Poster number: 2198

A Multitask Learning Framework for Speaker Change Detection with Content Information from Unsupervised Speech Decomposition

[1]Hang Su, [1]Danyang Zhao, [1]Long Dang, [2]Minglei Li, [1]Xixin Wu, [1]Xunying Liu, [1]Helen Meng

[1]The Chinese University of Hong Kong [2]Huawei Cloud

1. Introduction

Task

- Determining speaker change time boundaries in recorded speech

Motivation

- Speaker Change Detection (SCD) benefits speaker diarization, speaker tracking and transcribing audio with multiple speakers
- Current state-of-the-art SCD system may still improve
  - Speaker information in training data and content information in dialog have not been fully utilized

Goal

- Improve the state-of-the-art SCD system in terms of :
  - Utilize speaker information in training data
  - Add content information extracted from discussion dialog audio

2. Baseline System

Baseline System (see Figure 1):

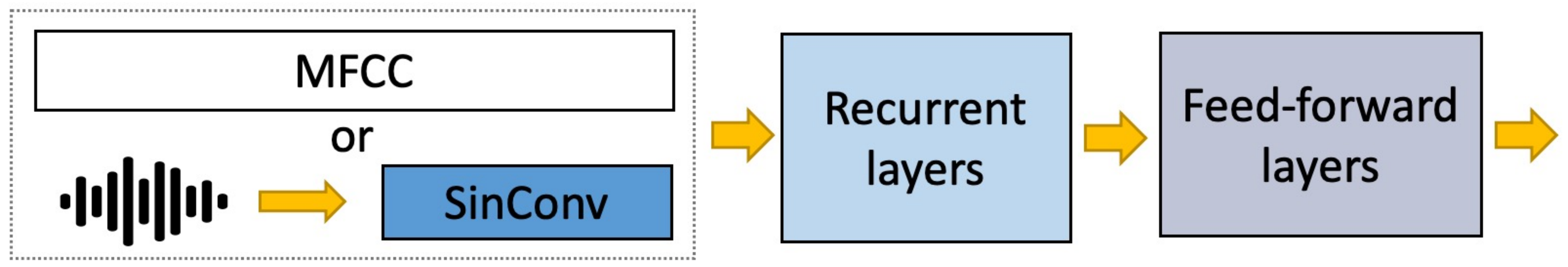

Figure 1: Model architecture of the baseline system

Training

- Figure 2 shows training process
  - Splitting audio sequence
  - Add random noise
  - Feed into model
  - Predict 1/0 (change/not change)
  - Cross Entropy loss

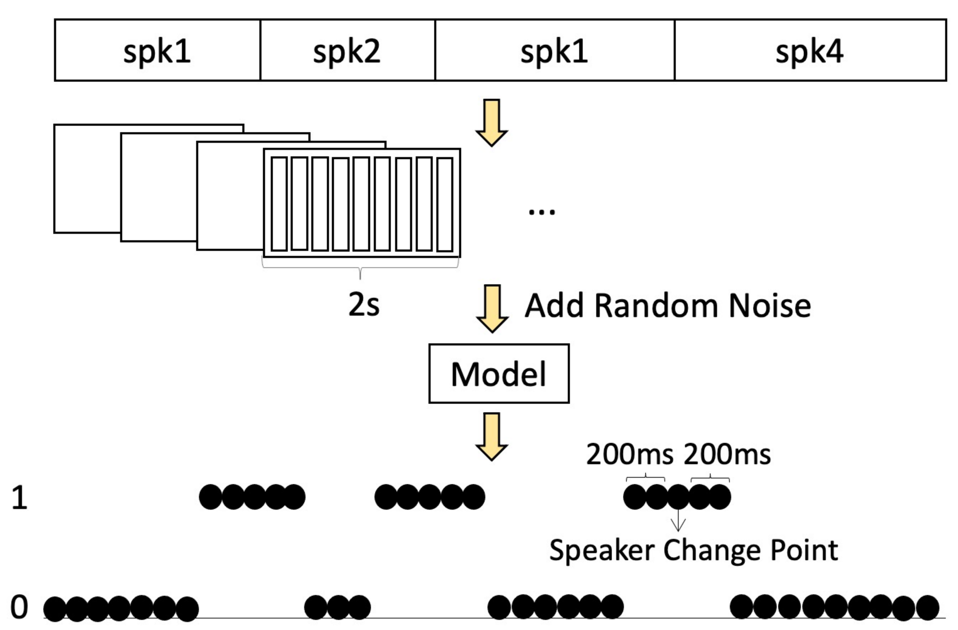

Figure 2: Training process for speaker change detection

Prediction

- Figure 3 shows the prediction process
  - Splitting audio sequence
  - Feed into the trained model
  - Predict 1/0 (change/not change)
  - Take the average
  - Decide the boundary

