ASR小组总结报告

2014年11月-2015年1月

张之勇，刘超，赵梦原，殷实，张雪薇，曾翔宇，林一叶，王冕

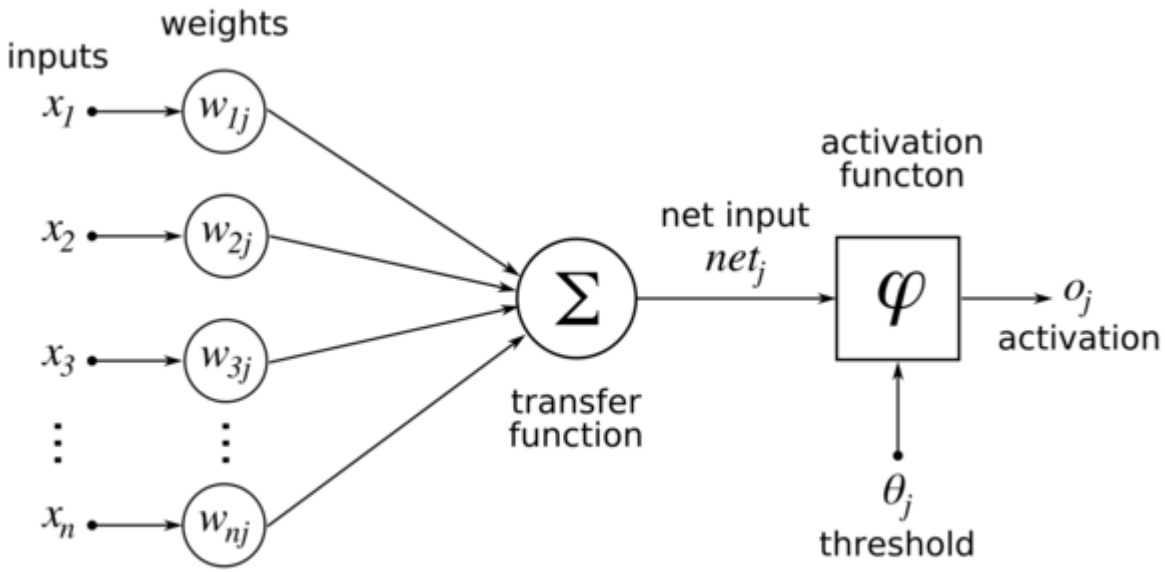
2015.01.24

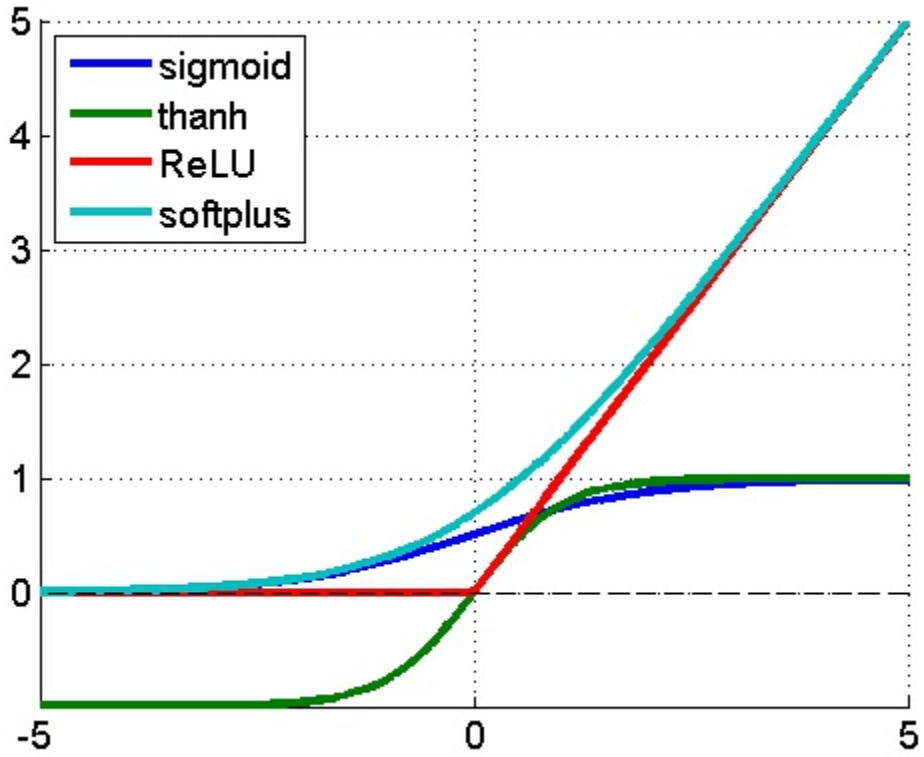
OverView

- Researches
  - Neural Network Activation Function
  - Dropout
  - Rate Of Speech(**ROS**) based Deep Neural Network
  - Sparse Deep Neural Network**(SparseDNN)**/DNN visualization
  - Bi-lingual Automatic Speech Recognition
