PADERBORN UNIVERSITY
The University for the Information Society

Communications Engineering

From SAD to ASR on the Fearless Steps Data

Bibash Thapaliya, Dheeraj Rajashekar Poolavaram, Padmanaban Krishnan, Saad Bin Abdul Mannan, Vivek Kandimalla

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

May 28, 2021

Table of contents

Introduction

Overview

- The work has been divided into three sub-tasks:
  1. Speech Activity Detection
  2. Speaker Identity Detection
  3. Automatic Speech Recognition
- The speech should be distinguished from the non-speech content.
- The speaker identity of the detected speech has to be determined.
- The speaker identities are incorporated in the recognition system to achieve better transcription results.

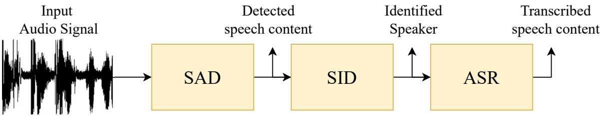


Figure: Speech Recognition System

Database

Audio Data

1. Fearless Steps-02 Corpus from NASA Apollo space missions
2. 100 hours of data from 19000 hours
3. Task dependent datasets

Data-set	Stream (min)	No. of examples	No. of segments	No. of speakers
Train	30	125	27336	218
Dev	30	30	6373	218

Table: Fearless Steps-02 audio data-sets

Speech Activity Detection

Introduction

- Speech Activity Detection is a task of detecting speech from non speech.
- In this current work, we present an NN-based SAD system.

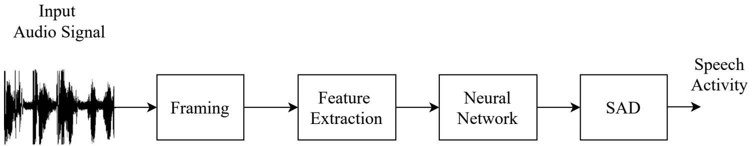


Figure: Work flow of SAD system

Data Preparation

- MFCC features extracted for 0.5 seconds audio segment.
- 13 Cepstral coefficients.
- 23 time frames.
- Targets estimated for 0.5 seconds.

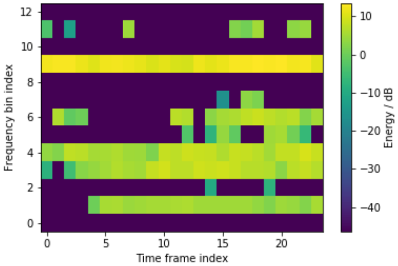


Figure: Mel-Spectrogram of 0.5 seconds

Speech Activity Detection

Data Preparation

- MFCC features extracted for 4 seconds audio segment.
- 13 Cepstral coefficients.
- 199 time frames.
- Frame-wise targets.

Aspects of Segmentation

- Targets.
- Precision of Detection.

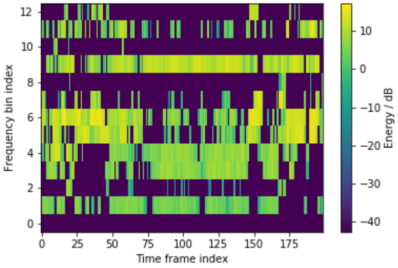


Figure: Mel-Spectrogram of 4 seconds

Simple CNN Architecture

- Compare the result with baseline paper.
- CNN model to understand the motive of the ResNet model.
- The output of the last layer contains 128 channels.
- Max pooling operation performed.
- Fully connected layer with Sigmoid activation function.

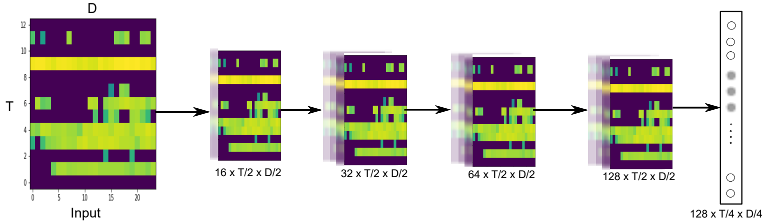


Figure: CNN based SAD

Architecture

ResNet-LSTM based SAD

- ResNet18: To overcome gradient vanishing problem.
- 1-D mean statistics pooling.
- Bi-LSTM: Continuous nature of an audio data.
- Binary Cross Entropy with SGD optimizer.

