From SAD to ASR on the Fearless Steps Data

Bibash Thapaliya, Dheeraj Rajashekar Poolavaram, Saad Bin Abdul Mannan

Department of Communications Engineering – Paderborn University
Prof. Dr.-Ing. Reinhold Haeb-Umbach

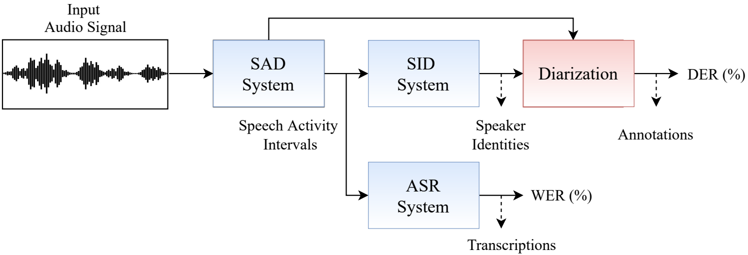
Nov 05, 2021

Table of contents

Introduction

Overview

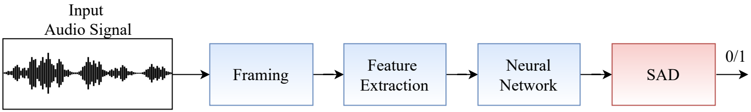
- The work has been divided into the following sub-tasks:
  1. Speech Activity Detection (SAD)
  2. Speaker Identity Detection (SID)
  3. Speaker Diarization (SD)
  4. Automatic Speech Recognition (ASR)



Speech Activity Detection

SAD System

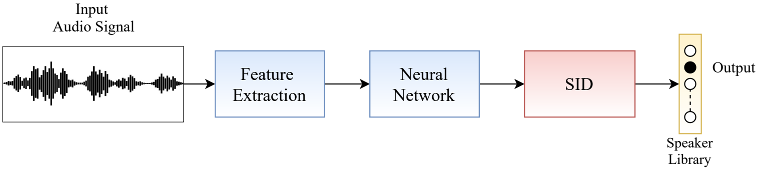
- Task of detecting speech from non speech
- An LSTM-based ResNet Architecture
- Achieves a **3.32%** DCF as compared to the Baseline of **1.12%**



Speaker Identity Detection

SID System

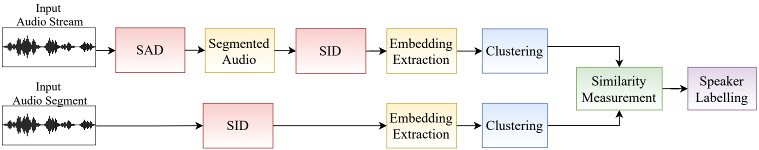
- Task of identifying a speaker from attributes of voices
- Trained using a Deep ResNet vector model



Speaker Diarization

Introduction

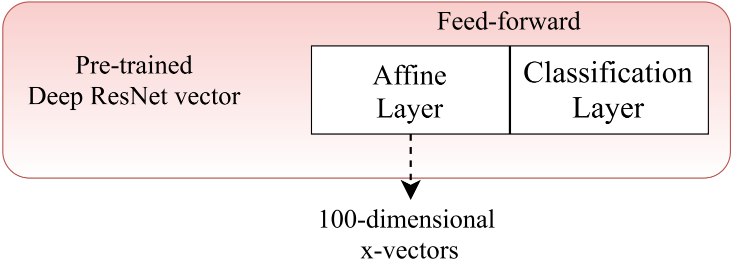
- *Who spoke when?*
- In this work, we present (un)supervised way of SD system



Speaker Diarization

x-vectors

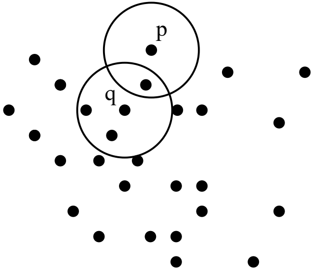
- Extracted from affine layer
- Variable length utterance to fixed dimensional embedding
- Discriminating feature vectors

