d-Vector Approach for Speaker Diarization

Research Paper : DEEP NEURAL NETWORKS FOR SMALL FOOTPRINT TEXT-DEPENDENT SPEAKER VERIFICATION

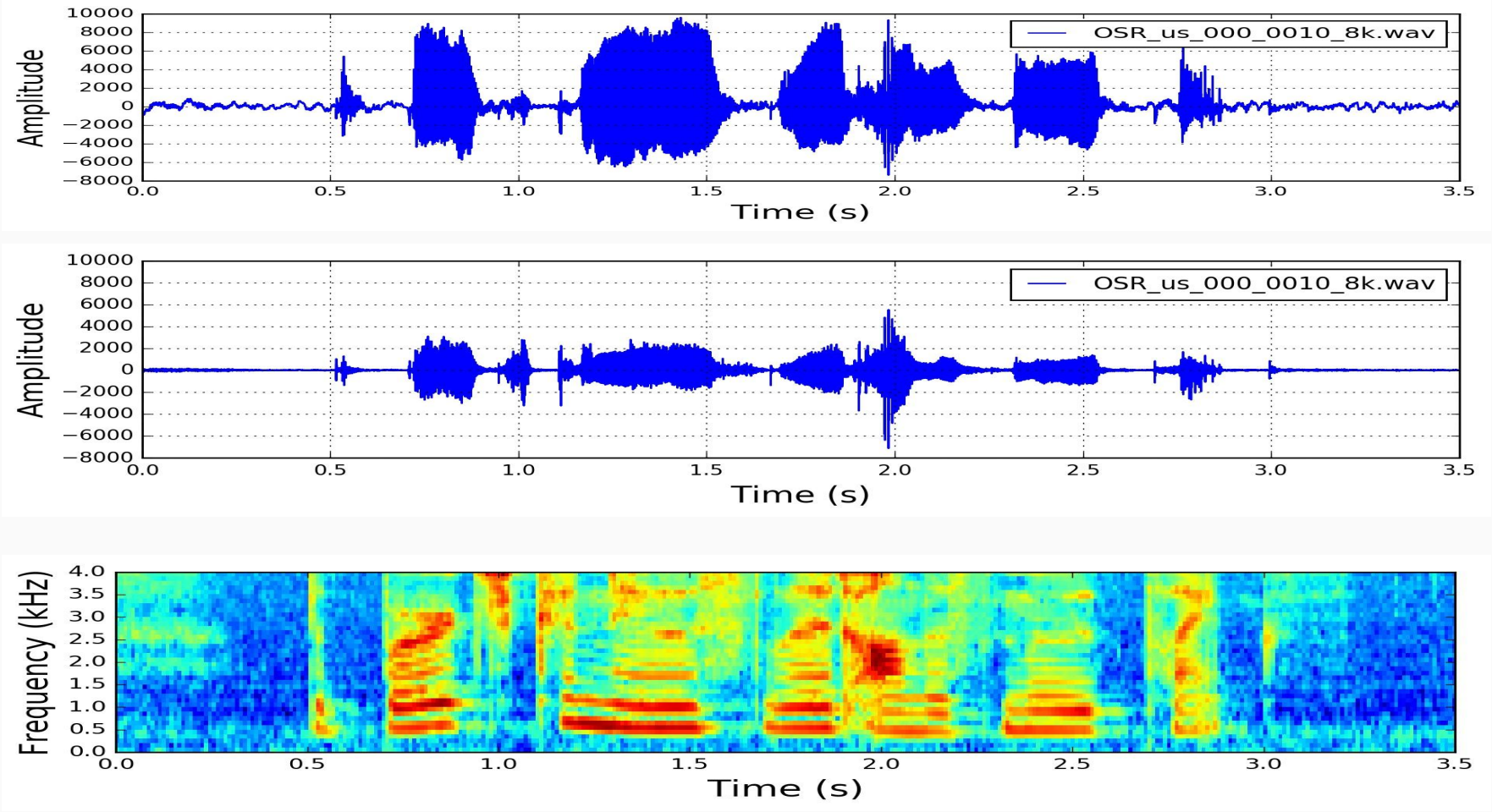
Piyush Tiwary
B.Tech Electrical Engineering 2017-21

Speaker Diarization

- In most real-world scenarios speech does not come in well defined audio segments with only one speaker. In most of the conversations more than one person are involved.
- In many application we would like identify multiple speakers in a conversation.
- Speaker Diarization is the solution for all these problems.
- Speaker Diarization answers the question “**Who Spoke When?**”.
- Application where Speaker Diarization is used -
  - In Meetings
  - Automatic Transcripts (Used by Youtube)
  - Google Assistant

How do we a Speech Signal to a Deep Learning Framework?

