Speaker Recognition in the Wild

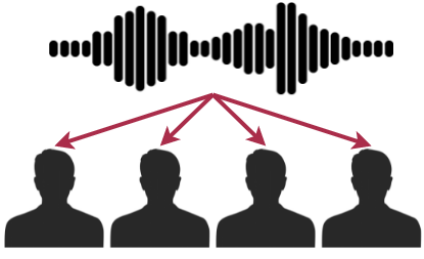
Audio Pattern Recognition Project

Gabriele Cerizza

Università degli Studi di Milano

TASK OVERVIEW

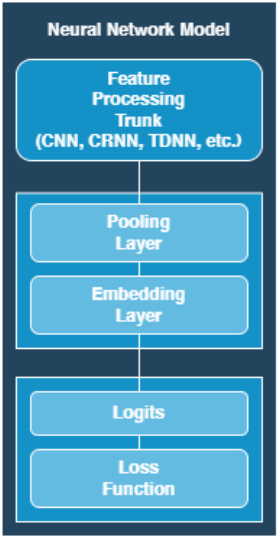
Speaker Identification



Who is the speaker?

Speaker Verification

Is this speaker A?

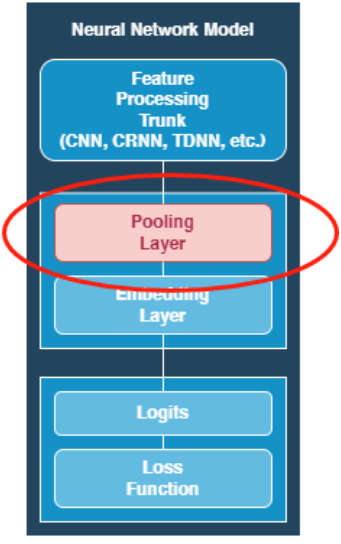


- **First block (“Trunk”)**
  - Takes as input acoustic features (MFCCs or Mel spectrograms) and outputs frame-level features
- **Second block**
  - Pooling layer to aggregate frame-level features of varying length into fixed dimensional utterance-level features
  - Fully connected layer to produce embeddings
- **Third block**
  - Projection into a dimension whose size is the number of speakers
  - Loss function

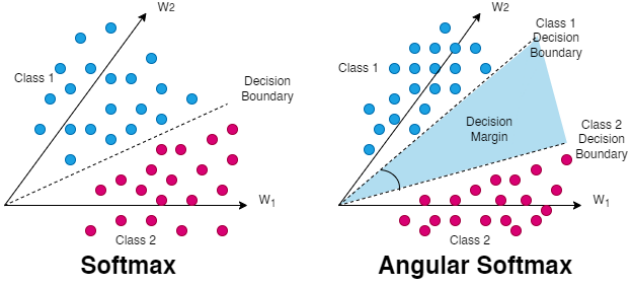
- **Scoring**
  - PLDA
  - Cosine distance
- **Score normalization**
  - Adaptive S-Norm
- **Metrics**
  - EER
  - MinDCF

$$C_{miss} \times P_{target} \times P_{miss}(\theta) + C_{fa} \times (1 - P_{target}) \times P_{fa}(\theta)$$

- **Temporal Average Pooling (TAP)**
- **Self-Attentive Pooling (SAP)**
- **Other Pooling Layers**
  - Self Multi-Head Attention Pooling
  - Attentive Statistics Pooling (ASP)



- **Softmax**
  - Does not enforce intra-class compactness and inter-class separation
- **Angular Additive Margin Softmax (AAM Softmax)**
  - Improves compactness and separation
  - Difficult to train with and sensitive to parameters
- **Sub-center Angular Additive Margin Softmax (SC AAM Softmax)**
  - More robust against noisy data



EXPERIMENT

» **VoxCeleb1 Dataset**

› Collected "in the wild" from Youtube

› Background noise, music, overlapping speech and varying room acoustics

› Heterogeneous age, gender and nationality

» **We used a subset due to hardware constraints**

› 100 speakers, retaining gender ratio

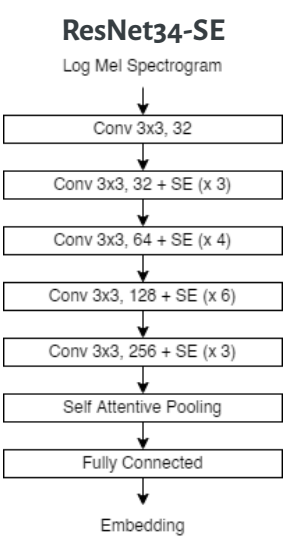
› Official training, validation and test splits

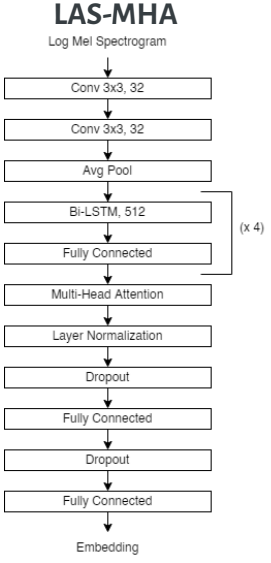
› For verification, 10 speakers not present in the training set

