- End-effect problem can be solved by using simple first order differences at the start and end frame.

$$\begin{aligned} \Delta\Delta c_i(n) &= \Delta c_i(n+1) - \Delta c_i(n), \quad & n < N \\ \Delta\Delta c_i(n) &= \Delta c_i(n) - \Delta c_i(n-1), \quad & n \geq T - N \end{aligned} \tag{2}$$

Excitation source features

- Residual MFCC

$$e[n] = s[n] - \tilde{s}[n]$$

$$\tilde{s}[n] = \sum_{k=1}^{p} a_k s(n-k)$$

p is the LP order

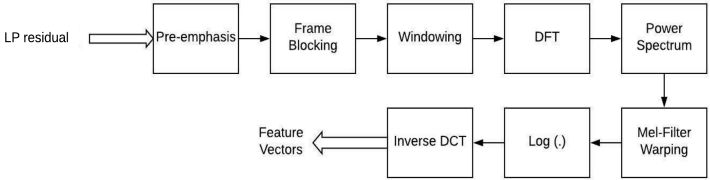

Figure: Block diagram for the extraction of RMFCC from LP residual.

Model Training

1. Vector Quantization (VQ)
2. Gaussian Mixture Model (GMM)
3. Gaussian Mixture Model and Universal background model (GMM-UBM)
4. I-vector approach
5. Neural network based approaches

Vector Quantization

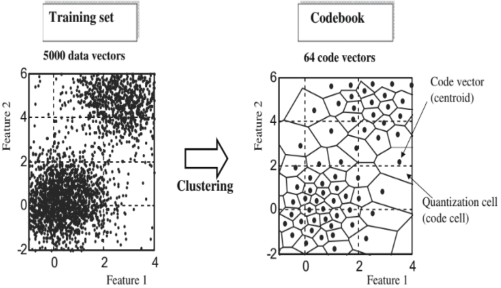

Figure: Codebook construction for vector quantization using the K-means algorithm.

$$D_Q(X, R) = \frac{1}{T} \sum_{t=1}^{T} \min_{1 \leq k \leq K} d(x_t, r_k)$$

Courtesy: Kinnunen, T. and Li, H., 2010. An overview of text-independent speaker recognition: From features to supervectors. Speech communication, 52(1), pp.12-40.

Gaussian Mixture Model

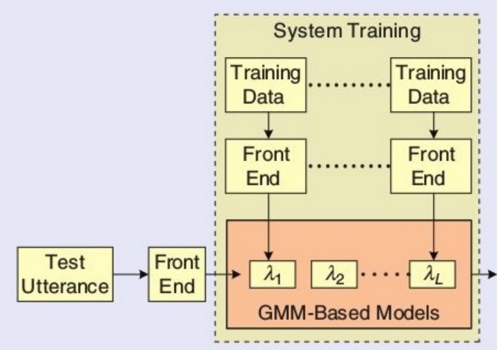

Figure: Block diagram of GMM [1]

Gaussian Mixture Model

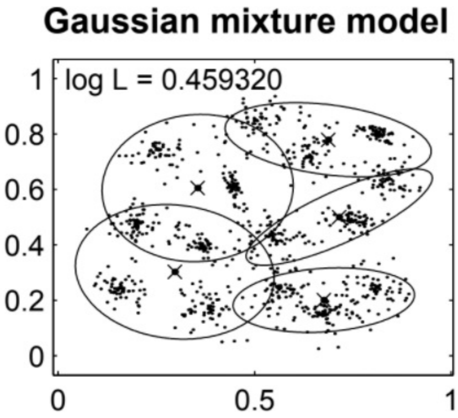

Figure: GMM modeling

Courtesy: Kinnunen, T., 2003. Spectral features for automatic text-independent speaker recognition. Licentiate's thesis.

