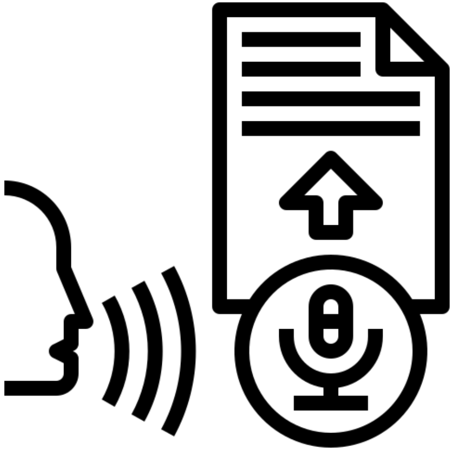
Comparing data augmentation and training techniques to reduce bias against non-native accents in hybrid speech recognition systems

Yixuan Zhang

The Bias

: An ASR tends to generate more accurate predictions on certain groups within a dataset while making more errors on others.



Possible Origins of The Bias:

- Imbalanced training data set
- Mismatch between the test data and the training data
- Vocal characteristics of certain speaker group
- Specific architectures & algorithms used during ASR system development

Group	Read			HMI		
	F	M	Avg	F	M	Avg
DC	34.8	35.7	35.3	43.5	43.3	43.4
DT	16.5	20.1	18.4	34.4	36.2	35.3
DOA	22.3	27.9	24.2	37.8	42.5	39.5
AvgD	24.4	28.1	$\boxed{26.1}$	38.4	41.7	$\boxed{39.8}$
NNC	54.3	55.9	55.1	60.9	62.1	61.6
NNA	57.3	56.1	56.9	61.2	61.5	61.3
AvgN	55.8	56.0	$\boxed{55.9}$	61.1	61.7	$\boxed{61.4}$
Avg	35.4	37.2	36.2	46.5	49.0	47.5

Relate Work

Table.1 WERs per age group on JASMIN-CGN, with TDNN-BLSTM acoustic model [1]

[1]: Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. *Quantifying Bias in Automatic Speech Recognition*. 2021. arXiv preprint arXiv:2103.15122

To offset the bias brought by the lack of accented speech data:

- ➔ ***Increasing the amount*** **of non-native speech**
  data augmentation, synthesize speech, ...
- ➔ ***Improving the learning efficiency*** **of the model when learning from the limited non-native speech resources**
  transfer learning, pre-trained model, ...