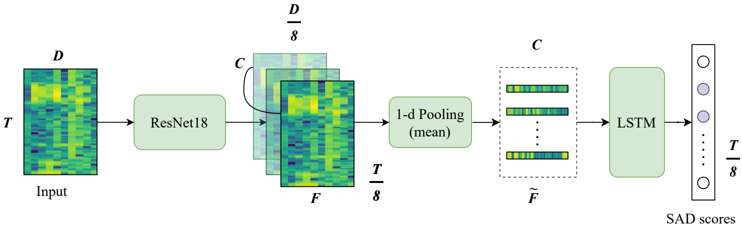


Figure: ResNet-LSTM based SAD. This image is adapted from baseline

Evaluation

Results

- $\text{DCF} = 0.75 \times \text{FNR} + 0.25 \times \text{FPR}$
- DCF with optimal threshold.
- Better performance with frame wise estimation.

Model	Segmentation (s)	DCF (%)
CNN	0.5	8.23
	4	2.44
ResNet	0.5	7.82
	4	3.32
Baseline	-	1.123

Table: Results of SAD architectures

Speaker Identity Detection

Introduction

- Speaker Identity Detection is the task of identifying a speaker from attributes of voices.
- In this work, we present a NN-based SID system.

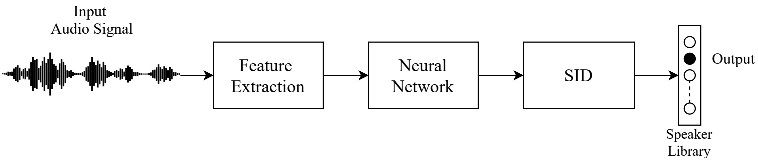


Figure: Block diagram of SID

Speaker Identity Detection

Data Preparation

- 64-dimensional Mel-energy filterbanks extracted from audio segments = Features.
- Every unique speaker assigned to a one-hot binary vector = Targets.

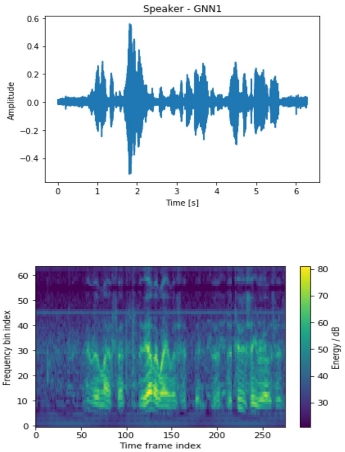


Figure: Example audio signal and its corresponding Mel-spectrogram.

Speaker Identity Detection

Architectures

- Simple CNN models to understand the motive behind Deep ResNet Vector-based system.
- Fully Connected Linear layers with a Softmax activation function.

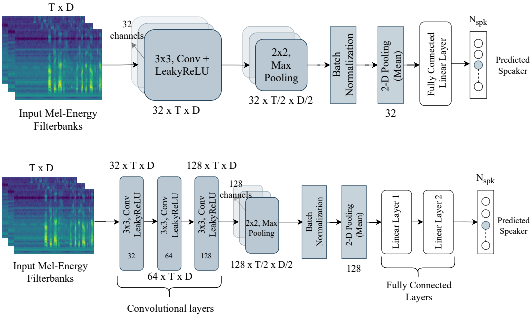
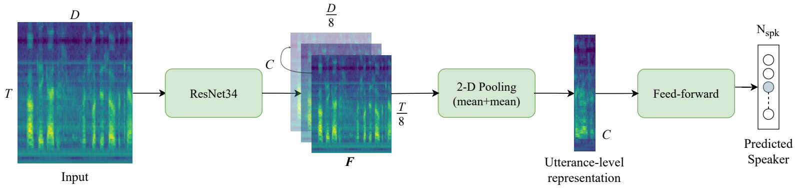


Figure: Simple Convolutional Networks

Speaker Identity Detection

Deep ResNet Vector-based system

- ResNet34: Channel widths set to $\{32, 64, 128, 256\}$.
- 2-D mean statistics pooling.
- Network Output: Predicted speaker.