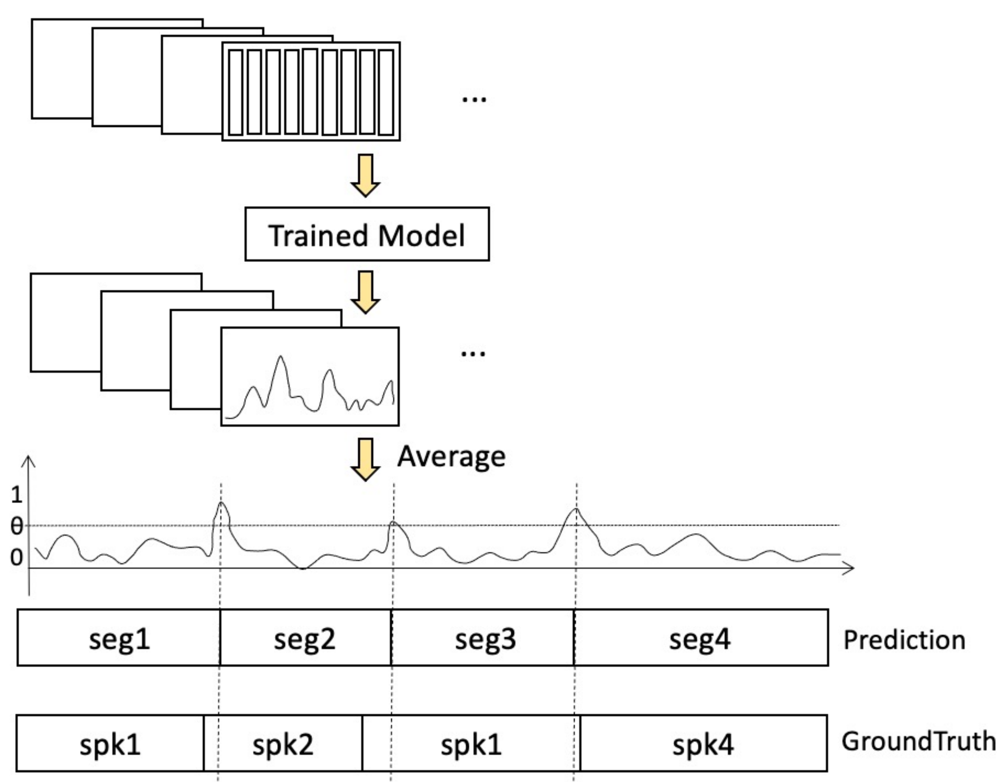

Figure 3: Prediction process for speaker change detection

3. Proposed Approach

Proposed Approach (see Figure 4)

Multitask Learning

- Goal: Utilize speaker information
- Add a "Speaker Branch"
  - Predict speaker (Cross Entropy loss)
  - Distinguish speakers (Triplet loss)

Unsupervised Speech Decomposition

- Goal: Add content information
- Pretrain a decomposition model
  - Decompose spoken information into pitch, rhythm, timbre and content
  - Encoder → Decoder with MSE loss