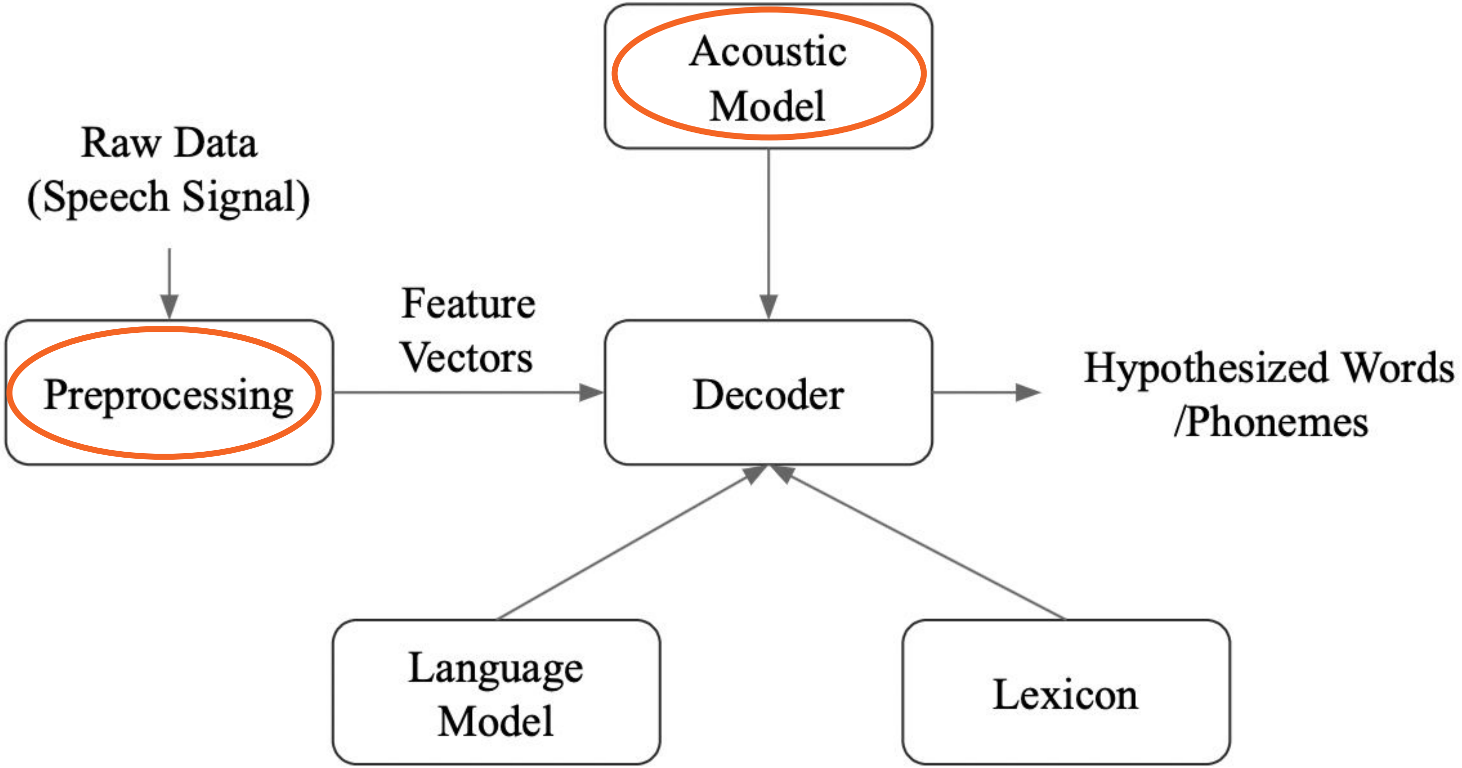
