D-Vector Embedding: TE2E to GE2E

Fig. We use multi-layer LSTM network as audio feature extractor.
- Tuple E2E [2]: Simulate enroll-verify runtime logic during training. However, tuples are randomly selected, and most tuples are easy—inefficient.
- Generalized E2E loss [1]: For each speaker, focus on its most offensive impostor in the batch.

Clustering Algorithms

- **Naive online**
  - Thresholding the cosine similarities to centroids.

- **K-Means offline**
  - K-Means++ for cluster initialization.
  - Find k using Mean Squared Cosine Distances (MSCD): \( k = \arg\max_{k} \text{MSCD}(k) \)

- **Spectral offline (Winner!**) (Naive-decomposition affinity matrix. Run K-Means on dimensionality-reduced embeddings. Find k using the max eigen-gap criterion.

Affinity matrix refinement: The key to the success of spectral clustering.

- Gaussian blur: Smooth the data, and reduce the effect of outliers.
- Row-wise thresholding: Zero-out affinities between different speakers.
- Symmetrization: Restore matrix symmetry.
- Diffusion: Sharpen affinity section boundaries of distinct speakers.
- Row-wise max normalization: Avoid undesirable scale effects.

Fig. We applied a sequence of refinement operations on the affinity matrix.

Experiment Results

| Table (Up), DER (%) on CALLHOME American English 2-speaker subset (CH-109). |
|-------------------------------|-------------------------------|
| **Method** | **Confusion** | **PA** | **Miss** | **FA** | **Prec.** | **Rec.** |
| Our Model | 9.70 | 2.51 | 4.06 | | | |
| Callisto [4] | 13.7 | - | - | | | |
| Shum [5] | 12.1 | - | - | | | |
| Senoussaoui [6] | 12.1 | - | - | | | |
| Sell [7] [8] | 13.7 (105) | - | - | | | |
| Romero [8] [9] | 12.8 (109) | - | - | | | |

References