- WE DON'T!
- We extract features from Speech Signal itself which can we fed to our Deep Learning framework.
- These features are also called Filter Banks or mel - Filter Banks or log - Filter Banks.
- All of these are variants of same concepts each have there own pros & cons.
- Main of these is to reduce the noise in speech (improve Speech to Noise Ratio) and balance the frequency spectrum (boosting signal's weaker/higher frequency part).



Effect of extracting Filter Banks

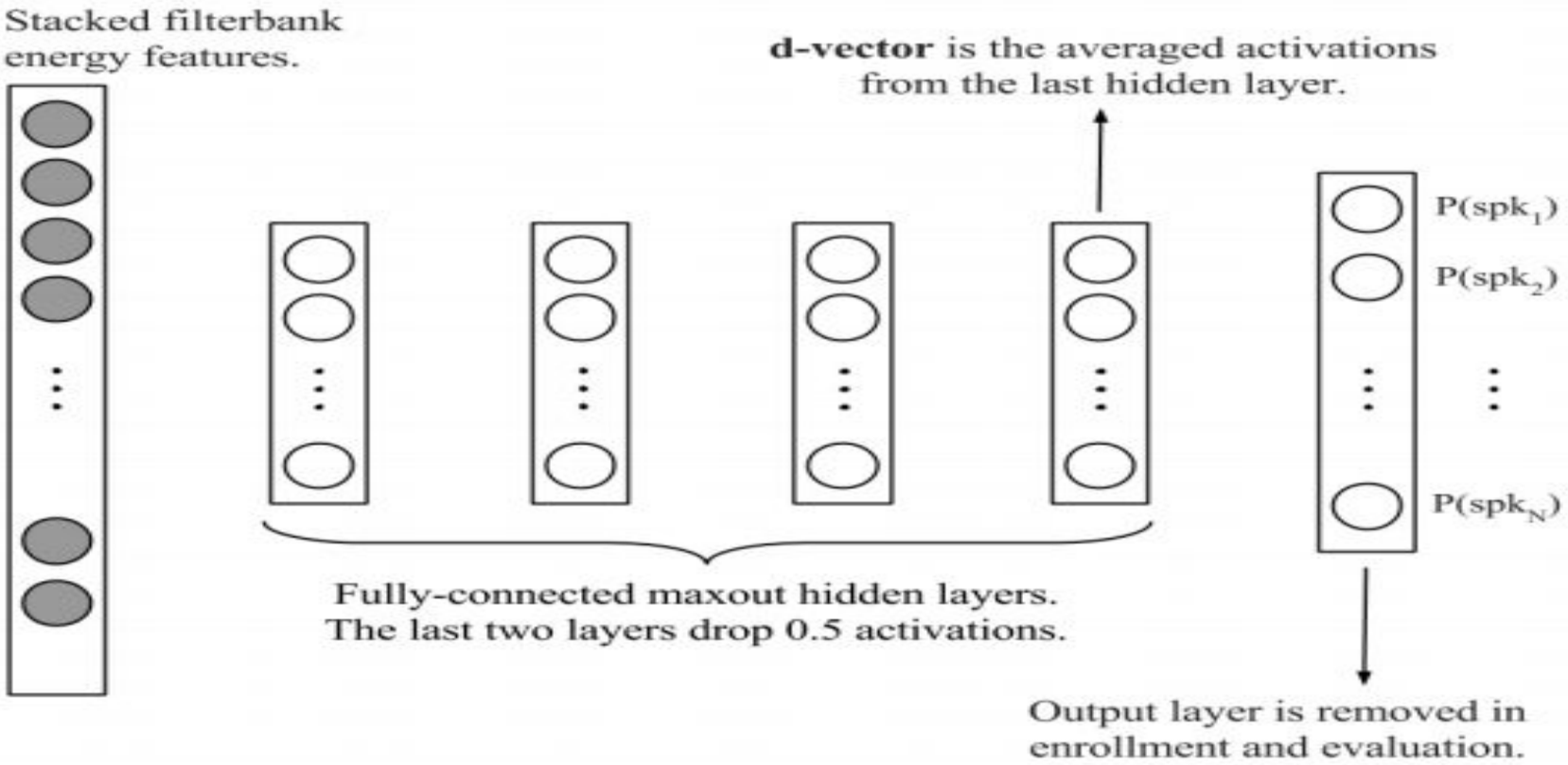
A Typical Diarization System

1. **Speech Segmentation -** Input audio is segmented into small segments.
2. **Audio Embedding Extraction -** Specific features such as MFCCs, Speaker factors or *i*-vectors are extracted from the segments.
3. **Clustering -** Where the number of speakers is determined and extracted audio embeddings are clustered into these speakers optimally.
4. **Resegmentation -** Clustering results are further refined to produce final diarization results.