  - Convolutional Neural Network(**CNN**)
  - Recurrent Neural Network(**RNN**)
  - Deep AutoEncoder (**DAE**)/Feature mapping
  - Voice Activity Detection(VAD)
  - VoicePrint Recognition(VPR)
- Projects
- Papers
  - Noisy DNN training

Researches

Activation Function





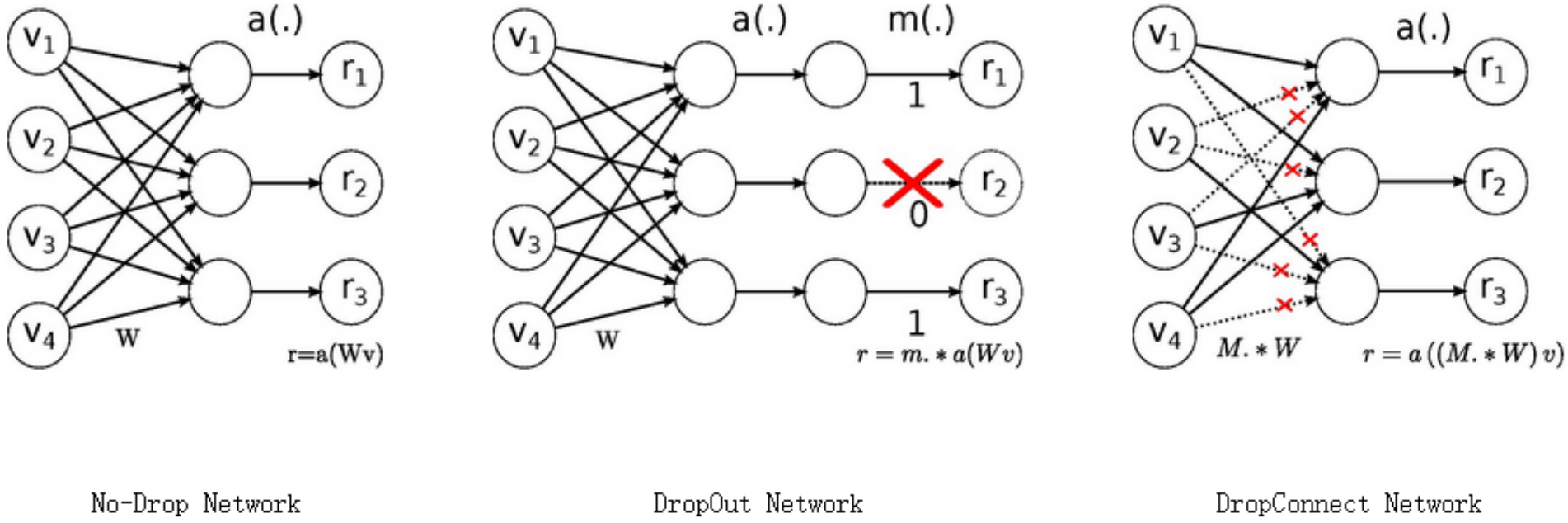
- Sigmoid
- Tanh
- **ReLU**
- **Maxout**
- **SoftMaxout**
- **P-norm**