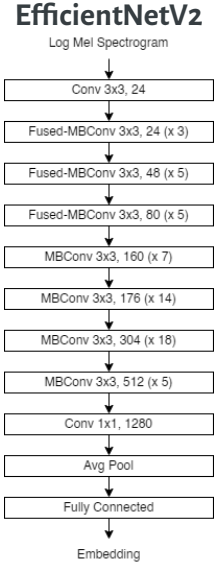
		Full Dataset	Identif. Set	Verif. Set
Speakers No.		1,251	100	10
Samples No.		152,123	12,123	1,896
Gender	*Male*	0.55	0.51	0.50
	Female	0.55	0.52	0.50
Nationality[a]	*USA*	0.65	0.52	0.41
	UK	0.21	0.12	0.32
	Italy	0.01	0.05	0.07
	Russia	0.07	0.02	0.02
Seconds	*Mean*	8.12	7.59	5.67
	Std	2.55	3.55	4.55

[a]Only the four most frequent nationalities are listed.

- **Features**
  - 80-dimensional log Mel spectrograms
  - Window length of 25 ms, frame-shift of 10 ms
  - Cepstral mean normalization
- **Offline data augmentation**
  - Speed perturbation
  - Background noise
  - Babble
  - Reverberation
- **Online data augmentation**
  - SpecAugment: time and frequency masking


ResNet34-SE
Log Mel Spectrogram
Conv 3x3, 32
Conv 3x3, 32 + SE (x 3)
Conv 3x3, 64 + SE (x 4)
Conv 3x3, 128 + SE (x 6)
Conv 3x3, 256 + SE (x 3)
Self Attentive Pooling
Fully Connected
Embedding


LAS-MHA
Log Mel Spectrogram
Conv 3x3, 32
Conv 3x3, 32
Avg Pool
Bi-LSTM, 512
Fully Connected
(x 4)
Multi-Head Attention
Layer Normalization
Dropout
Fully Connected
Dropout
Fully Connected
Embedding


EfficientNetV2
Log Mel Spectrogram
Conv 3x3, 24
Fused-MBConv 3x3, 24 (x 3)
Fused-MBConv 3x3, 48 (x 5)
Fused-MBConv 3x3, 80 (x 5)
MBConv 3x3, 160 (x 7)
MBConv 3x3, 176 (x 14)
MBConv 3x3, 304 (x 18)
MBConv 3x3, 512 (x 5)
Conv 1x1, 1280
Avg Pool
Fully Connected
Embedding

- **Large margin fine-tuning**
  - Training in two steps
  - Random crops: from 2 to 4 seconds
  - SC AAM Softmax: margin from 0.1 to 0.3, scale from 15 to 30
  - SpecAugment only in the first step
- **Evaluation**
  - Full-length utterances for verification
  - Random 4 seconds crops for identification

Model	Year	Training Set	Top1 Acc.	Top5 Acc.	F1 Score	EER(%)	MinDCF[a]
Nagrani et al.	2017	VoxCeleb1	80.5	92.1	-	-	-
Nagrani et al.	2017	VoxCeleb1	-	-	-	7.8	0.71 (0.01)
Cai et al.	2018	VoxCeleb1	89.9	95.7	-	-	-
Cai et al.	2018	VoxCeleb1	-	-	-	5.27	0.439
Cai et al.	2018	VoxCeleb1	-	-	-	4.46	0.577
Chung et al.	2018	VoxCeleb1	-	-	-	7.8	0.71 (0.01)
Okabe et al.	2018	VoxCeleb1	-	-	-	3.85	0.406 (0.01)
Hajibabaei, Dai	2018	VoxCeleb1	94.6	98.1	-	4.69	0.453 (0.01)
Hajibabaei, Dai	2018	VoxCeleb1	92.8	97.5	-	4.30	0.413 (0.01)
India et al.	2019	VoxCeleb1	-	-	-	4.0	0.0045 (0.01)
Chung et al.	2019	VoxCeleb1	89.00	96.15	-	5.37	-
Chung et al.	2019	VoxCeleb1	89.00	95.94	-	5.26	-
Hajavi, Etemad	2021	VoxCeleb1	-	-	-	3.14	-
Kwon et al.	2021	VoxCeleb2	-	-	-	0.73	0.056 (0.05)
Kwon et al.	2021	VoxCeleb2	-	-	-	0.74	0.061 (0.05)
Thienpondt et al.	2021	VoxCeleb2	-	-	-	0.64	0.0700 (0.01)
Thienpondt et al.	2021	VoxCeleb2	-	-	-	0.56	0.0743 (0.01)
Zhao et al.	2021	VoxCeleb2	-	-	-	0.5249	0.0498 (0.01)
Zhao et al.	2021	VoxCeleb2	-	-	-	0.5594	0.0480 (0.01)

[a] If provided, we noted the P_{target} value within parentheses.

Thank You