ROBOVOX Speaker Verification Challenge

Signal Processing Cup
ICASSP 2024, Seoul, South Korea

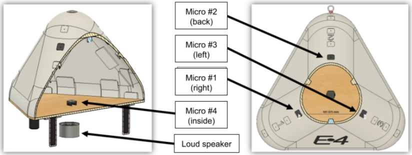
Atharv Ramesh[1,2], Anirudh Srinivasan[1,3], Tejadhith Sankar[1,2], Vaideeshwaran AP[2], Sreekanth Sankala[2], K Sri Rama Murty[2], Himanshu Gupta [4]

Presenting Authors[1]
Department of Electrical Engineering[2], Computer Science and Engineering[3], Artificial Intelligence[4]
Indian Institute of Technology Hyderabad

Table of Contents

Problem Description

- Text Independent Far Field Speaker Verification

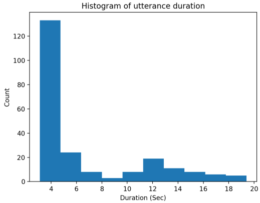


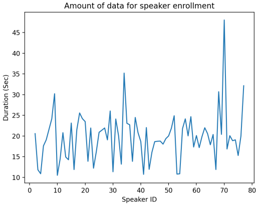
Robovox (E4): a mobile robot

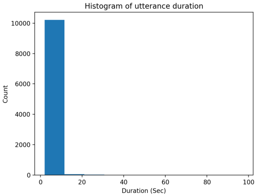
- Challenges
  - Ambient noise, Reverberation, Babble noise, Far-field speech
  - 1m, 2m, and 3m. Hall, Open space, Small and medium rooms.
  - Near the wall, Center of the room, or corner.
  - Robot internal noise, Angle between speaker and robot

Evaluation Dataset

- Robovox Dataset - Conversations between robot and speakers
  - Sampling frequency: 16000, Language: French
  - Enrollment - No.of Speakers: 75, Utts per speaker: 3, Avg dur: 6.5 sec
  - Test - No.of utterances: 10332, Avg duration: 3.37 sec
  - Sample Data - 30 minutes of data from 2 speakers

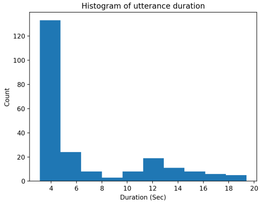


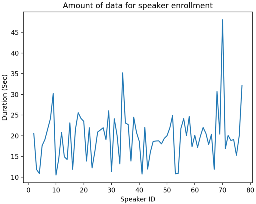


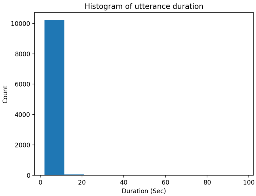


Evaluation Dataset

- Robovox Dataset - Conversations between robot and speakers
  - Sampling frequency: 16000, Language: French
  - Enrollment - No.of Speakers: 75, Utts per speaker: 3, Avg dur: 6.5 sec
  - Test - No.of utterances: 10332, Avg duration: 3.37 sec
  - Sample Data - 30 minutes of data from 2 speakers



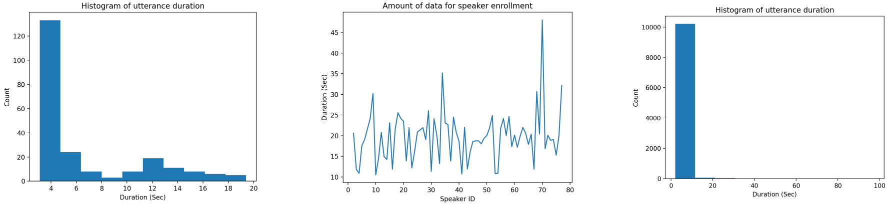




- Task1: **Far-field single-channel tracks:** Ch 5 vs Ch 4
- Task2: Far-field multi-channel tracks: Ch 5 vs. All excluding Ch 5