Activation Function

- AURORA4 -- clean/15h

Condition/testcase	Clean	Airport	Babble	Car
Sigmoid(lr0.008)	6.04	29.91	27.76	16.37
ReLU(lr0.0008)	6.17	28.10	27.46	14.97
Maxout(lr0.001_gs6)	6.04	25.17	24.31	14.28
SoftMaxout(lr0.001_gs6)	5.92	27.08	25.15	15.56
P-norm(lr0.008_gs6_p2)	6.17	27.51	24.98	15.40

Dropout

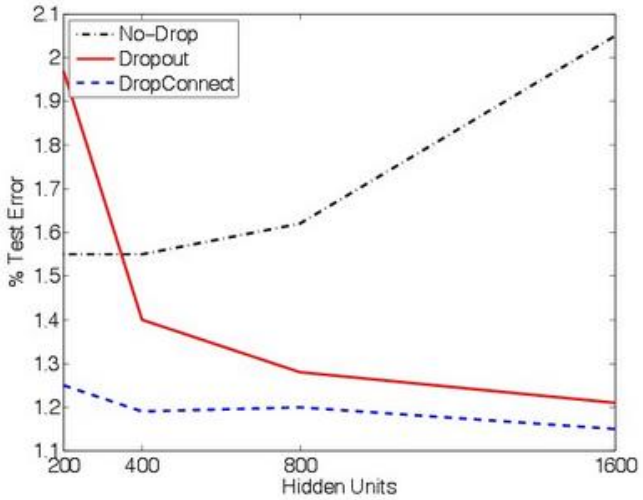


No-Drop Network DropOut Network DropConnect Network

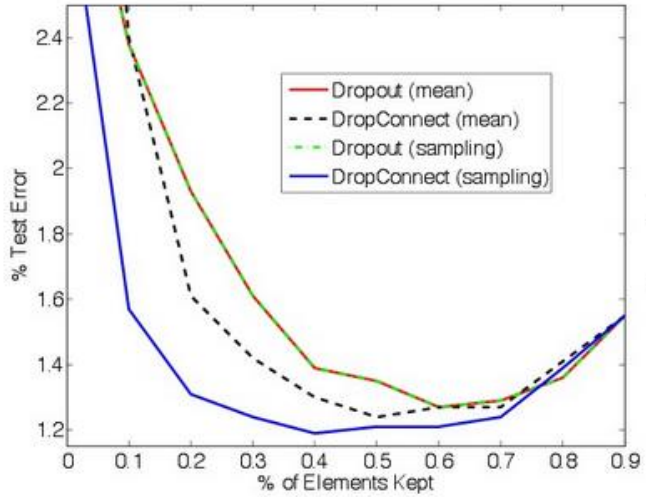
Dropout: randomly selected subset of activations are set to zero within each layer

DropConnect: randomly selected subset of *weights* within the network to zero

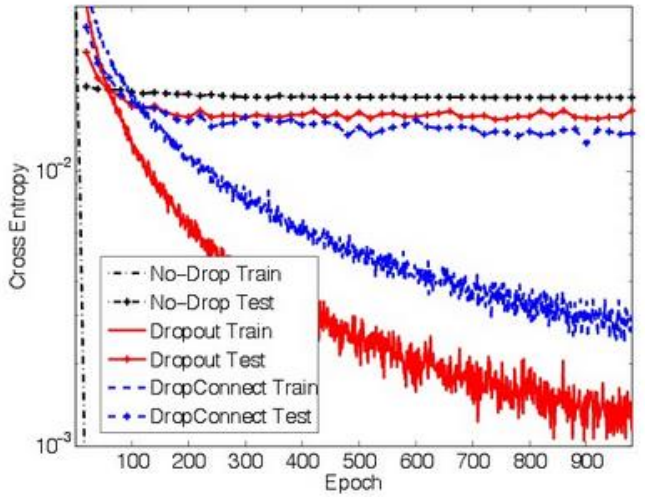
Dropout



(a)Prevent overfitting as the size of connected layers increase



(b)Varying the drop-rate in a 400-400 network



(c)Convergence properties of the train/test sets

Condition/testcase	Clean	Airport	Babble	Car
Sigmoid(lr0.008)	6.04	29.91	27.76	16.37
Dropout-0.8	5.94	24.94	23.67	15.77
Dropou0.8+Maxout_gs6	6.26	23.80	21.40	15.50