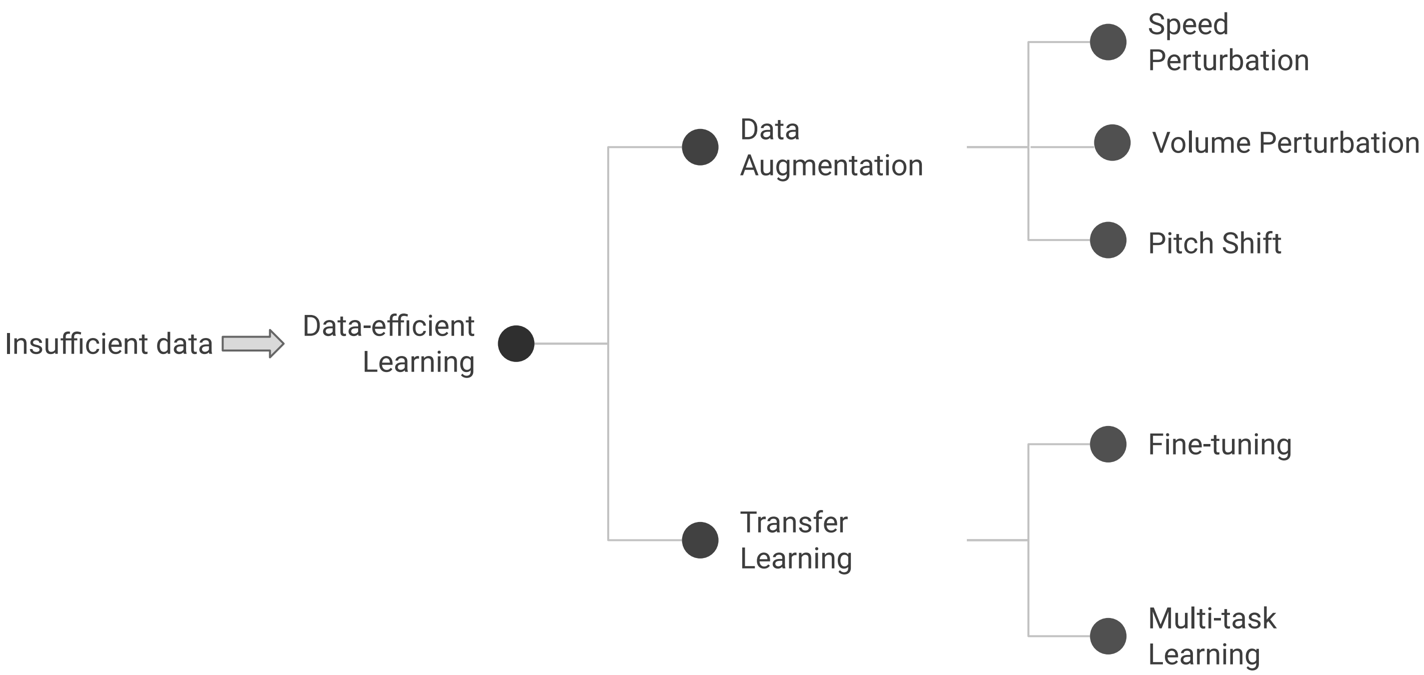