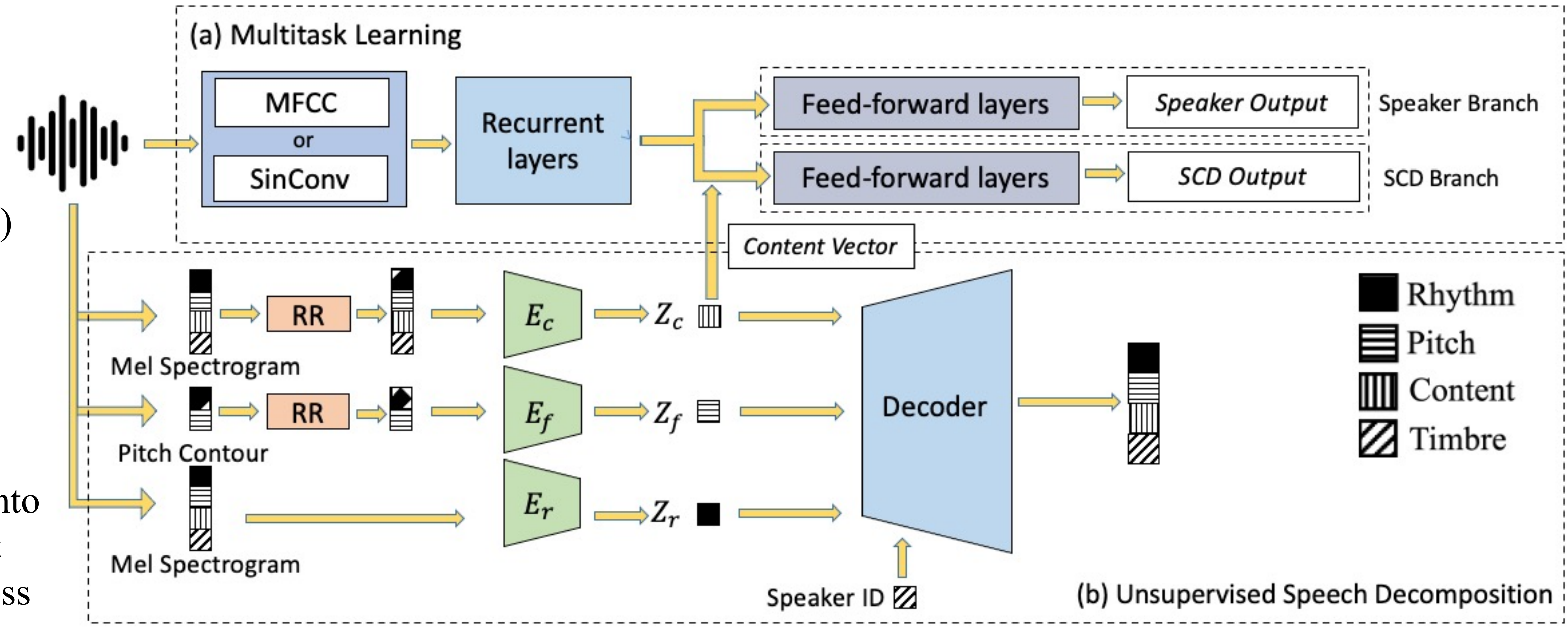

Figure 4: Framework of proposed approach

Training and Prediction

- Splitting audio sequence
- Feed into both Multitask Learning model and pre-trained Unsupervised Speech Decomposition (USD) model
- Obtain content vector from USD model, then feed into Multitask Learning model
- Predict 1/0 (change/not change) in Multitask Learning model

4. Experiments

Dataset – AMI corpus

- Collection of conversational recordings in meeting domain
- 4~5 speakers in each conversation
- 70 hours for training, 15 hours for validation, 15 hours for test

Evaluation Metric

- Coverage : r: reference segment; h: hypothesis segment

$$coverage(R,H) = \frac{\sum_{r \in R} max_{h \in H} |r \cap h|}{\sum_{r \in R} |r|}$$

- Purity : dual metric of coverage where role of h and r interchanged
- F1 : harmonic average of coverage and purity

Results

- Table 1 shows the results of using MFCC as the input
- Table 2 shows the results of using waveform as the input

	Validation			Test		
	Purity	Coverage	F1	Purity	Coverage	F1
Baseline	85.01	79.90	82.27	86.54	80.72	83.53
Pretrain+finetune	85.08	80.78	82.87	87.04	81.18	84.01
Multitask (spk id)	85.07	79.98	82.44	86.84	82.97	84.34
Multitask (triplet)	85.02	81.14	83.03	86.04	83.31	84.65
Triplet + content	85.04	81.68	**83.33**	86.16	84.56	**85.35**

Table 1: Results of using MFCC as the input

	Validation			Test		
	Purity	Coverage	F1	Purity	Coverage	F1
Baseline	85.38	89.49	87.39	85.62	89.71	87.62
Pretrain+finetune	85.00	90.51	87.67	85.16	90.92	87.95
Multitask (spk id)	85.00	91.74	88.24	85.61	91.04	88.24
Multitask (triplet)	85.26	91.49	88.27	85.66	91.02	88.26
Triplet + content	85.00	91.92	**88.32**	85.68	91.75	**88.61**

Table 2: Results of using waveform as the input

5. Conclusions

- Utilize speaker information with proposed multitask learning architecture to improve performance of SCD
- Add spoken content vectors extracted from pre-trained unsupervised speech decomposition model to further improve performance of SCD task
- Proposed approach achieved new state-of-the-art result on the AMI dataset for SCD task

6. Acknowledgements

This work is partially supported by the CUHK TDLEG Grant (2016-2019) and a grant from the CUHK Stanley Ho Big Data Decision Analytics Research Centre.