ROS

- Add speech rate feature to DNN training
- Test-case: Random selected from 1000h / TC 19h
- Results

Model/testcase	Random_select	TC
ROS	35.18	34.55
Non-ROS	35.22	35.04

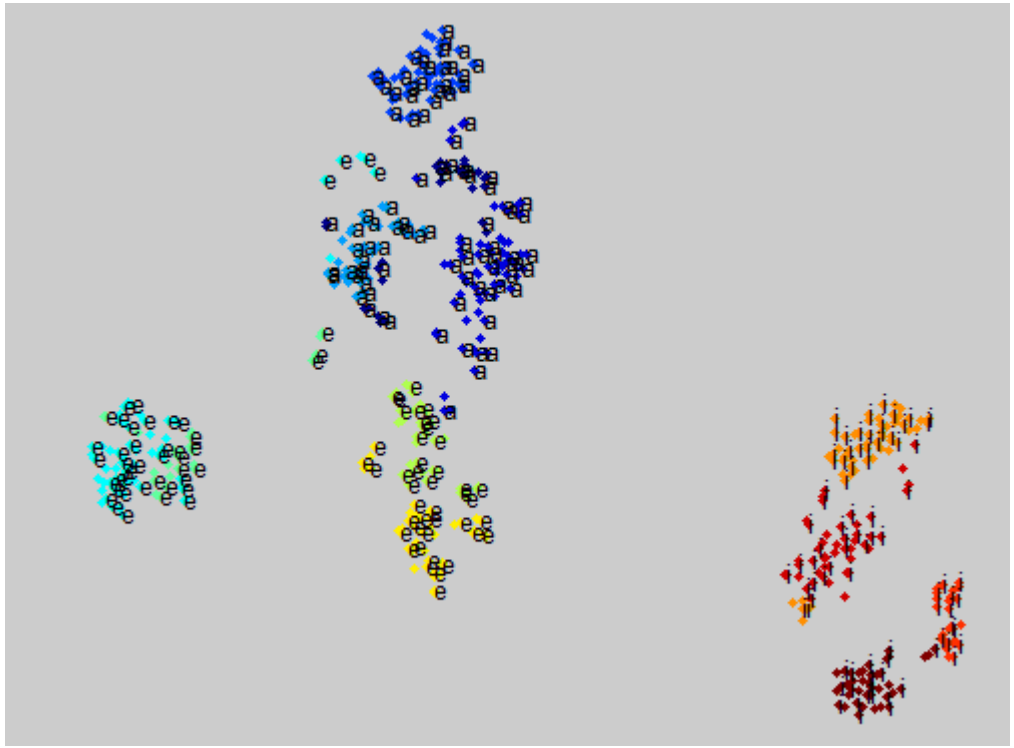
Sparse-DNN and visualization

- OBD-based sparse(AURORA4)