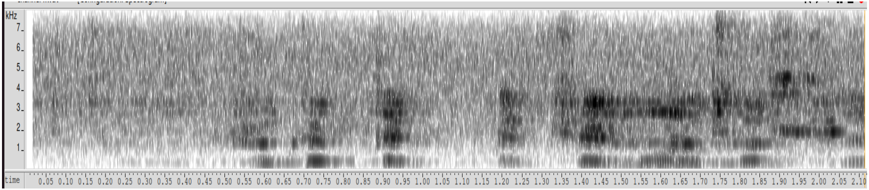
Evaluation Dataset

- Robovox Dataset - Conversations between robot and speakers
  - Sampling frequency: 16000, Language: French
  - Enrollment - No.of Speakers: 75, Utts per speaker: 3, Avg dur: 6.5 sec
  - Test - No.of utterances: 10332, Avg duration: 3.37 sec
  - Sample Data - 30 minutes of data from 2 speakers

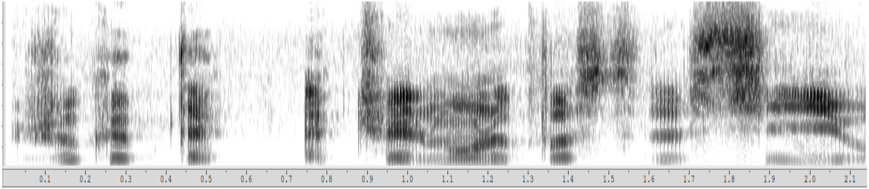


- Task1: **Far-field single-channel tracks:** Ch 5 vs Ch 4
- Task2: Far-field multi-channel tracks: Ch 5 vs. All excluding Ch 5
- Evaluation Metrics
  - Minimum Decision Cost Function (min_DCF)
  - Equal Error Rate (EER)

Spectrograms

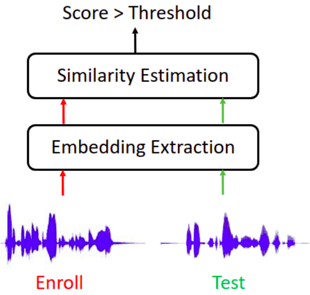


Channel 4



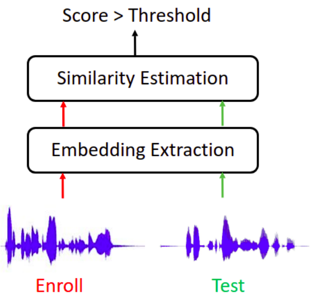
Channel 5

Basic Pipeline of Speaker Verification

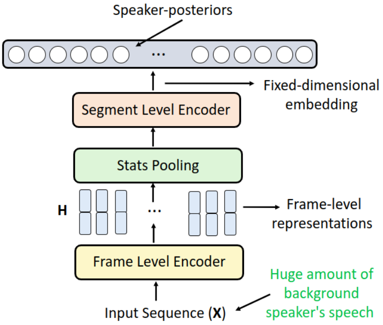


Speaker Verification

Basic Pipeline of Speaker Verification

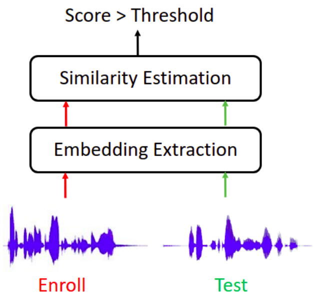


Speaker Verification

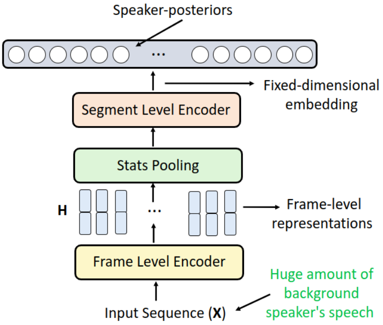


Speaker Embedding Extractor

Basic Pipeline of Speaker Verification



Speaker Verification

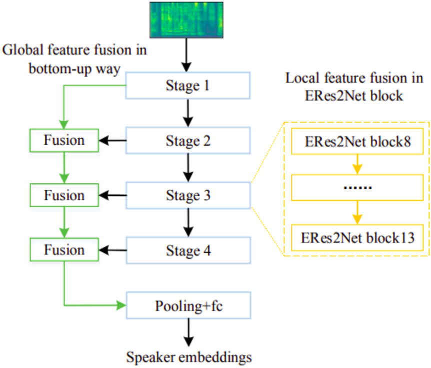


Speaker Embedding Extractor

- Frame-level encoder: DNN, CNN, ResNet, Transformer encoders, etc.
- Stats pooling: Mean, Standard deviation (Equal importance to all frames), Self-attention (Relative importance to the frames), etc

