IEEE ICASSP 2021 No. 3488

Development of the CUHK Elderly Speech Recognition System for Neurocognitive Disorder Detection Using the DementiaBank Corpus

Zi Ye, Shoukang Hu, Jinchao Li, Xurong Xie, Mengzhe Geng, Jianwei Yu, Junhao Xu, Boyang Xue, Shansong Liu, Xunying Liu, Helen Meng

{zye,skhu,jcli,mzgeng,jwyu,jhxu,byxue,ssliu,xyliu,hmmeng}@se.cuhk.edu.hk xr.xie@link.cuhk.edu.hk

The Chinese University of Hong Kong, Hong Kong, China

1. Introduction

❑ Motivation

- Early diagnosis of Neuro-cognitive Disorder (NCD), e.g. Alzheimer Disease, is crucial for timely treatment and intervention
- Automated speech technologies for large scale screening
- Most previous works rely on manually generated transcripts
- Automatic speech recognition (ASR) systems targeting elderly speech is essential

❑ Challenge

- Large mismatch between normal speech and elderly speech with increased voice perturbation, articulatory imprecision, etc.
- Lack of large amounts of elderly speech recordings for NCD

❑ Our Work: CUHK Elderly Speech Recognition System

- ASR system for automatic NCD tests built on the DementiaBank Pitt corpus incorporating a series of modelling techniques (Fig. 1)
  - Re-segmentation, augmentation, adaptation of audio data
  - Transformer language model combined with 4-gram
- Evaluation with word error rate (WER) and NCD detection results

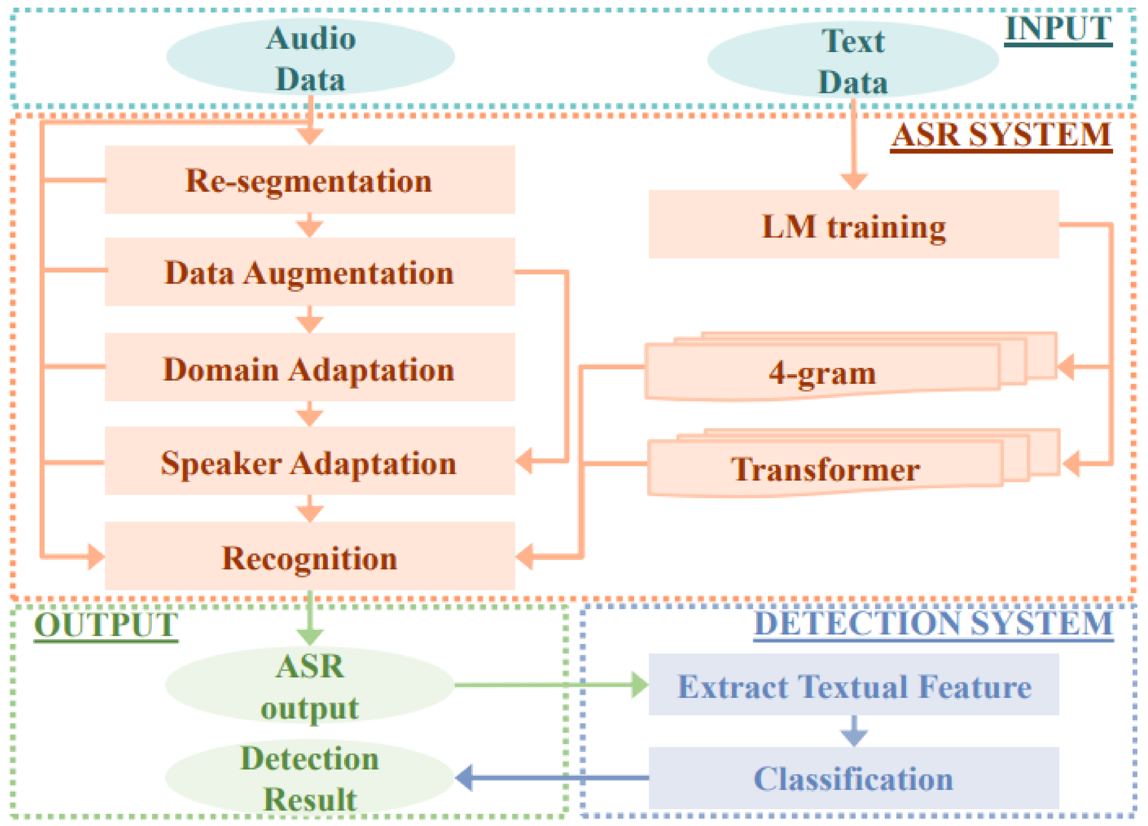


Fig. 1: The development stages of the speech recognition system combined with NCD detection system introduced in this paper