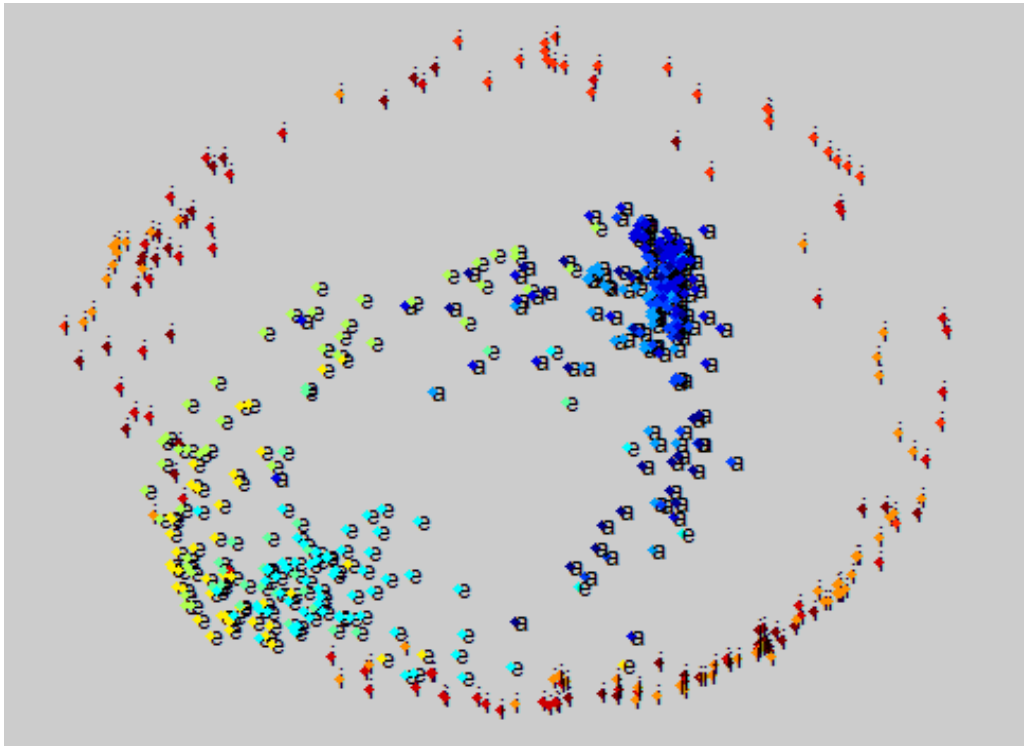
layers#	units#	sparsity	1 iter test WER%	1 + 4 iters test WER%
4	2048	100.0	11.62	11.31
4	2048	88.8	12.27	11.34
4	2048	76.0	12.42	11.28
4	2048	49.8	13.28	11.12
4	2048	39.4	16.18	11.21
4	2048	23.7	80.85	15.93

Sparse-DNN and visualization

- t-SNE/vMF Visualization



t-SNE without SoftMax



vMF-vMF

Bi-lingual ASR

- Bi-Softmax

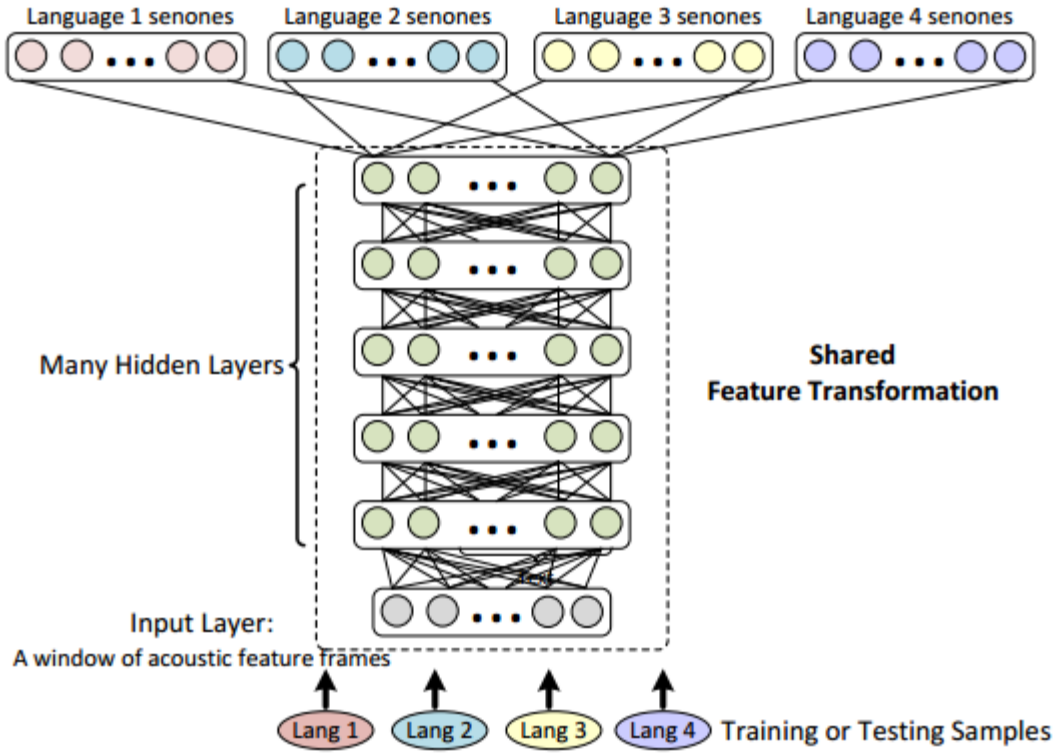


Figure 1: Architecture of the shared-hidden-layer multilingual DNN

- *Chinese and English code-switch SR*
- *Batch-level randomization*
- *Independent Softmax*
- *Simultaneous Competetive decoding*