Enhanced Res2Net Architecture

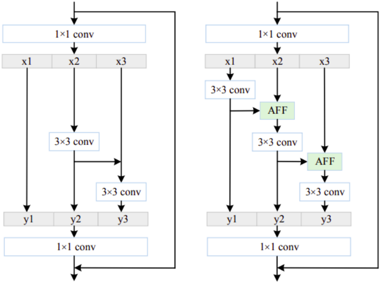
- ERes2Net: An Enhanced Res2Net with Local and Global Feature Fusion for Speaker Verification



Overview of the enhanced Res2Net framework

ERes2Net Architecture: Key Features

- Attentively fuses the local and global information



Local feature fusion

$$\mathbf{y}_i = \begin{cases} \mathbf{x}_i & i = 1 \\ \mathbf{K}_i(\mathbf{x}_i) & i = 2 \\ \mathbf{K}_i(\mathbf{x}_i + \mathbf{y}_{i-1}) & i \geq 2 \end{cases}$$

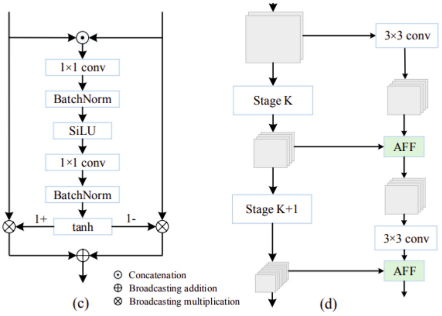
$$U = tanh(\mathbf{W}_2 \cdot g(\mathbf{W}_1([\mathbf{x}_i, \mathbf{y}_{i-1}])))$$

$$\mathbf{y}_i = \begin{cases} \mathbf{K}_i(\mathbf{x}_i) & i = 1 \\ \mathbf{K}_i(\color{red}{(U(\mathbf{x}_i, \mathbf{y}_{i-1}) + 1)} \cdot \mathbf{x}_i) & i > 1 \\ +\color{green}{(1 - U(\mathbf{x}_i, \mathbf{y}_{i-1}))} \cdot \mathbf{y}_{i-1})) & \end{cases}$$

ERes2Net Architecture: Key Features

Stage	Structure	Output size
	$3 \times 3, 32$	$T \times 80 \times 32$
Stage 1	$\begin{bmatrix} 1 \times 1 & 32 \\ 3 \times 3, 16, s = 2 \\ 1 \times 1 & 64 \end{bmatrix} \times 3$	$T \times 80 \times 64$
Stage 2	$\begin{bmatrix} 1 \times 1 & 64 \\ 3 \times 3, 32, s = 2 \\ 1 \times 1 & 128 \end{bmatrix} \times 4$	$T/2 \times 40 \times 128$
Stage 3	$\begin{bmatrix} 1 \times 1 & 128 \\ 3 \times 3, 64, s = 2 \\ 1 \times 1 & 256 \end{bmatrix} \times 6$	$T/4 \times 20 \times 256$
Stage 4	$\begin{bmatrix} 1 \times 1 & 256 \\ 3 \times 3, 128, s = 2 \\ 1 \times 1 & 512 \end{bmatrix} \times 3$	$T/8 \times 10 \times 512$
	Temporal statistics pooling	
	Fully connected layer	
	Softmax layer	

