Speaker Recognition in the Wild

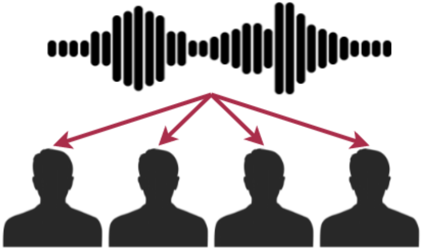
Audio Pattern Recognition Project

Gabriele Cerizza

Università degli Studi di Milano

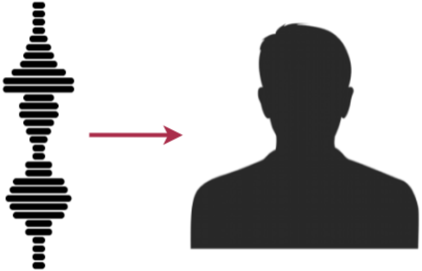
TASK OVERVIEW

Speaker Identification

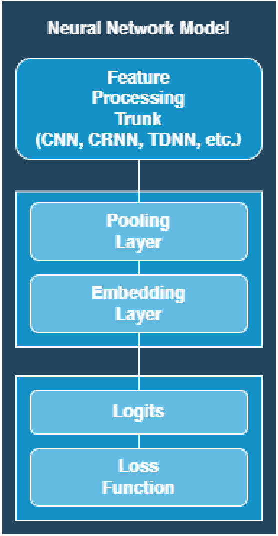


Who is the speaker?

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

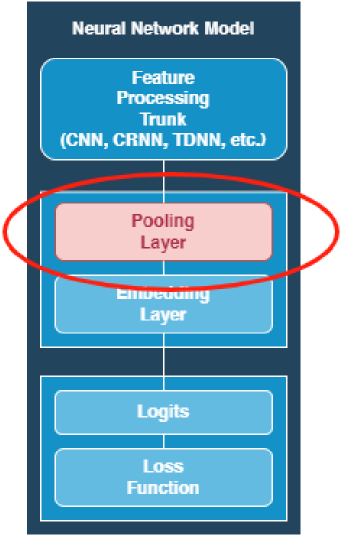
$$C_{\text{miss}} \times P_{\text{target}} \times P_{\text{miss}}(\theta) + C_{\text{fa}} \times (1 - P_{\text{target}}) \times P_{\text{fa}}(\theta)$$

» **Temporal Average Pooling (TAP)**

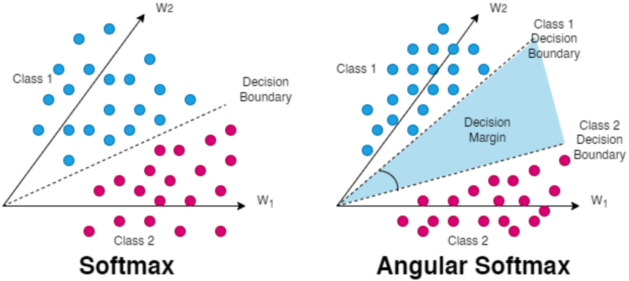
» **Self-Attentive Pooling (SAP)**

» **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

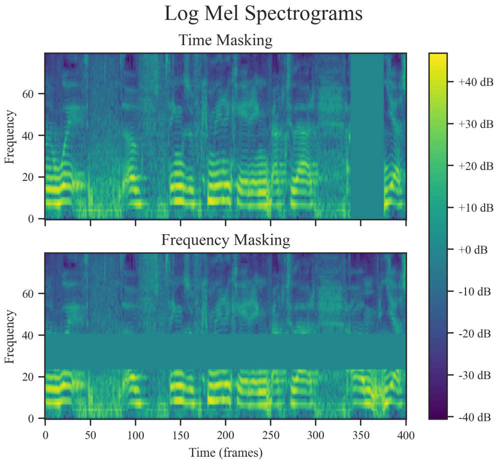
		Full Dataset	Identif. Set	Verif. Set
Speakers No.		1,251	100	10
Samples No.		153,516	13,042	758
Gender	*Male*	0.55	0.55	0.50
	Female	0.45	0.45	0.50
Nationality[a]	*USA*	0.64	0.59	0.80
	UK	0.17	0.19	0.10
	Canada	0.04	0.02	-
	Australia	0.03	0.04	-
Seconds	*Mean*	8.25	8.20	8.05
	Std	5.31	5.35	5.14

