Semi-Supervised DNN Training in Meeting Recognition

Pengyuan Zhang[1]***, Yulan Liu***[2]***, Thomas Hain***[2]

[1]*Key Laboratory of Speech Acoustics and Content Understanding, IACAS, China;* [2]*University of Sheffield, UK.*

[1]pzhang@hccl.ioa.ac.cn [2]{acp12yl, t.hain}@sheffield.ac.uk

Motivation

Due to domain specificity, there are low resource scenarios where annotated training data can be especially expensive to obtain. Existing research based on advanced DNN front-end utilized semi-supervised training to improve the recognition performance of a seed system which is trained with limited amount of annotated data.In this work, semi-supervised training of two typical low-resource scenarios was explored. The performance of semi-supervised training with confidence score based hypothesis transcription selection is verified and extended with analysis on hypothesis label accuracy. By comparing hypothesis labels of different resolution, the semi-supervised training is further improved with an optimal balance between label resolution and accuracy achieved at monophone level.

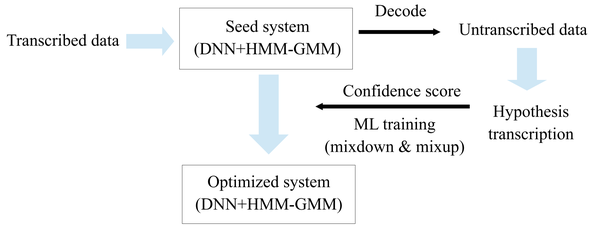
Two Typical Low-resource Scenarios

Scenario 1 (S1): Among the data that matches the decoding task, only a limited amount of data is transcribed.

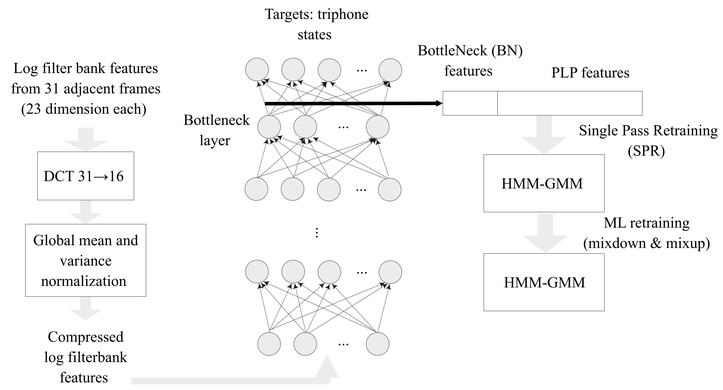
Scenario 2 (S2): None of the data that matches the decoding task is transcribed at all. However some slightly mismatched data within the same domain is transcribed.

Scenario	Dataset	Dur.	#Utt.	#Words	Corpus	Role in S1
S1	acntrain	15.8h	12876	152876	AMI	Transcribed training data.
	acotrain	72.0h	60297	710850	AMI	Untranscribed training data.
	acftest	6.1h	4633	54820	AMI	Test set.
S2	icsi	10.0h	7268	126487	ICSI	Transcribed training data.
	acftrain	87.8h	73173	863726	AMI	Untranscribed matching data.
	acftest	6.1h	4633	54820	AMI	Test set.

Semi-Supervised DNN Training



DNN front-end configuration: 368 dimensional input layer + 1745 dimensional hidden layer $\times$ 3 + 26 dimensional bottleneck layer + output layer



Confidence Score

The confident utterances are selected out from the decoded data based on the confidence score which is combines the DNN posteriors and HMM-GMM decoding results in two ways:

	Agreement-based	Posterior-based
word	$C_{\text{agr}}(w_i,t_s,t_e)=\frac{c_{\text{agr}}(t_s,t_e)}{N_{w_i}}$	$C_{\text{pos}}(w_i,t_s,t_e)=\frac{1}{t_e-t_s+1}\sum_{k=t_s}^{t_e}\log p_k$
utterance	$C_{\text{agr}}(u)=\frac{1}{K}\sum_{i=1}^{K}C_{\text{agr}}(w_i,t_s,t_e)$	$C_{\text{pos}}(u)=\frac{1}{K}\sum_{i=1}^{K}C_{\text{pos}}(w_i,t_s,t_e)$

According to the triphone state frame error rate of the selected utterances (below), C_{pos} is slightly better than C_{agr} thus is used in all experiments.