Architecture



Global feature fusion

System Specifications

- Feature Extraction:
  - 80 dim Log-Mel Filter Bank energies
  - 25 ms window size, 10 ms shift
  - 3 sec cropped/padded utterances
- Data Augmentation:
  - RIRS(room impulse response), MUSAN (additive noise)
  - speed perturbation (0.9,1,1.1)
- Optimization
  - Loss Function : Angular Additive Margin Softmax (AAM-Softmax)
    - margin 0.3
    - scale : 32
  - SGD with momentum (0.2), Weight Decay (1e-4)
  - Cosine annealing Scheduler with linear warm-up schedule
- Cosine Similarity for scoring

Training Dataset

- VoxCeleb: 16 kHz sampled signal
  - Celebrity interviews in YouTube: 16 kHz sampled signal
  - 7000 speakers, Around 2000 hours of speech
  - Gender balanced data
- CN-Celeb: 16 kHz sampled signal
  - Chinese Speaker Recognition Corpus
  - 3000 Chinese celebrities and 1200+ hours of speech
  - Gender balanced data

Training Dataset

- VoxCeleb: 16 kHz sampled signal
  - Celebrity interviews in YouTube: 16 kHz sampled signal
  - 7000 speakers, Around 2000 hours of speech
  - Gender balanced data
- CN-Celeb: 16 kHz sampled signal
  - Chinese Speaker Recognition Corpus
  - 3000 Chinese celebrities and 1200+ hours of speech
  - Gender balanced data
- **3D-Speaker**: Both 16 kHz & 48 kHz sampled signal
  - Multi-device, Multi-distance, and Multi-dialect
    - Distance: 0.1 meter to 4 meter
    - Dialect: 13 Chinese Dialects
    - Devices: 8 devices
  - 10000 speakers, Around 1124 hours of speech
  - Combination of Far Field and Near-field speech

Performance Evaluation - Training Dataset influence

- Multichannel Data used as Validation
- ERes2Net model trained with different data sets
  - Performance on Channel 5 - Channel 5 is better with VoxCeleb data
  - Performance on Channel 5 - Channel 4 is better with 3D-speaker data

Enrollment - Test	VoxCeleb	CN-Celeb	3D Speaker
Channel 5 - Channel 5	**4.05**	5.2	7.57
Channel 5 - Channel 4	12.4	14.71	**10.77**

Table: Performance evaluation on RoboVox data

Performance Evaluation - Training Dataset influence

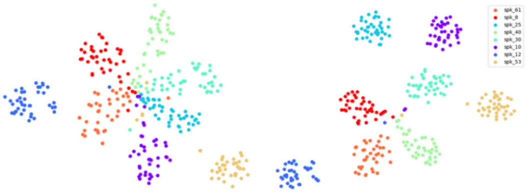
- Multichannel Data used as Validation
- ERes2Net model trained with different data sets
  - Performance on Channel 5 - Channel 5 is better with VoxCeleb data
  - Performance on Channel 5 - Channel 4 is better with 3D-speaker data

Enrollment - Test	VoxCeleb	CN-Celeb	3D Speaker
Channel 5 - Channel 5	**4.05**	5.2	7.57
Channel 5 - Channel 4	12.4	14.71	**10.77**

Table: Performance evaluation on RoboVox data

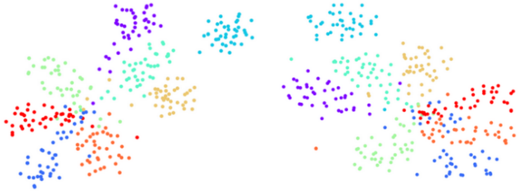
- VoxCeleb dataset contains more of **near-field** data
- 3D Speaker dataset includes a significant amount of **far-field** data

t-SNE Plots



3D Speaker (Channel 5)

Voxceleb(Channel 5)



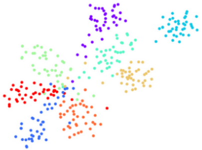
3D Speaker (Channel 4)

Voxceleb(Channel 4)

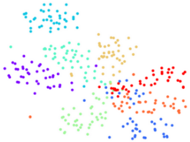
t-SNE Plots

3D Speaker (Channel 5)

Voxceleb(Channel 5)