Testing Gaussian Mixture Model

- The objective is to identify the hypothesized model from a set of models $\{S_1, S_2, ..., S_M\}$ given a set of testing vectors $X = \{x_1, x_2, .., x_T\}$.
- The identified model can be written as:

$$\begin{aligned} \hat{S} =& \arg \max_{1 \le i \le M} P(S_i|X) \\ =& \arg \max_{1 \le i \le M} \frac{P(X|S_i)}{P(X)} P(S_i) \end{aligned} \tag{3}$$

- Using ML detection criteria (i.e Assuming equal probability occurrence of all the models) and the statistical independence of the testing vectors.

$$\hat{S} = \arg \max_{1 \le i \le M} \sum_{j=1}^{T} \log(P(x_j|S_i)) \tag{4}$$

GMM-UBM

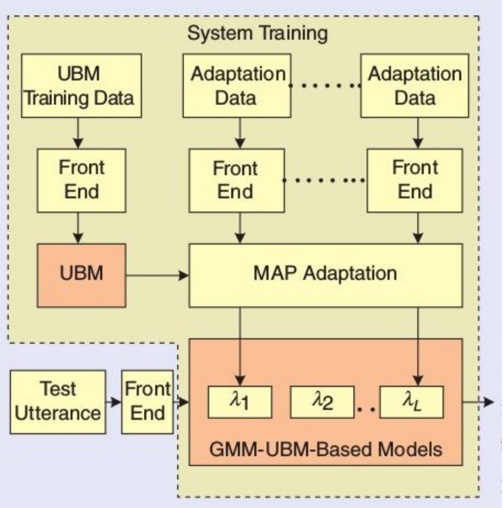

Figure: Block diagram of GMM-UBM [1]

MAP Adaptation

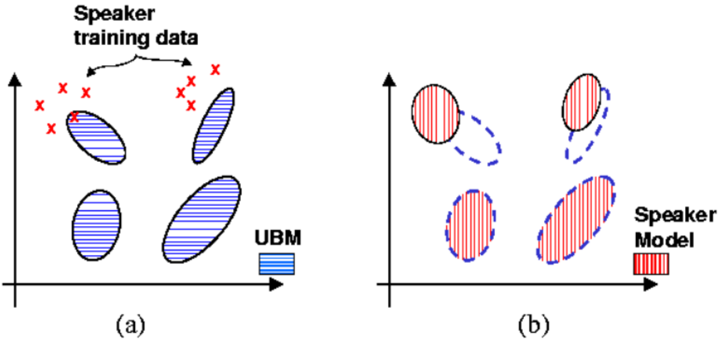

Figure: (a) The training vectors (x's) are probabilistically mapped into the UBM (prior) mixtures. (b) The adapted mixture parameters are derived using the statistics of the new data and the UBM (prior) mixture parameters. The adaptation is data dependent, so UBM (prior) mixture parameters are adapted by different amounts [2].

Testing of GMM-UBM

- Given a set of testing vectors $\{x_1, x_2, \ldots, x_T\}$, UBM model S^{ubm} and the adapt models of each speaker $\{S_i^{adapt}\}_{i=1}^{M}$, the identified speaker model can be evaluated as:

$$\hat{S} = \text{arg} \max_{1 \leq i \leq M} \sum_{j=1}^{T} \left[\log(P(x_j | S_i^{adapt})) - \log(P(x_j | S^{ubm}) \right] \tag{5}$$

where M is the total no of languages used in the system.

I-vector

- I-vector Based language recognition system:
  1. Total variability factors (w): represent each speech utterance.
  2. w is known as the i-vector.

$$M = m + Tw$$

$M =$ Utterance super-vector
$m =$ UBM-mean super-vector
$T =$ Total variability matrix
$w =$ I-vector