Speaker Identity Detection

Evaluation

- SGD optimizer with Cross Entropy loss function.
- A learning-rate scheduler employed.
- Top-5 Accuracy: A classification to be correct if any of the five predictions match the target.
- Achieves a **88.70%** as compared to the Baseline of **90.78%**
- Use of optimal thresholding for data balancing as future work.

Model	Accuracy	Macro F1-Score	Top-5 Accuracy
Simple Convolutional Network 1	14.27	6.33	38.90
Simple Convolutional Network 2	20.76	10.87	46.94
Simple Convolutional Network 3	31.30	18.90	58.16
Deep ResNet Vector	61.94	45.81	82.30
Deep ResNet Vector (LR)	**67.71**	**52.71**	**88.70**

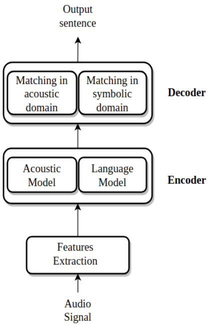
Automatic Speech Recognition

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

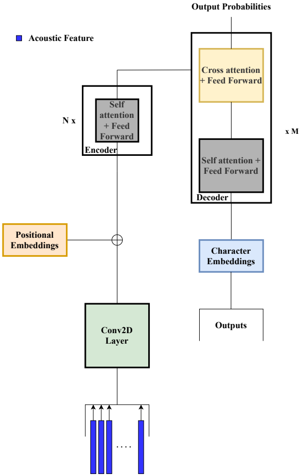
Speaker Adaptation

- Incorporates speaker information in the model.
- In this work, we used one-hot speaker embeddings to incorporate speaker information.

Automatic Speech Recognition

Architecture

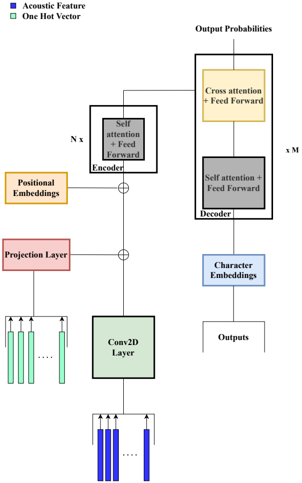
- The Transformer architecture consists of an encoder-decoder network.
- Uses Attention mechanism to find features relevant to the context.



Automatic Speech Recognition

Embed

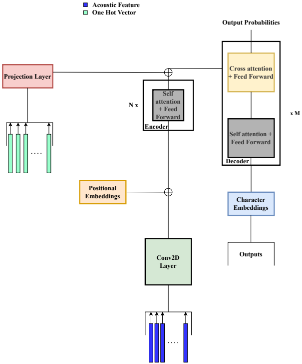
The speaker vectors are added to the encoder's input.



Automatic Speech Recognition

Encode

The speaker vectors are added to the final encoder's output.



Automatic Speech Recognition

Evaluation

- Word Error Rate (WER) is used to determine the performance of the system.
- Adam optimizer with a CTC loss function
- Use of speaker identities from the SID task for future work.

Model Implementation	WER %
Baseline without LM	41.3
Baseline	36.8
Embed [1]	89
Encode	33

Future Work

[1] The Embed model was retrained after report submission and got a WER of 32%

Conclusion

- Work split into three sub-tasks - SAD, SID, and ASR.
- Fearless Steps Database from NASA's Apollo-11 space mission.
- Several architectures used to examine the three tasks.
- The achieved results along with the architecture:
  1. SAD - ResNet-LSTM, DCF of 3.32%
  2. SID - Deep ResNet Vector, Top-5 Accuracy of 88.70%
  3. ASR - Embed, WER of 32%.

Future Scope

Integration of the sub-systems.

Thank you for listening!

Questions?