2. Task Description

❑ Audio Data from DementiaBank Pitt corpus (~ 33 hrs)

- Containing Interview recordings between 292 elderly participants and their investigators as one of the largest public dataset for NCD
- Partitioned based on the public ADReSS corpus for consistency

❑ Text Corpora for Language Model (vocabulary of ~3.6k words)

- Small 4-gram built with DementiaBank Pitt corpus only
- Large 4-gram built on the transcriptions from LDC Switchboard and Fisher corpus, LDA Gigaword corpus, other related corpora (Holland, Kempler, Lanzi) from DementiaBank in addition to Pitt

❑ Baseline System

- Lattice-free maximum mutual information ((LF-MMI) trained factored time delay neural network (TDNN) acoustic models

3. ASR System Development

❑ Audio Re-segmentation

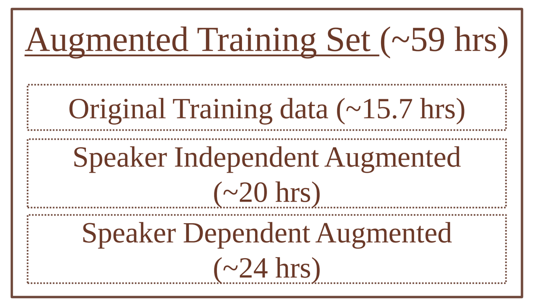
- GMM-HMM model (2k triphone states with 32 Gaussian per state) was used for alignment, then excessive silences (>200ms) at the beginning or at the end were removed and the utterances with long internal silence (>1s) were split
- About 40% of the data was removed as shown in Table. 1
- WER obtained from the resulting system (Sys.5) compared to the baseline system (Sys. 4) in Table. 2
  - 1.87% overall reduction
  - 5.88% reduction for participants in the evaluation set

Table 1: The Number of participants (PAR) and the number of hours of the speech of the participant and the investigator (INV), before (Col. 3-5) and after (Col. 6-8) audio re-segmentation, in the Pitt corpus

	#PAR	Before Resegmentation			After Resegmentation		
		PAR	INV	Total	PAR	INV	Total
Train	244	17.65h	9.51h	27.16h	9.71h	6.03h	15.74h
Dev.	43	2.96h	1.79h	4.75h	1.40h	1.12h	2.52h
Eval.	48	0.88h	0.19h	1.07h	0.53h	0.09h	0.62h

❑ Data Augmentation

- Speed perturbation based data augmentation
  - Speaker independent factors for participant {0.9, 1.0, 1.1}
  - Speaker dependent factors for investigator {0.84, 0.95, 1.0, 1.08, 1.27}
- The training dataset was expanded to about 59 hrs from 15.74 hrs by a factor of 4
- Absolute WER reduction from the system before augmentation (Sys. 5) to the resulting system (Sys. 6) in Table. 2
  - 2.94% overall reduction
  - 2% reduction for participants in the evaluation set



❑ Cross-domain Adaptation

- 1000-hr LibriSpeech corpus trained LF-MMI TDNN was Cross-domain Bayesian adapted to the 59-hr augmented Pitt data
- Resulting System (Sys. 7) decreased overall WER by 1.04% absolute (1.7% absolute for participants in evaluation set) as shown in Table. 2 compared to the system without adaptation (Sys. 6)

❑ Speaker adaptation

- Learning Hidden Unit Contribution (LHUC) speaker adaptation with Bayesian estimation to account for the model uncertainty caused by the limited data
- Absolute WER reduction obtained from the resulting systems compared to the system without adaptation (Sys. 6) in Table. 2
  - 2.7% (2.89% for participants in eval set) with speaker adaptation (Sys. 8)
  - +0.3% for participants in eval set combined with domain adaptation (Sys. 9)

❑ Language Model (LM)