- Intersession compensation: LDA, WCCN, NAP
- Cosine distance score:

$$score(w_1, w_2) = \frac{< w_1 w_2 >}{|w_1||w_2|}$$

Block diagram of I-vector

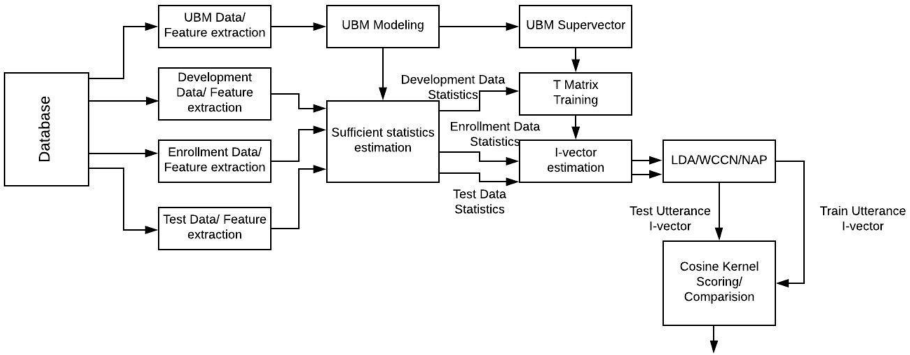

Figure: Basic block diagram of speaker identification using i-vector.

Deep neural network (DNN) based speaker recognition system

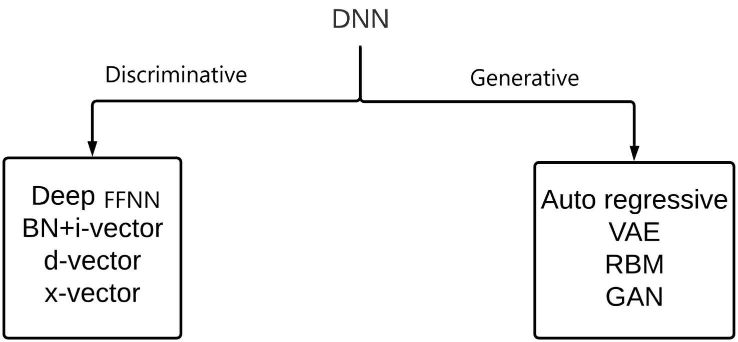

Figure: Overview of DNN

Feedforward neural network

- Used as a classifier

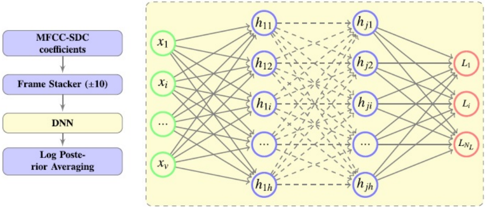

Figure: Feedforward network based classifier [3].

- Testing:

$$score_l = \frac{1}{N}\sum_{t=1}^{N} logp(L_l|x_t, \Theta) \quad (6)$$

Bottleneck Feature

- Bottleneck feature based I-vector framework:

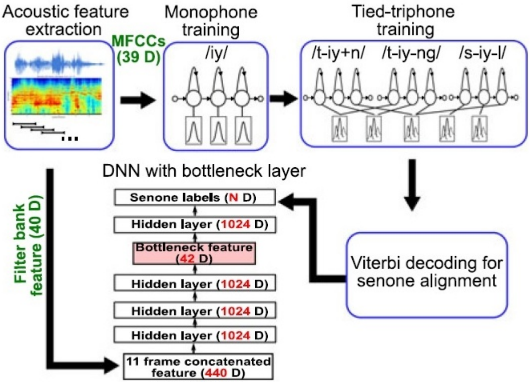

Figure: Bottleneck feature extraction [4].

Bottleneck Feature

- After BNF extraction I-vector framework is used for classification.

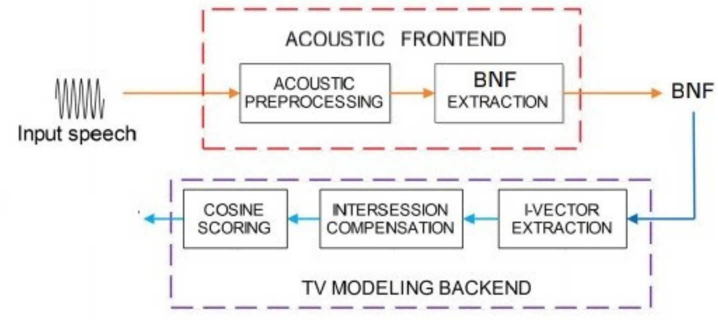

Figure: Bottleneck feature-I-vector based speaker identification system [5].