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

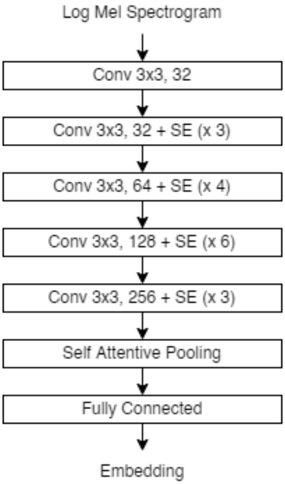
» **Features**
  - 80-dimensional log Mel spectrograms
  - Window length of 25 ms, frame-shift of 10 ms
  - Cepstral mean normalization

» **Offline data augmentation**
  - Speed perturbation
  - Background noise
  - Babble
  - Reverberation

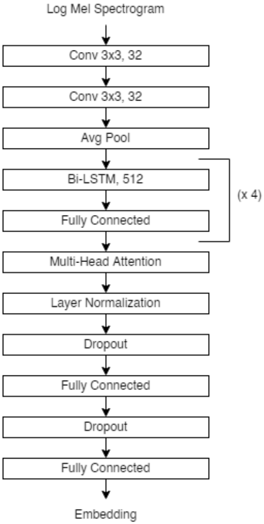
» **Online data augmentation**
  - SpecAugment: time and frequency masking



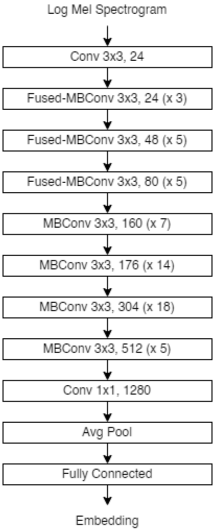
ResNet34-SE



LAS-MHA



EfficientNetV2



- **Large Margin Fine-Tuning**
  - Training in two steps
  - Random crops: from 2 to 4 seconds
  - SC AAM Softmax: margin from 0.1 to 0.15, scale from 15 to 20
  - SpecAugment only in the first step
- **Evaluation**
  - Full-length utterances for verification
  - Random 4 seconds crops for identification

Model	Year	Training Set	Top1 Accuracy	Top5 Accuracy	F1 Score	EER(%)	MinDCF[a]
Nagrani et al.	2017	VoxCeleb1	80.50	92.10	-	-	-
Nagrani et al.	2017	VoxCeleb1	-	-	-	7.80	0.710 (0.01)
Cai et al.	2018	VoxCeleb1	89.90	95.70	-	-	-
Cai et al.	2018	VoxCeleb1	-	-	-	5.27	0.439
Cai et al.	2018	VoxCeleb1	-	-	-	4.46	0.577
Okabe et al.	2018	VoxCeleb1	-	-	-	3.85	0.406 (0.01)
Hajibabaei, Dai	2018	VoxCeleb1	**94.60**	**98.10**	-	4.69	0.453 (0.01)
Hajibabaei, Dai	2018	VoxCeleb1	92.80	97.50	-	4.30	0.413 (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	-
Thienpondt et al.	2021	VoxCeleb2	-	-	-	0.64	0.070 (0.01)
Thienpondt et al.	2021	VoxCeleb2	-	-	-	0.56	0.074 (0.01)
Zhao et al.	2021	VoxCeleb2	-	-	-	**0.52**	0.050 (0.01)
Zhao et al.	2021	VoxCeleb2	-	-	-	0.56	**0.048 (0.01)**
SVM (our baseline)	2022	VoxCeleb1 Subset	13.98	-	11.46	-	-
ResNet34-SE (ours)	2022	VoxCeleb1 Subset	64.21	84.21	59.28	14.50	0.893 (0.01)
LAS-MHA (ours)	2022	VoxCeleb1 Subset	47.97	66.62	42.94	21.69	0.980 (0.01)
EfficientNetV2 (ours)	2022	VoxCeleb1 Subset	67.67	85.11	**64.26**	16.44	0.974 (0.01)

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

Thank You