- **d-Vectors** are part of 2nd stage (Audio Embedding Extraction).

d - Vectors

- d-Vectors are also a kind of extracted feature embedding except it is extracted from a Deep Neural Network.
- We feed the stacked filterbank energy features to a DNN and train it.
- Once the DNN is has been trained successfully, we use the accumulated output activations of the last hidden layer as new speaker representation.
- That is, for every frame of a given utterance belonging to a new speaker, we compute the o/p activation of last hidden layer using standard feed-forward network in the trained DNN, and accumulate these activations to form a new compact representation of that speaker, the “d-Vector”.



DNN Model for Speaker Verification

Why last Hidden Layer?

1. Intuitively, the last hidden layer is the one which has learnt the most amongst all the layers, because it has learnt from all it’s previous layer and is able to extract deeper features useful for Speaker Verification.
2. It is observed that it gives better generalization to unseen speakers.
3. It is able to tackle the noise present in the data.

Complete Working of d-Vectors

1. **Development** : Training of model on large collection of data. This is done to create background models to define speaker manifold.
2. **Enrollment** : New speakers are enrolled by deriving speaker specific information to obtain speaker dependent models. Speakers in Enrollment and Development stage are not overlapped.
3. **Evaluation** : Each test utterance is evaluated using enrolled speaker models and background models.

Continued...

During enrollment, the speaker model is computed as the average of activations derived from d-Vectors.

In evaluation phase, we make decision based on distance between target d-Vectors and the test d-Vectors.

Benefits of using d-Vectors

- Since the time when d-Vectors are introduced entire Speaker Verification related industries have shifted from *i*-vectors to *d*-vectors for Speaker embeddings.
- Reasons being high ‘Equal Error Rate (EER)’, low complexity of model, and robustness against noise.
- Experimental results show that the performance of the d-vector SV system is reasonably good compared to an i-vector system, and system fusion achieves much better results than the standalone i-vector system.
- A simple sum fusion of these two systems can improve the i-vector system performance in all operating points.

Continued...

- The EER of the combined system is 14% and 25% better than our classical i-vector system in clean and noisy conditions respectively.
- Furthermore, the d-vector system is more robust to additive noise in enrollment and evaluation data.
- At low False Rejection operating points, the d-vector system outperforms the i-vector system.

#Gaussians	i-vector Dim	LDA Dim	#Params	EER (raw)	EER (t-norm)
1024	300	200	12.2M	2.92%	2.29%
256	200	100	2.1M	3.11%	2.92%
128	100	100	540K	3.50%	2.83%