- Transformer LM, trained on 2.4M-word transcriptions from Pitt, Switchboard and Fisher and then Bayesian adapted to the Pitt transcripts, before being used to rescore the 4-gram LM decoded output
- Resulting System (Sys.10) further reduced the overall WER by 0.92% absolute (0.65% absolute for participants in evaluation set) compared to Sys.9 in Table. 2

4. System Performance

❑ ASR Performance

Table 2: WER(%) obtained using the baseline systems with or without i-Vector and optionally using the small or large 4-gram (Sys. 1-4); WER(%) of the systems improved through different stages: audio re-segmentation (Sys. 5); data augmentation (Sys. 6); domain adaptation (Sys. 7); speaker adaptation (Sys. 8-9); transformer LM re-scoring (Sys. 10)

Sys.	I-Vector	Audio Re-segment	Speed perturb	Bayesian TDNN Adaptation: Domain	Bayesian TDNN Adaptation: Speaker	Language Model	Dev. PAR	Dev. INV	Eval. PAR	Eval. INV	All
1	×	×	×	×	×	small 4-gram	53.48	22.65	43.54	29.06	38.53
2	√					small 4-gram	52.93	23.16	45.96	27.62	38.87
3	×					large 4-gram	52.87	23.15	43.00	28.18	38.37
4	√					large 4-gram	51.70	23.13	44.89	26.85	38.18
5	√	√	×	×	×	large 4-gram	51.51	21.57	39.01	20.64	36.31†
6	√	√	√	×	×	large 4-gram	46.76	19.97	37.01	18.20	33.37†
7	√	√	√	√	×	large 4-gram	45.56	19.19	35.31	19.31	32.33†
8				×	BLHUC-SAT		42.95	18.24	34.12	17.87	30.67†
9				√	BLHUC-SAT		43.74	18.06	33.82	**16.65**	30.82†
10	√	√	√	√	BLHUC-SAT	large 4-gram + Transformer	**42.12**	**17.61**	$\boxed{\textbf{33.17}}$	17.20	$\boxed{\textbf{29.90}}^{\dagger}$

❑ NCD Detection Performance

- NCD Detection Task
  - Textual features extracted from the baseline or best recognition outputs (Sys. 4 and Sys. 10) for Pitt evaluation set, including
    - ✓ 1035-dim TF-IDF features (sparse) encoding word frequency information
    - ✓ 768-dim BERT based features (dense) may capturing additional long-range contextual information
  - Support Vector Machines (SVM) based detection system
- Detection Results (Table. 3)
  - Detection accuracy improved from 0.79 to 0.88 with WER reduced from 44.89% to 33.17% for participants in evaluation set using BERT based features
  - The best ASR system outputs (Sys. 10) gave NCD detection accuracy comparable to that obtained using manual (ground truth) speech transcription

Table 3: ASR WER(%) and NCD detection results in terms of accuracy, precision, recall, F1 score and area under curve (AUC) obtained using the manual transcripts, the baseline or the best ASR outputs (Sys. 4 & 10 in Table 2) for participants of the evaluation set

Sys.	Feature	WER	Acc.	Pre.	Rec.	F1	AUC
Manual	TF-IDF	N/A	0.71	0.73	0.67	0.70	0.83
4	TF-IDF	44.89	0.69	0.74	0.58	0.65	0.85
10	TF-IDF	33.17	0.69	0.74	0.58	0.65	0.82
Manual	BERT	N/A	**0.88**	**0.91**	0.83	0.87	0.89
4	BERT	44.89	0.79	0.72	**0.96**	0.82	0.87
10	BERT	33.17	$\boxed{\textbf{0.88}}$	0.82	**0.96**	**0.88**	**0.92**

4. Conclusion

❑ Tailored ASR system for elderly speech for NCD detection

- Overall WER reduction of 11.72% absolute (26.11% relative) for elderly participants in evaluation set was obtained in system development
- Comparable NCD detection results to that using manual transcription

❑ Future Plan

- Analysis on individual ASR modelling techniques' effect on NCD detection
- Tighter integration between the ASR system and NCD detection model

This research is supported by Hong Kong RGC GRF grant No.14200218, 14200220, TRS T45-407/19N, Innovation & Technology Fund grant No. ITS/254/19, and SHIAE grant No. MMT-p1-19.