Bi-lingual ASR

- Chinese-English phone share strategy

Model/testcase	English	Chinglish	Chinese
English-baseline	5.65	81.75	-
Chinglish-baseline	43.77	31.25	-
Chinese-baseline	-	-	31.05
Chinese/Chinglish no-share	54.62	43.35	33.76
Chinese/English no-share	12.48	82.86	32.25
Chinese/English initial-share	13.93	83.79	33.51
Chinese/English initial-share-std	15.65	83.15	34.43
Chinese/English initial-final-share	16.80	84.53	32.41

CNN

Testcase/model	CNN(2C+4D)	DNN(4-1200)
Clean	5.73	6.04
Airport	29.07	29.91
Babble	27.59	27.76
Car	17.25	16.37

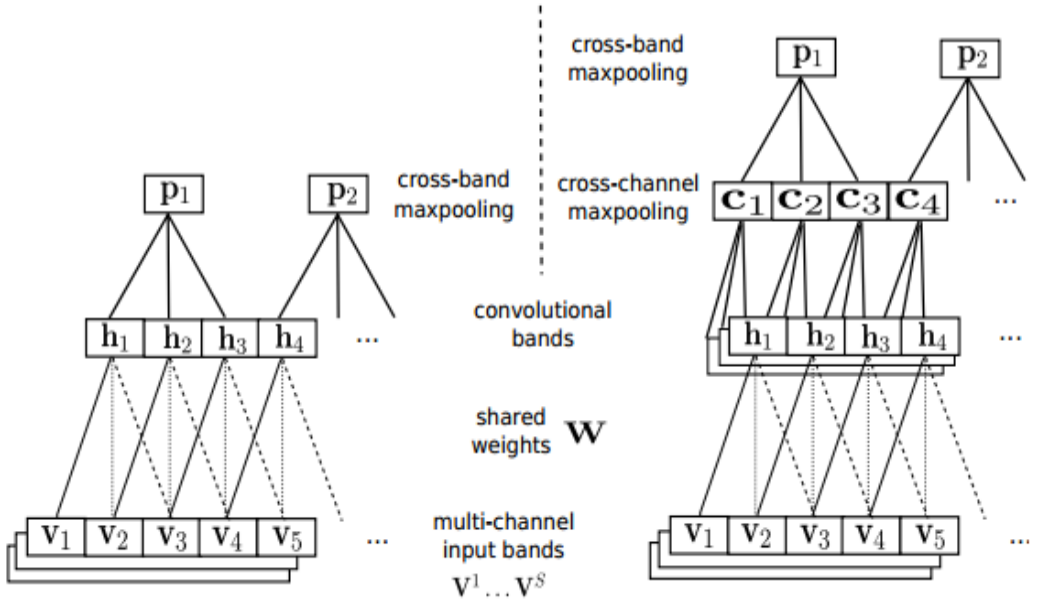


Fig. 1. Frequency domain max-pooling multi-channel CNN layer (left), and a similar layer with cross-channel max-pooling (right).

RNN

Model/metric	clean	Training-time/epoch
DNN-baseline	12.70	14min*14
RNN_800(512)_timeshift5	15.40	99min*14
RNN_512(256)_timeshift5	16.14	54min*15
RNN_512(256)-512(256)_timeshift5	14.87	132*18