Fig.1. The architecture of a hybrid ASR system

Experimental Setup



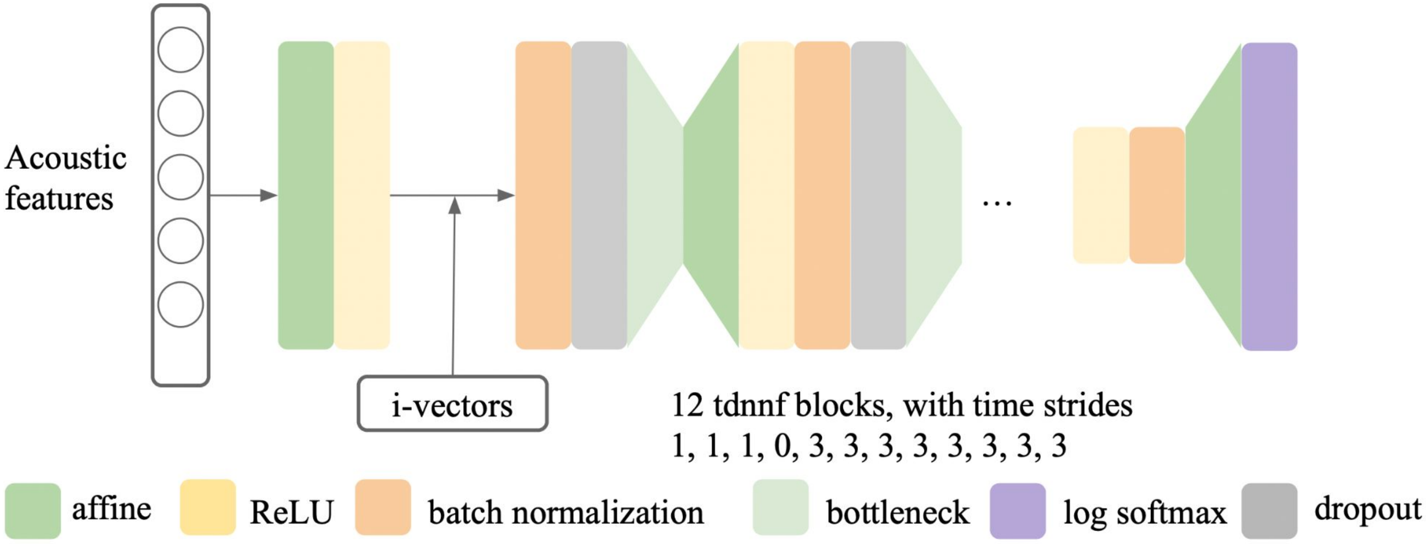


Fig.2. The architecture of the TDNNF AM

Acoustic Model

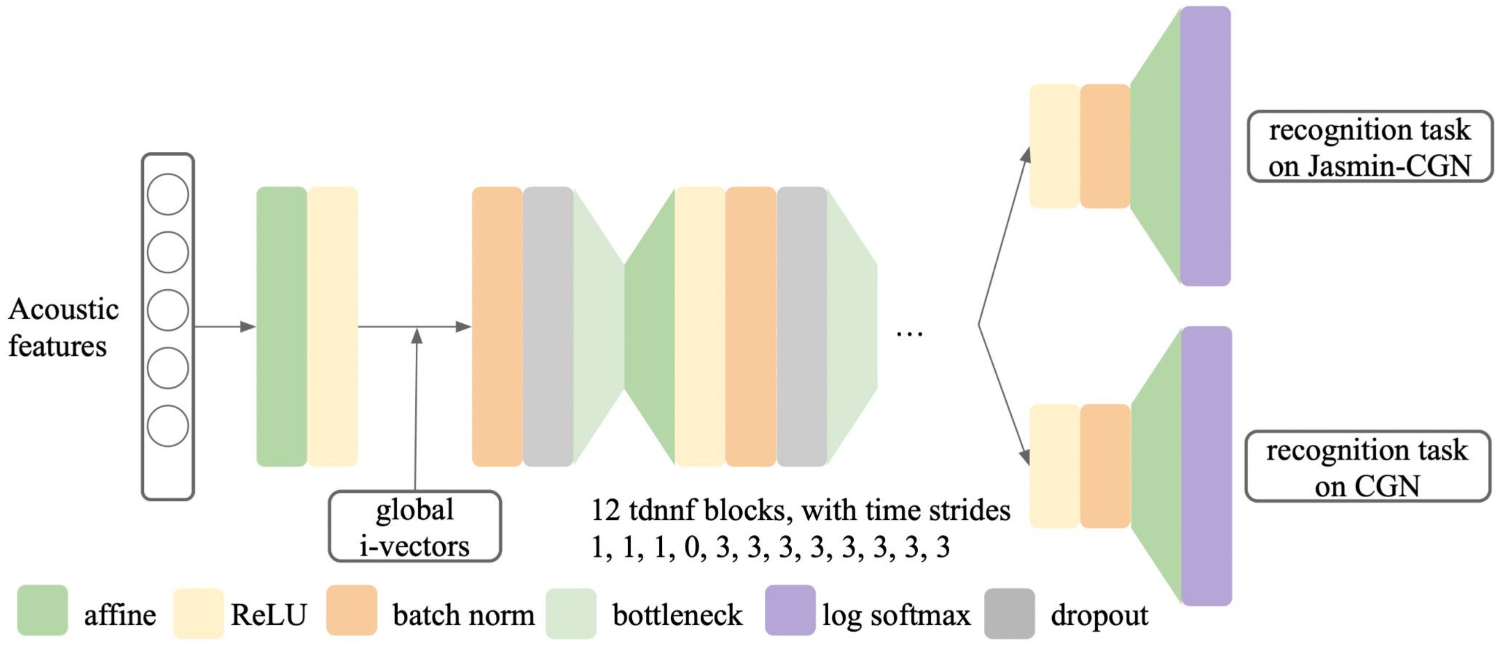


Fig.3. Multi-task Learning

Training Strategy

Datasets

Only speech recorded in The Netherlands has been used.

➔ **CGN**

423h training data

➔ **JASMIN-CGN**

36.12h training data; *1.45h* native read speech test data, *1.63h* non-native read speech test data, *0.68h* native HMI speech test data, *0.36h* non-native HMI speech test data.