X-vector

- X-vector: The DNN, trained to discriminate between languages
- Variable length utterances: fixed dimensional embedding
- Architecture: time delay neural network (TDNN)
- Sub-sampling: reduce the computation during training.

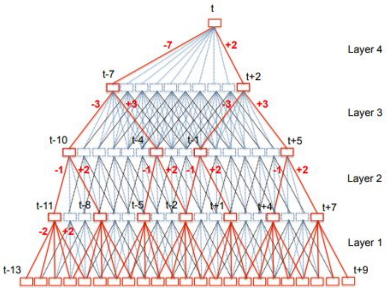

Figure: Computation in TDNN with sub-sampling (red) and without sub-sampling (blue+red) [6].

X-vector

- X-vector architecture configuration:

Layer	Layer Context	Total Context	*Input* $\times$ *Output*
layer 1	[t-2,t+2]	5	$5F \times 512$
layer 2	{t-2,t,t+2}	9	1536×512
layer 3	{t-3,t,t+3}	15	1536×512
layer 4	{t}	15	512×512
layer 5	{t}	15	512×1500
Stats Pooling	[0,T)	T	$1500T \times 3000$
Segment 6	{0}	T	3000×512
Segment 7	{0}	T	512×512
Softmax	{0}	T	$512 \times L$

- Data augmentation: provides robustness against noise.
- After X-vector is extracted cosine distance measure (like I-vector) is used for classification.

GAN Architecture

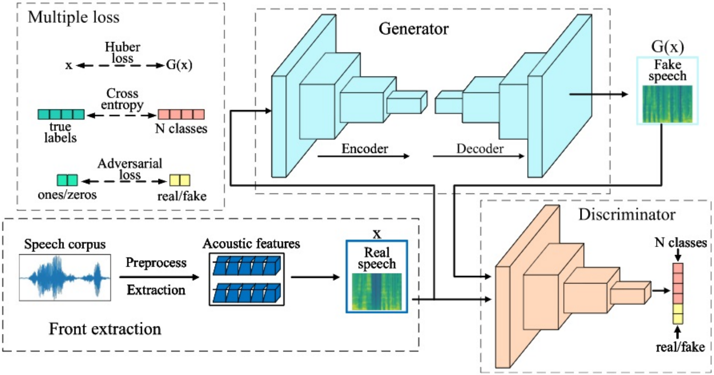

Figure: Speaker GAN architecture [7]

Database

- TIMIT
- NIST Speaker recognition series
- SITW (open source)
- Voxcleb (open source)
- IITG-MV

- References

E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu, "Language identification: A tutorial," *IEEE Circuits and Systems Magazine*, vol. 11, no. 2, pp. 82–108, 2011.

D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

I. Lopez-Moreno, J. Gonzalez-Dominguez, D. Martinez, O. Plchot, J. Gonzalez-Rodriguez, and P. J. Moreno, "On the use of deep feedforward neural networks for automatic language identification," *Computer Speech & Language*, vol. 40, pp. 46–59, 2016.

Q. Zhang and J. H. Hansen, "Language/dialect recognition based on unsupervised deep learning," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 5, pp. 873–882, 2018.

B. Jiang, Y. Song, S. Wei, J.-H. Liu, I. V. McLoughlin, and L.-R. Dai, "Deep bottleneck features for spoken language identification," *PloS one*, vol. 9, no. 7, p. e100795, 2014.

D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors," in *Odyssey: The Speaker and Language Recognition Workshop, Les Sables d'Olonne*, 2018.

L. Chen, Y. Liu, W. Xiao, Y. Wang, and H. Xie, "Speakergan: Speaker identification with conditional generative adversarial network," *Neurocomputing*, vol. 418, pp. 211–220, 2020.