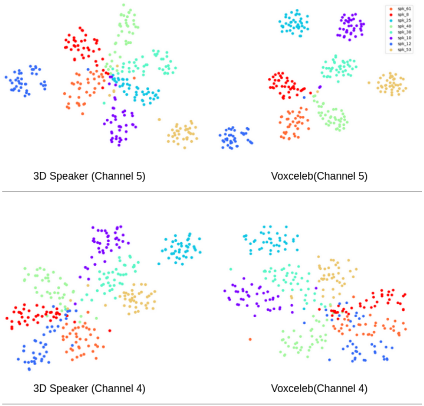
3D Speaker (Channel 4)



Voxceleb(Channel 4)

- No. of Speakers $= 8$
- Utterance per speaker $= 40$

t-SNE Plots



- No. of Speakers = 8
- Utterance per speaker = 40
- Better clustering
  - Voxceleb: Channel 5
  - 3D-Speaker: Channel-4

Performance Evaluation - Model influence

- Processing spectral magnitude features (Log-Mel filterbank energies)
  - ECAPA, SEResNet34V2, ERes2Net
- Direct processing of wave samples
  - wavLM self-supervised model pre-trained with 94k hours of speech.
  - ECAPA model as backend classifier on wavLM representations

Evaluation Type	wavLM	ECAPA	SEResNet34V2	**ERes2Net**
5 vs 5	6.16	7.14	7.2	**5.9**
5 vs 3	14.6	14.2	14.1	**11.2**
5 vs 4	17.7	17.1	16.8	**12.8**

Table: Performance evaluation on RoboVox data

Score Normalisation Results

- Adaptive Score Normalisation

Table: EER and EER (after score norm)

Model	EER	Avg MinDCF
ERes2Net Large	11.22	0.63
ERes2Net Large (after)	9.93	0.59

- Significant improvement observed in multi-channel data

Score Fusion

Table: EER and MinDCF Performance of different model on Multichannel Data

S.No	Model	EER	MinDCF (Day)	MinDCF (Night)	Avg MinDCF
1	ERes2Net-Large	10.77	0.40	0.99	0.69
2	ERes2Net-base	11.91	0.45	0.99	0.72
3	ResNet34	11.54	0.41	0.90	0.65
4	ECAPA-TDNN	11.76	0.44	0.99	0.72
5	CAM++	10.61	0.41	0.912	0.66
	Fusion(All)	9.33	0.35	0.92	0.63
	Fusion (1+3+5)	**9.26**	**0.35**	**0.809**	**0.57**

- Day ($P_{target} = 0.8$, $C_{Miss} = 1$, $C_{FA} = 20$)
- Night ($P_{target} = 0.01$, $C_{Miss} = 10$, $C_{FA} = 100$)
- Note: Scoring is done here after centering

Data Augmentation & Speech Enhancement

- Speech enhancement as a preprocessing step
  - Facebook denoiser, CMGAN, TVCN
  - Speech enhancement on clean data acts as identity transformation
  - Performance on noisy data: $\approx 13\%$
- Data augmentation to mimic the RoboVox challenge
  - Estimate channel 4 data from channel 5

$$x_4[n] = h[n] * x_5[n] + v[n] \tag{1}$$

  - Use $\hat{h}[n]$, $\hat{v}[n]$ and generate augmented noisy sample

$$\hat{x}_4[n] = \hat{h}[n] * x_5[n] + \hat{v}[n] \tag{2}$$

  - Include $\hat{x}_4[n]$ in the training data.
  - Wiener Filter Approach: 11.7% (+1%)

Summary & Future work

- Summary
  - Features: Log-Mel filter bank energies
  - Speaker embedding extractor: ERes2Net architecture
  - Training database: 3D-speaker database with 10,000 speakers
  - Score normalization & Speech enhancement as preprocessor
- Future work
  - Speech Enhancement: De-noising followed by de-reverberation models
  - Data Augmentation: Optimal estimation of channel 4 from channel 5
  - Distance and Device invariant features: Adversarial Training to learn it (metadata available in 3D-Speaker Dataset)

Thank You
Any Questions?