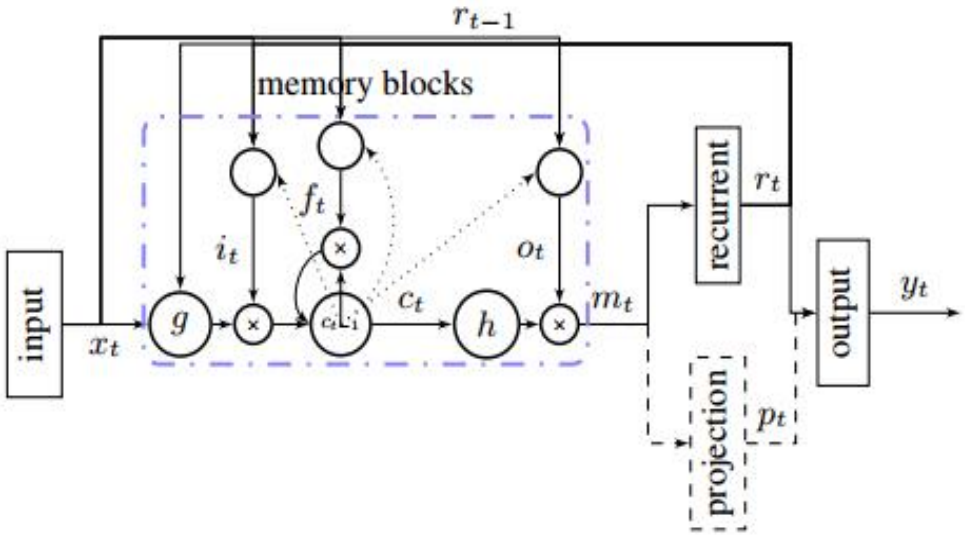


Fig. 1. LSTM based RNN architectures with a recurrent projection layer and an optional non-recurrent projection layer. A single memory block is shown for clarity.

De-noising-DAE

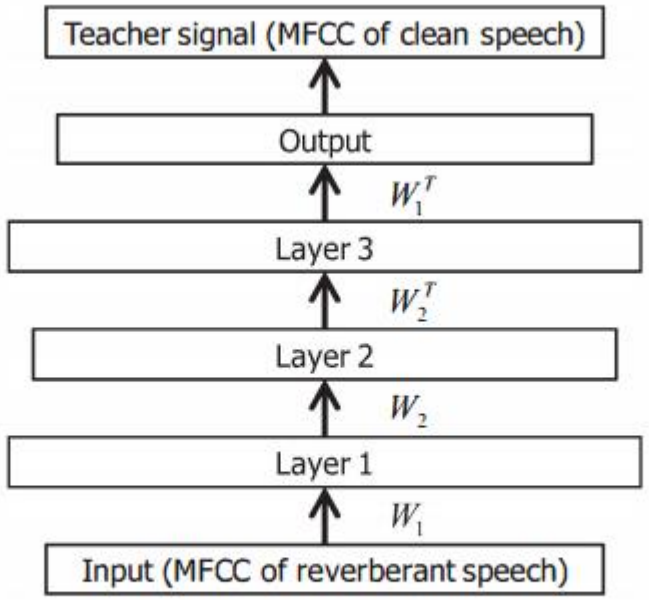


Figure 1: Topology of stacked denoising autoencoder for cepstral-domain dereverberation.

- DNN-DAE
- CNN-DAE
- RNN-DAE

De-noising-DAE

- DNN-DAE
  - noise

model/testcase(WER)	test_clean_wv1	test_airport_wv1	test_babble_wv1	test_car_wv1
std-xEnt-sigmoid-baseline	6.04	29.91	27.76	16.37
std+dae_cmvn_noFT_2-1200	7.10	15.33	16.58	9.23
std+dae_cmvn_splice5_2-100	8.19	15.21	15.25	9.31

  - Music

	clean	airport	bubble	car	piano	violin	symphony	rap	piano2
baseline	5.98	30.26	27.82	15.56	52.99	42.08	61.46	57.10	50.86
piano	8.91	22.20	21.50	15.38	8.17	31.80	56.65	63.63	34.22
violin	7.58	22.79	23.65	19.12	36.82	7.41	53.54	63.69	32.56
symphony	8.07	22.45	21.25	14.47	32.31	30.18	11.02	66,47	27.82
rap	12.11	34.90	29.84	21.48	56.76	54.57	66.87	51.49	60.22

