Research Frontier in Speech Technology

Dong Wang

2013/09/09

Deep Neural Network

Systems (Features: static+Δ+ΔΔ)	Word error rate
Best GMM-HMM (MFCCs; fMPE+BMMI)	34.7%
DNN (MFCCs)	31.6%
DNN (256 log FFT bins)	32.3%
DNN (29 log filter-banks)	30.1%
DNN (40 log filter-banks)	**29.9%**
-Static 40-log-filter-banks only (11-frames)	31.1%
-Static 40-log-filter-banks only (19-frames)	30.5%

L. Deng, J. Li, J. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero. "Recent advances of deep learning for speech research at Microsoft," ICASSP, 2013.

Deep Neural Network

- New optimization approaches
  - Online stochastic gradient descent (BP)
  - GPU training/testing
  - Google DistBelief: distributed asynchronous update; parameter-dependent learning rate
  - IBM Hessian-free approach. Second order semi-online optimization

NEW TYPES OF DEEP NEURAL NETWORK LEARNING FOR SPEECH RECOGNITION AND RELATED APPLICATIONS: AN OVERVIEW, ICASSP 2013.

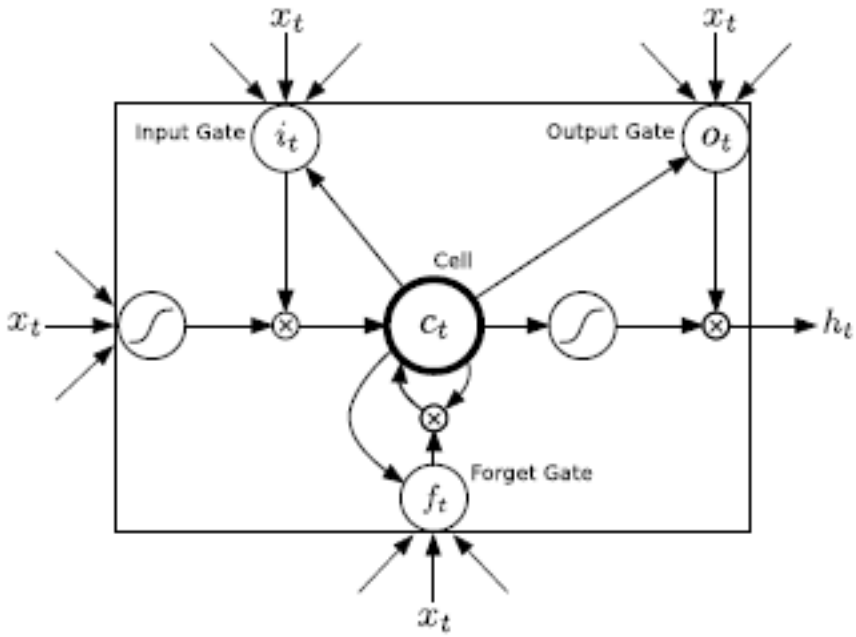
Deep Neural Network

- New optimization approaches (2)
  - Gradient clipping & Nesterov acceleration, Montreal.
  - Dropout and sparsity, Toronto
  - Rectify activations
  - Search for hyper-parameters

Recurrent network

$$h_t = \mathcal{H}\left(W_{xh}x_t + W_{hh}h_{t-1} + b_h\right)$$
$$y_t = W_{hy}h_t + b_y$$

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right)$$
$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right)$$
$$c_t = f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_\right.$$
$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right)$$
$$h_t = o_t \tanh(c_t)$$



Recurrent network

NETWORK	WEIGHTS	EPOCHS	PER
CTC-3L-500H-TANH	3.7M	107	37.6%
CTC-1L-250H	0.8M	82	23.9%
CTC-1L-622H	3.8M	87	23.0%
CTC-2L-250H	2.3M	55	21.0%
CTC-3L-421H-UNI	3.8M	115	19.6%
CTC-3L-250H	3.8M	124	18.6%
CTC-5L-250H	6.8M	150	18.4%
TRANS-3L-250H	4.3M	112	18.3%
PRETRANS-3L-250H	**4.3M**	**144**	**17.7%**

A. Graves, A. Mohamed, and G. Hinton. “Speech recognition with deep recurrent neural networks,” ICASSP, 2013.

Recurrent network

- Recurrent networks
  - Long-Short-Term Memory(LSTM)
  - Rectify activation & Dropout
  - Leaky integration $h_{t,i} = \alpha_i h_{t-1,i} + (1 - \alpha_i) F_i(h_{t-1}, x_t).$
  - Gradient clipping
  - Output modeling
  - Nesterov accelerated gradient (NAG)