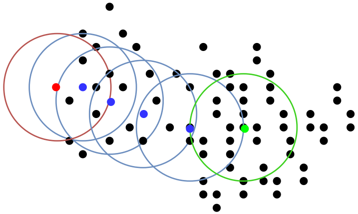


Clustering Algorithms

Feature clustering algorithms

- Density based algorithms
  1. DBSCAN
  2. Mean Shift



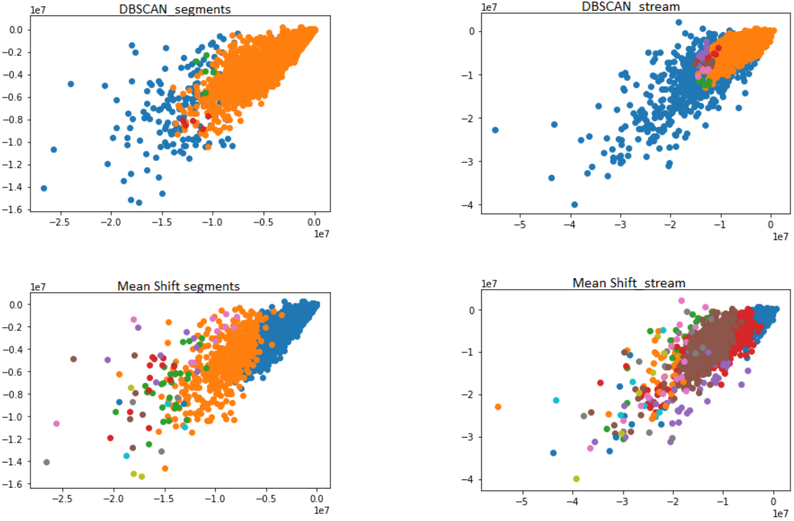


Measurements

Similarity Measurements

- Euclidean Distance
- Manhattan Distance
- Cosine Similarity
- Jaccard Similarity

Results: DBSCAN and Mean Shift

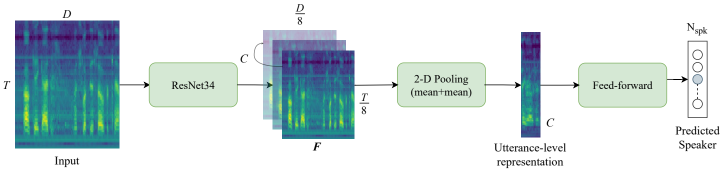


Reasoning

Why x-vector approach didn't work?

- Cluster centers of different classes not distinguishable
- Quality of x-vectors from pre-trained SID model
- Imbalanced dataset

Speaker Identity Detection



Evaluation

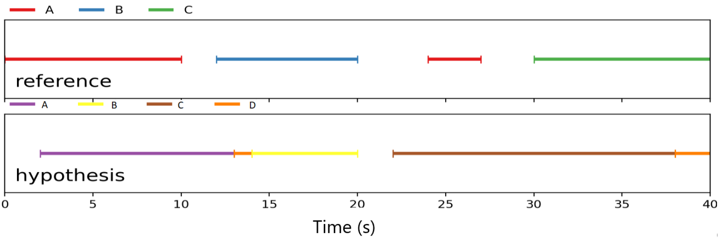
- Log-Mel filterbanks instead of Mel filterbanks
- Achieves a **91.65%** as compared to the Baseline of **90.78%**

Model	Accuracy	Weighted F1-Score	Top-5 Accuracy
Deep ResNet Vector (earlier)	67.71	67.01	88.70
Deep ResNet Vector (now)	**76.89**	**75.53**	**91.65**
Baseline	-	-	90.78

Speaker Diarization

Annotations

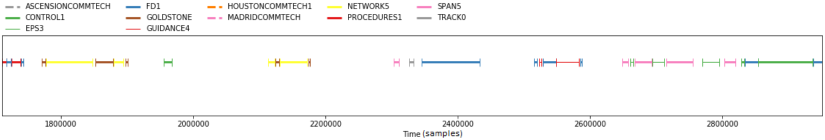
- Temporal speech activities and the corresponding speaker labels
- $\text{DER} = \frac{\text{false alarm} + \text{missed detection} + \text{confusion}}{\text{total duration}}$



Speaker Diarization

Reference Annotations

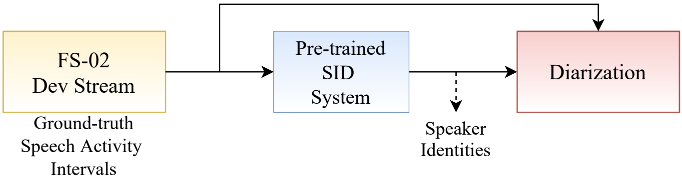
- FS-02 ground-truth speech intervals and corresponding speaker identities
- 5.579 hours of speech content



Speaker Diarization

Hypothesis Annotations-I

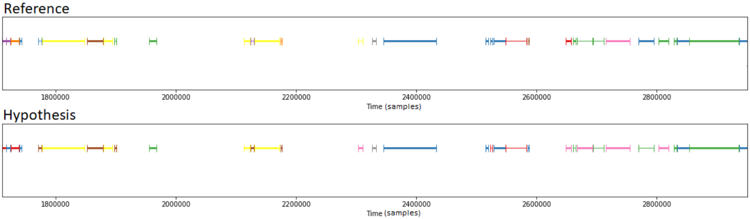
- Groundtruth FS-02 Dev stream audio
- Speaker predictions from pre-trained Deep ResNet Vector