Testsets

5 sets of target data consisting of both native and non-native Dutch are created

➔ **6 speakers per age group, 3 females and 3 males**

Each one has 2 types of recordings, human-machine interactive (HMI) speech and read speech

➔ **Native speakers whose home language is only Dutch without any second home language**

➔ **Non-native speakers whose home languages are picked to be as inclusive as possible**

Result

Method	Datasets	R_D	R_{NN}	H_D	H_{NN}	B_R	B_H
in-domain	C_{train}, J_{train}	17.97	31.65	28.8	37.95	13.68	9.15
	$C_{train}, J_{train} + SP$	17.55	30.13	29.47	36.65	12.58	7.18
	$C_{train}, J_{train} + VP$	20.49	32.54	29.9	37.65	12.05	7.75
	$C_{train}, J_{train} + PS$	17.26	30.04	28.59	36.33	12.78	7.74
	$C_{train}, J_{train} + SP + VP + PS$	16.82	30.04	27.95	**34.66**	13.22	6.71

Result

Method	Datasets	R_D	R_{NN}	H_D	H_{NN}	B_R	B_H
fine-tune	J_{train}	15.61	31.09	45.24	53.7e	15.48	8.48
	$J_{train} + SP$	15.31	30.89	45.1	52.81	15.58	7.71
	$J_{train} + VP$	15.66	31.45	46.46	53.96	15.79	7.5
	$J_{train} + PS$	13.85	30.3	47.06	54.55	16.45	7.49
	$J_{train} + SP + VP + PS$	**12.64**	29.91	43.79	50.1	17.27	**6.31**

Result

Method	Datasets	R_D	R_{NN}	H_D	H_{NN}	B_R	B_H
multi-task	C_{train}, J_{train}	21.11	34.8	29.05	35.98	13.69	6.93
	$C_{train}, J_{train} + SP$	20.03	34.05	28.67	35.37	14.02	6.7
	$C_{train}, J_{train} + VP$	20.84	33.73	29.01	35.86	12.89	6.85
	$C_{train}, J_{train} + PS$	18.79	27.88	28.29	35.06	**9.09**	6.77
	$C_{train}, J_{train} + SP + VP + PS$	17.05	**27.87**	**28.03**	34.99	10.82	6.96

Result

Method	Datasets	R_D	R_{NN}	H_D	H_{NN}	B_R	B_H
in-domain	C_{train}, J_{train}	17.97	31.65	28.8	37.95	13.68	9.15
	$C_{train}, J_{train} + SP$	17.55	30.13	29.47	36.65	12.58	7.18
	$C_{train}, J_{train} + VP$	20.49	32.54	29.9	37.65	12.05	7.75
	$C_{train}, J_{train} + PS$	17.26	30.04	28.59	36.33	12.78	7.74
	$C_{train}, J_{train} + SP + VP + PS$	16.82	30.04	27.95	**34.66**	13.22	6.71
fine-tune	J_{train}	15.61	31.09	45.24	53.7e	15.48	8.48
	$J_{train} + SP$	15.31	30.89	45.1	52.81	15.58	7.71
	$J_{train} + VP$	15.66	31.45	46.46	53.96	15.79	7.5
	$J_{train} + PS$	13.85	30.3	47.06	54.55	16.45	7.49
	$J_{train} + SP + VP + PS$	**12.64**	29.91	43.79	50.1	17.27	**6.31**
multi-task	C_{train}, J_{train}	21.11	34.8	29.05	35.98	13.69	6.93
	$C_{train}, J_{train} + SP$	20.03	34.05	28.67	35.37	14.02	6.7
	$C_{train}, J_{train} + VP$	20.84	33.73	29.01	35.86	12.89	6.85
	$C_{train}, J_{train} + PS$	18.79	27.88	28.29	35.06	**9.09**	6.77
	$C_{train}, J_{train} + SP + VP + PS$	17.05	**27.87**	**28.03**	34.99	10.82	6.96

Could data augmentation help reduce the bias against non-native accented speech in ASR systems?

Yes

As shown in the table in the previous slide, for in-domain training the results on adding Jasmin-CGN and the perturbed data with different augmentation techniques, the WER and bias has been decreased as compared to the baseline.

Could fine-tuning and multi-task learning be effective in reducing bias against non-native accented speech when compared with standard training methods?

Yes

Among the techniques employed, fine-tuning and multi-task learning reduce the bias more than simply including the target non-native speech as in-domain data.

Thank you.