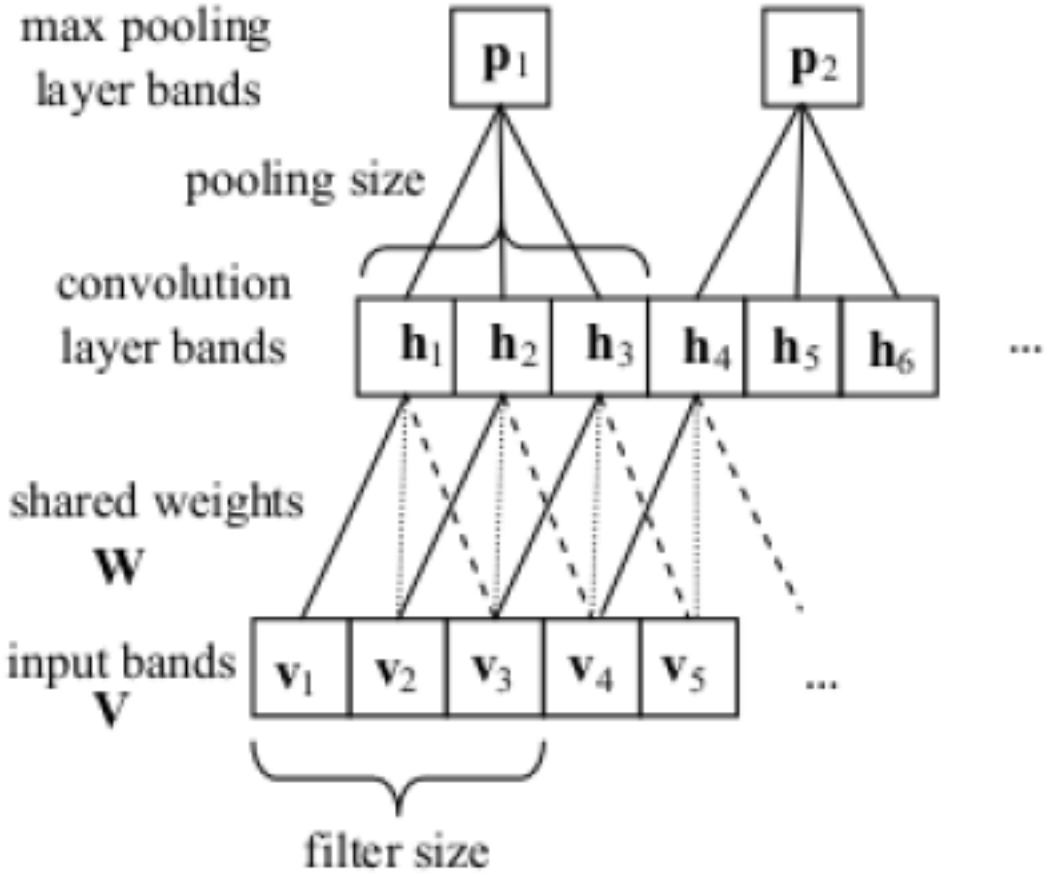
$$
\begin{aligned}
v_t &= \mu_{t-1} v_{t-1} - \epsilon_{t-1} \nabla f(\theta_{t-1}) \\
\theta_t &= \theta_{t-1} + v_t \\
&= \theta_{t-1} + \mu_{t-1} v_{t-1} - \epsilon_{t-1} \nabla f(\theta_{t-1})
\end{aligned}
$$