Speaker Diarization

Hypothesis Annotations-I

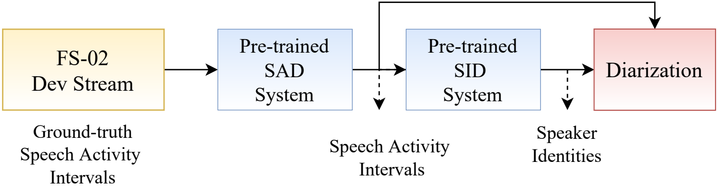
Metrics	Results (%)
False Alarm	0
Missed Detection	0
Confusion	19.39
Correct	80.60
DER	**19.39**



Speaker Diarization

Hypothesis Annotations-II

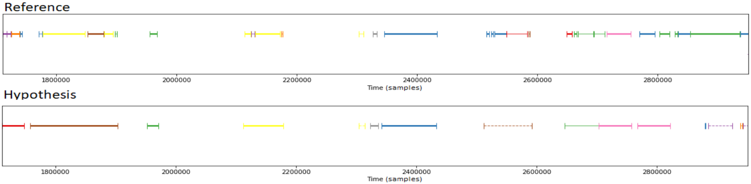
- FS-02 Dev stream audio through pre-trained SAD system
- Speaker predictions from pre-trained Deep ResNet Vector



Hypothesis Annotations-II

Metrics	Results (%)
False Alarm	23.76
Missed Detection	0.22
Confusion	32.80
Correct	66.97
DER	**56.79**

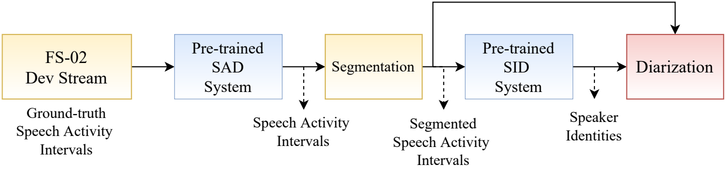
- Around 50% intervals as compared to ground-truth FS-02



Speaker Diarization

Hypothesis Annotations-III

- SAD speech intervals segmented

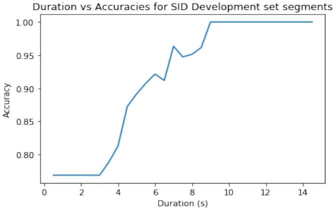


Hypothesis Annotations-III

Metrics	Results (%)
False Alarm	23.70
Missed Detection	0.22
Confusion	49.32
Correct	50.44
DER	**73.26**

SID performance

Longer segments, better predictions



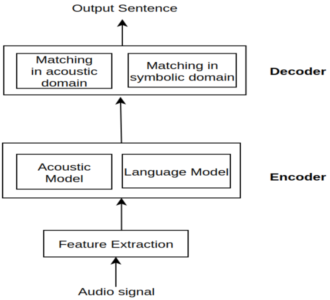
Automatic Speech Recognition

Introduction

- Automatic Speech Recognition (ASR) is a technology where speech signal is converted into text
- In this work, we use an end-to-end ASR system using the Transformer architecture

Main stages involved in ASR

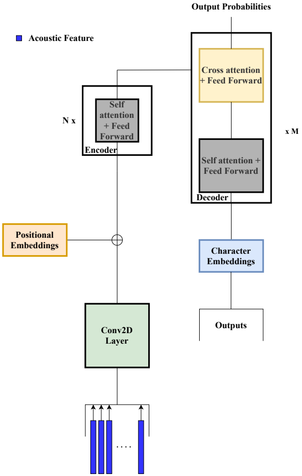
- Feature extraction
- Encoding
- Decoding



Automatic Speech Recognition

Architecture

- The Transformer architecture consists of an encoder-decoder network
- Uses Attention mechanism to find features relevant to the context