De-noising-DAE

- DNN-DAE
  - Echo

dae_training_noise = echo_50ms_0.8(50%) + clean(50%):

	clean	airport	bubble	car	xwlb_20dB	xwlb_5dB	echo_50ms_0.8
std-xEnt-sigmoid-baseline	5.98	30.26	27.82	15.56	11.35	51.35	17.71
std+dae_cmvn_noFT_2-1200	6.21	47.49	52.84	19.33	-	-	19.76
std+dae_cmvn_splice5_2-1200	6.38	40.40	46.42	18.83	-	-	8.89

  - Real-Environment

training = real_far + real_near + clean
target = clean

	clean	airport	bubble	car	echo_50ms_0.8	echo_100ms_0.8	real_far	real_near
std-xEnt-sigmoid-baseline	5.98	30.26	27.82	15.56	17.71	53.62	93.30	68.93
std+dae_cmvn_splice5_2-1200	6.93	33.89	34.94	14.15	19.71	55.43	67.09	27.59
std+dae_cmvn_splice10_2-1200	7.03	34.94	35.78	15.21	20.47	51.01	56.60	25.96

De-noising-DAE

- CNN-DAE
  - noise

dae+dnn_std_aurora4

type	clean	car	babble	airport	echo_100_0.8	echo_50_0.8
dae-dnn_cmvn_splice_2-100/noise	8.74	9.18	15.42	14.70		
dae-cnn_splice5/noise	7.25	8.26	13.71	13.80		
dae-cnn_splice5/echo_50_0.8	14.45	38.35	63.52	64.30	57.44	7.27
dae-cnn_splice5/echo_100_0.8	35.62	51.01	68.77	69.04	22.26	18.83
dae-cnn_splice10/echo_50_0.8	13.73	35.02	58.93	60.09	55.73	6.78
dae-cnn_splice10/echo_100_0.8	65.67	86.16	90.52	90.63	7.84	47.56
dae-dnn_cmvn_splice5_2-1200//my50	14.49	40.67	69.99	66.51	57.69	7.31
dae-dnn_cmvn_splice5_2-1200//my100	41.30	68.13	82.65	80.26	27.68	21.21
dae-dnn_cmvn_splice10_2-1200//my100	65.02	85.13	90.02	90.46	8.89	47.11
dae-cnn_splice5/far_train_si284	94.00	95.11	97.56	96.69	95.93	95.47

De-noising-DAE

- RNN-DAE
  - noise

model/testcase(WER)	test_clean_wv1	test_airport_wv1	test_babble_wv1	test_car_wv1
std-xEnt-sigmoid-baseline	6.04	29.91	27.76	16.37
std+dae_cmvn_noFT_1-100_timeshift5	27.28	19.27	20.11	15.18
std+dae_cmvn_splice5_1-100_timeshift5	13.18	16.20	15.99	10.91
std+dae_cmvn_splice10_1-100_timeshift5	13.16	16.58	16.98	11.29

VAD

- DNN posterior probability-based VAD
- Add new energy features: frame/harmonic/targer energy

SNR	[40,)	[30,40)	[20,30)	[10,20)	[0,10)	(,0)	(,)
能量	94	83	57	41	16	0	48.5
谱熵	95	82	46	40	15	0	46.33
基频	96	88	81	74	33	0	62
DNN	98	93	87	80	52	11	70.16
带噪DNN	98	96	88	83	76	39	80

VPR

- DNN based d-vector extraction

Model	Splice	Dim	3Enroll-12Eval
Ivector-baseline	-	-	2.92
DNN-4	10	500	6.67
CNN-2_DNN-2	10	600	5.83
CNN-2_DNN-2	10	500	5.42

Projects

Projects

- Sinovoice AM training
  - ✓ Uyghur acoustic model training
  - ✓ Low resource AM optimization (based on Chinese)
  - ✓ 4200h-8k traininng (no-delta / add-delta)
  - ✓ 3600h-8k training (no-delta / add-delta) with new training schedule
  - ✓ 1400h dropout / maxout / dropout+maxout training
  - ✓ 1400h-100h Bi-softmax
  - ✓ ……..

Papers & Patents

Papers

- Noisy Training for Deep Neural Network in Speech Recognition
- 基于深度神经网络的语音端点检测

Patens

- 基于汉语声学模型的维语声学模型训练