Selection (%)	100	90	80	70	60	50	40	30	20	10	5
C_{agr}	31.3	30.7	29.7	28.2	26.2	24.0	21.5	19.0	16.0	11.4	6.5
C_{pos}	31.3	30.5	29.4	27.8	25.8	23.4	20.8	18.1	14.9	9.2	5.5

	Training Data		%WER
	DNN	HMM-GMM	
S1	acntrain	acntrain	*29.7*
	acntrain	acntrain+acotrain.hyp	29.0
	acntrain	acntrain+acotrain.hyp0.7	28.8
	acntrain+acotrain.hyp	acntrain+acotrain.hyp	30.0
	acntrain+acotrain.hyp0.7	acntrain+acotrain.hyp0.7	**29.3**
S2	icsi	icsi	*40.4*
	icsi	acftrain.hyp	35.5
	icsi	acftrain.hyp0.7	35.8
	icsi+acftrain.hyp	icsi+acftrain.hyp	35.2
	icsi+acftrain.hyp0.7	icsi+acftrain.hyp0.7	**34.5**

- Adding hypothesis data improved HMM-GMM in both scenarios, with a larger improvement in S2.
- Adding hypothesis data does not guarantee to improve DNN front-end. In S2 evident improvement is observed while in S1 there is a slight degradation.

Hypothesis Label Resolution and Accuracy

Scenario	Seed data	%FER on	*TS*	*MS*	*M*	*PC*
S1	acntrain	acotrain	39.9	36.3	21.1	15.1
S2	icsi	acftrain	54.3	49.0	45.9	36.4

Hypothesis label frame error rate increases as the resolution increases.

Scenario	Training data	DNN targets			
		TS	*MS*	*M*	*PC*
S2	icsi	**40.4**	41.5	42.6	49.0
	icsi+acftrain.hyp	35.2	34.6	**33.8**	39.5
	icsi+acftrain.hyp0.7	34.5	34.2	**33.7**	39.4

- In ground-truth transcription the accuracy is guaranteed, thus a higher label resolution in DNN training leads to better recognition performance.
- In semi-supervised training, label errors in transcription are harmful, especially for discriminative DNN front-end. The best balance between label resolution and accuracy is met at monophone level.

Summary

- Two typical scenarios with low-resource are investigated with semi-supervised training methods.
- Adding data with hypothesis transcription in HMM-GMM training improved recognition performance by 0.7% absolute in S1 and 4.9% absolute in S2.
- Adding hypothesis transcription does not guarantee to improve DNN front-end. In S2 a 5.2% absolute or 12.9% relative WER reduction is observed over the seed system, while in S1 a 0.3% absolute degradation is observed.
- By selecting 70% hypothesis out of all based on confidence score, over the seed system there is a 0.4% absolute WER reduction in S1 and a 5.9% in S2.
- The frame error rate increases as the resolution of the hypothesis label increases. This harms the discriminative training of DNN front-end. By decreasing the label resolution, the best recognition performance is observed with monophones as the semi-supervised DNN training targets in S2. With the confidence score based hypothesis selection, using monophones training targets gave a 0.8% absolute WER reduction compared to using triphone states, and a 6.7% absolute or 16.6% WER reduction compared to the seed system.

Acknowledgements: Thanks to the EPSRC Programme Grant EP/1031022/1 (Natural Speech Technology project) and the Chinese Academy of Sciences Fellowship for Visiting Scholars for supporting the research.