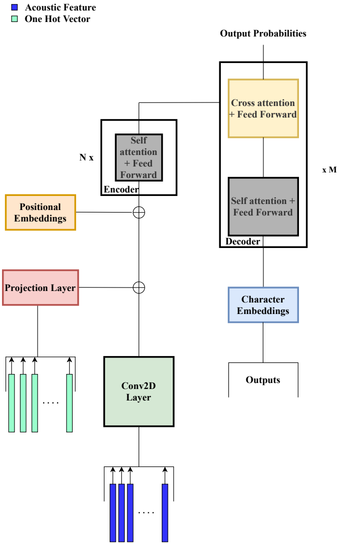
Automatic Speech Recognition

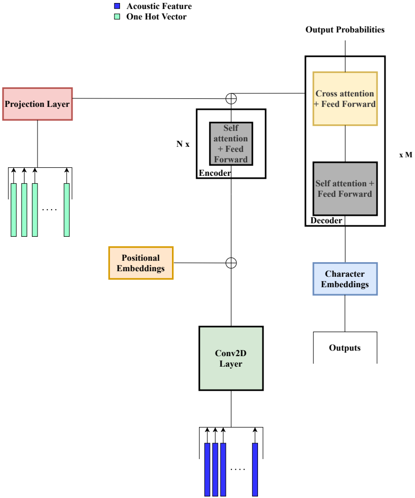
Speaker Adaptation

Incorporates speaker information in the model

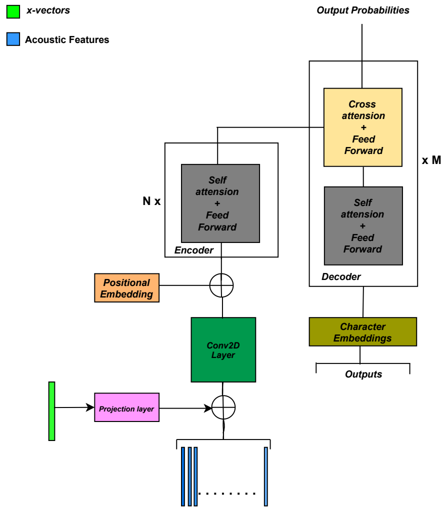
- one-hot speaker embedding
- x-vector speaker embedding

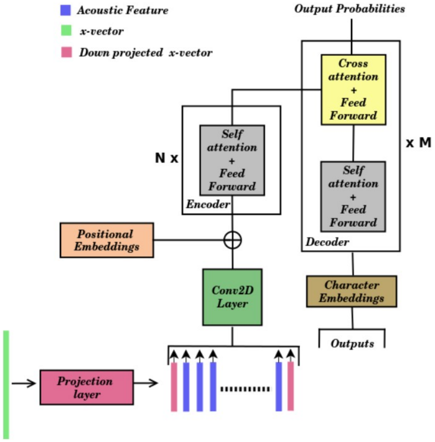
Automatic Speech Recognition



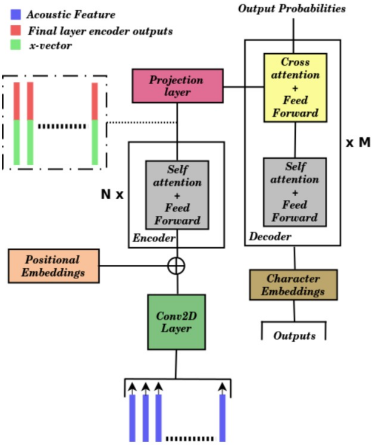


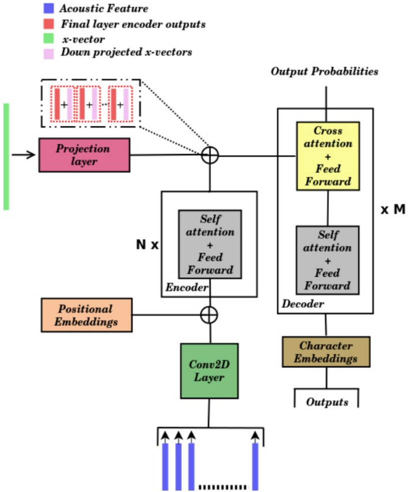
Automatic Speech Recognition





Automatic Speech Recognition





Automatic Speech Recognition

Evaluation

- Word Error Rate (WER): To determine the performance of the system
- Adam optimizer with a CTC loss function

Dataset	Number of Original Data	Number of Speakers	Number of Data used for one-hot Speaker Embedding	Number of Data used for x-vector Embedding
Train set	35,474	256	33,345	30,978
Dev set	9,203	201	9,029	8,462

Automatic Speech Recognition

Model Implementation (with one-hot vectors)	WER (%)
Base Model	**32.9**
Addition to Encoder	33
Addition to Embedding	**32.9**
Concatenation to acoustic features	44

Model Implementation (with x-vectors)	WER (%)
Base Model	38.2
Addition to Encoder	39.6
Concatenation to Encoder	39.4
Stacking on Top and Bottom of acoustic features frame	37.7
Addition to acoustic features	38.3

Conclusion

- Work split into sub-tasks - SAD, SID, SD and ASR.
- Fearless Steps Database from NASA's Apollo-11 space mission.
- The achieved results along with the architecture:
  1. SAD - ResNet-LSTM, DCF of 3.32%
  2. SID - Deep ResNet Vector, Top-5 Accuracy of 91.65%
  3. SD - SAD+SID, DER of 56.78%
  4. ASR - Transformer Model, WER of 32.9%
- *who spoke when and what was the content of the speech utterance.*

Future Scope

- Speaker Change Detection
- More datasets for a more robust system

Thank you for listening!

Questions?