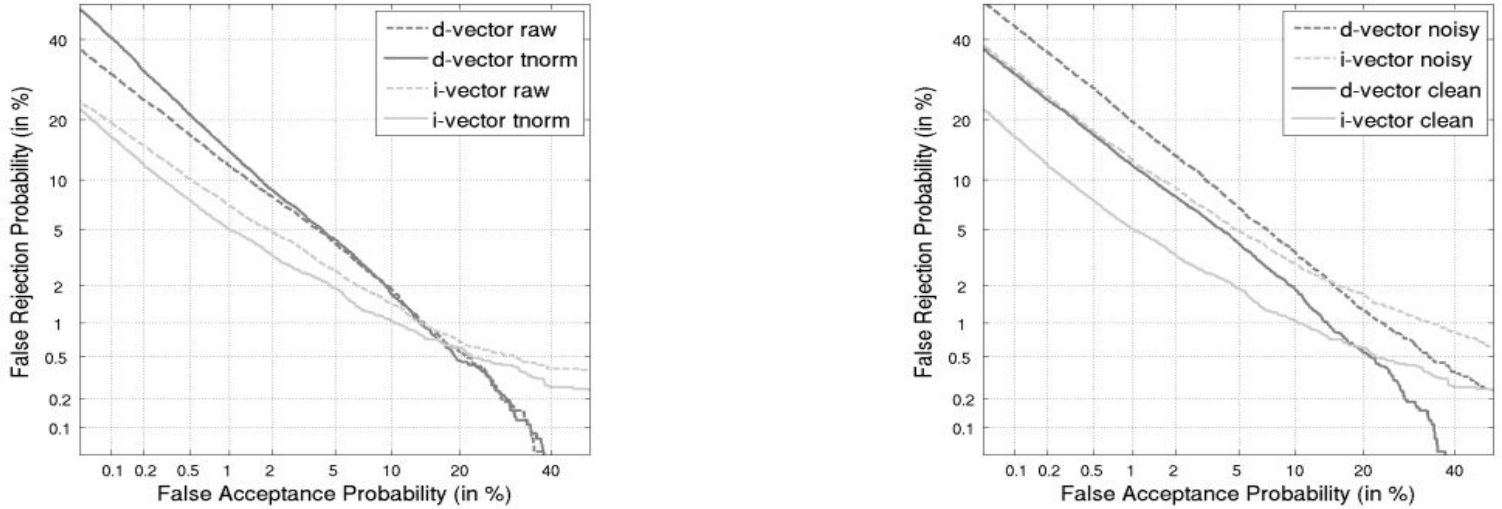
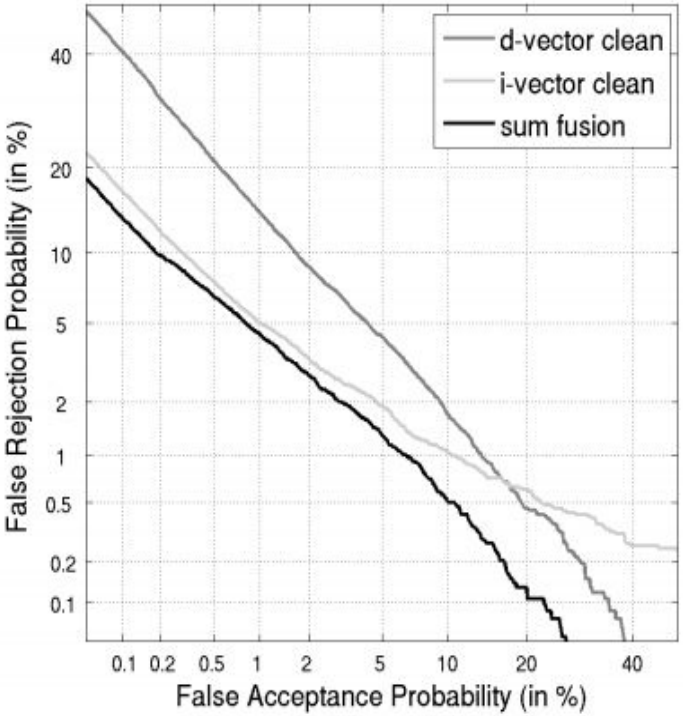


Fig. 2. Left: DET curve comparison between i-vector and d-vector speaker verification systems using raw and t-norm scores. Right: DET curve comparison of the two systems in clean and noisy conditions.

Experimental Results



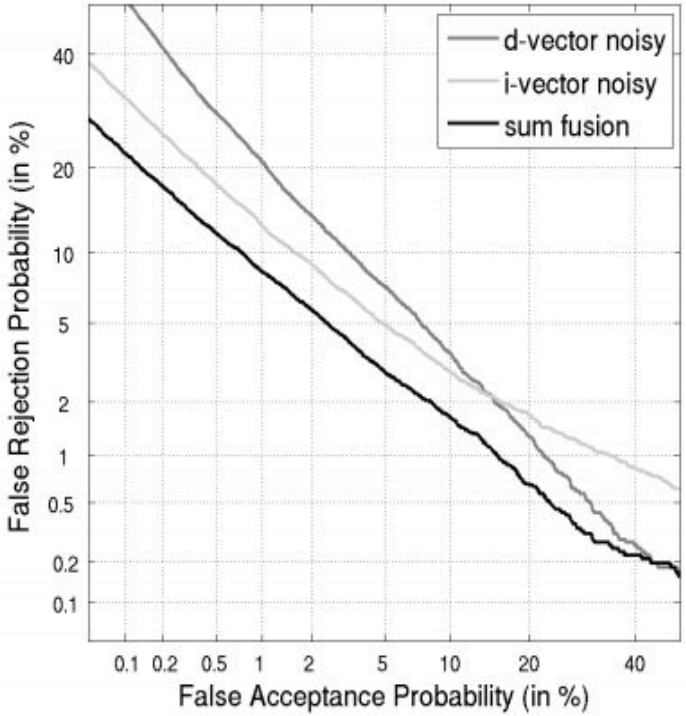


Fig. 3. DET curve for the sum fusion of the i-vector and d-vector systems in clean (left) and noisy (right) conditions.

Thanks!