Recurrent network

Model	Piano-midi.de			Nottingham			MuseData			JSB chorales		
	LL (train)	LL (test)	ACC % (test)	LL (train)	LL (test)	ACC % (test)	LL (train)	LL (test)	ACC % (test)	LL (train)	LL (test)	ACC % (test)
RNN (SGD)	-7.10	-7.86	22.84	-3.49	-3.75	66.90	-6.93	-7.20	27.97	-7.88	-8.65	**29.97**
RNN (SGD+C)	-7.15	**-7.59**	22.98	-3.40	-3.67	67.47	-6.79	-7.04	30.53	-7.81	-8.65	**29.98**
RNN (SGD+CL)	-7.04	**-7.57**	22.97	-3.31	-3.57	67.97	**-6.47**	**-6.99**	**31.53**	-7.78	-8.63	**29.98**
RNN (SGD+CLR)	**-6.40**	-7.80	**24.22**	**-2.99**	-3.55	**70.20**	-6.70	-7.34	29.06	**-7.67**	-9.47	**29.98**
RNN (SGD+CRM)	-6.92	-7.73	23.71	-3.20	**-3.43**	68.47	-7.01	-7.24	29.13	-8.08	-8.81	29.52
RNN (HF)	-7.00	**-7.58**	22.93	-3.47	-3.76	66.71	-6.76	-7.12	29.77	-8.11	**-8.58**	29.41
RNN-RBM	N/A	-7.09	**28.92**	N/A	-2.39	**75.40**	N/A	-6.01	**34.02**	N/A	-6.27	33.12
RNN-NADE (SGD)	-7.23	-7.48	20.69	-2.85	-2.91	64.95	-6.86	-6.74	24.91	-5.46	-5.83	32.11
RNN-NADE (SGD+CR)	-6.70	-7.34	21.22	-2.14	-2.51	69.80	-6.27	-6.37	26.60	-4.44	-5.33	34.52
RNN-NADE (SGD+CRM)	-6.61	-7.34	22.12	-2.11	-2.49	69.54	-5.99	-6.19	29.62	**-4.26**	**-5.19**	**35.08**
RNN-NADE (HF)	**-6.32**	**-7.05**	23.42	**-1.81**	**-2.31**	71.50	**-5.20**	**-5.60**	32.60	-4.91	-5.56	32.50

Yoshua Bengio, Nicolas Boulanger-Lewandowski, Razvan Pascanu, ADVANCES IN OPTIMIZING RECURRENT NETWORKS, ICASSP 2013.

Convolutive DNN



T. Sainath, A. Mohamed, B. Kingsbury, B. Ramabhadran, “Convolutional neural networks for LVCSR,” ICASSP, 2013

Convolutive DNN

# of Convolutional vs. Fully Connected Layers	WER
No conv, 6 full (DNN)	24.8
1 conv, 5 full	23.5
2 conv, 4 full	22.1
3 conv, 3 full	22.4

Number of Hidden Units	WER
64	24.1
128	23.0
220	22.1
128/256	21.9

Feature	WER
Mel FB	21.9
VTLN-warped mel FB	21.3
VTLN-warped mel FB + fMLLR	21.2
VTLN-warped mel FB + d + dd	20.7
VTLN-warped mel FB + d + dd + energy	21.0

Multilanguage Language Training

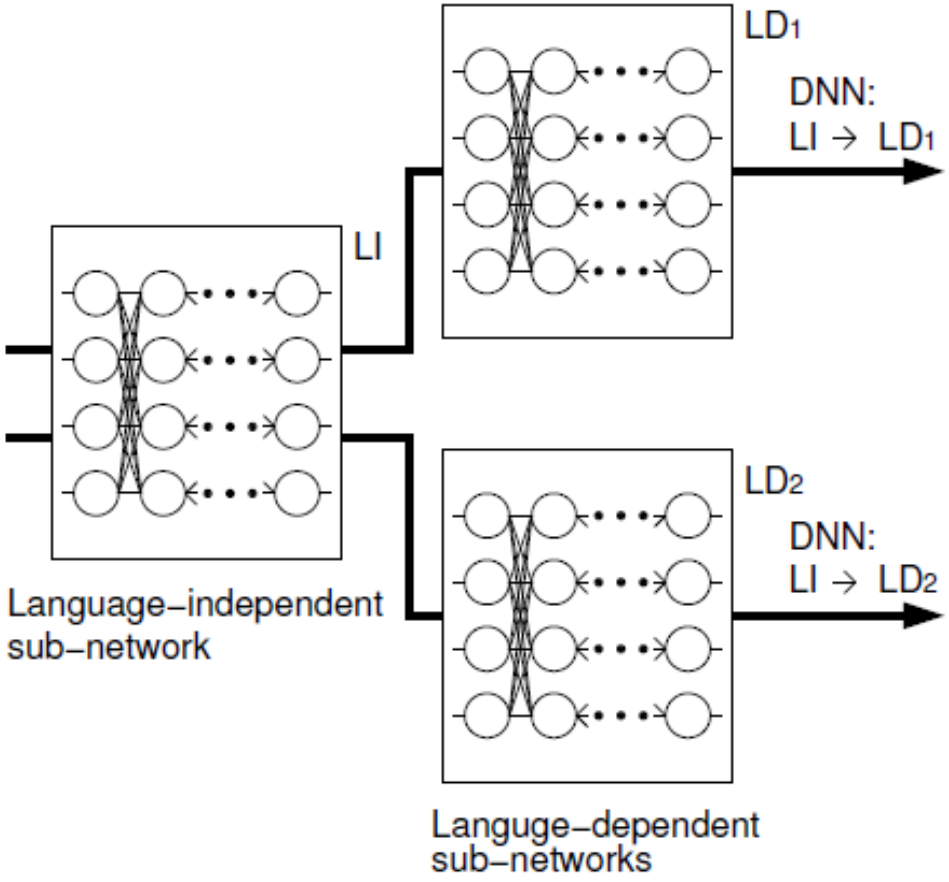


Table 1. Phoneme classification rate of baseline system (%)

# of layers	Lang.		
	Jp	En	Ch
2	71.69	51.09	60.37
4	73.28	51.51	61.58
6	75.13	54.50	62.58

Table 2. Phoneme classification performance (%) of DNNs ($\mathrm{LI}_{Jp,En}$→LD_{En} and $\mathrm{LI}_{Jp,En}$→LD_{Jp})

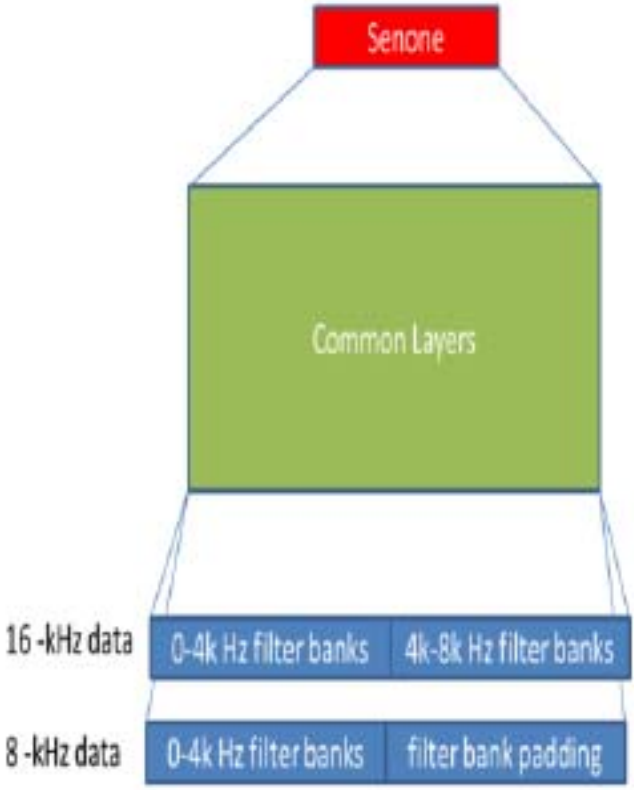
# of layers		Lang.	
LI	LD	En	Ja
1	5	54.37	74.73
2	4	54.61	74.86
3	3	54.63	74.90
4	2	54.59	74.82
5	1	52.50	73.38

Table 3. Phoneme classification performance (%) of DNNs ($\mathrm{LI}_{Jp,Ch}$→LD_{Ch} and $\mathrm{LI}_{Jp,Ch}$→LD_{Jp})

# of layers		Lang.	
LI	LD	Ch	Ja
1	5	62.89	74.60
2	4	62.93	74.67
3	3	63.10	74.85
4	2	63.08	74.70
5	1	62.05	74.08

AUTOMATIC LOCALIZATION OF A LANGUAGE-INDEPENDENT SUB-NETWORK ON DEEP NEURAL NETWORKS TRAINED BY MULTI-LINGUAL SPEECH, Shigeki Matsuda et al. ICASSP 2013.

Multi Channel Training



Training Data	Test WER (Wideband)	Test WER (Narrowband)
Wideband only	30.0%	71.2%
Narrowband only	-	29.0%
Wideband+Narrowband	**28.3%**	29.3%

L. Deng, J. Li, J. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero. "Recent advances of deep learning for speech research at Microsoft," ICASSP, 2013.

DNN Adaptation

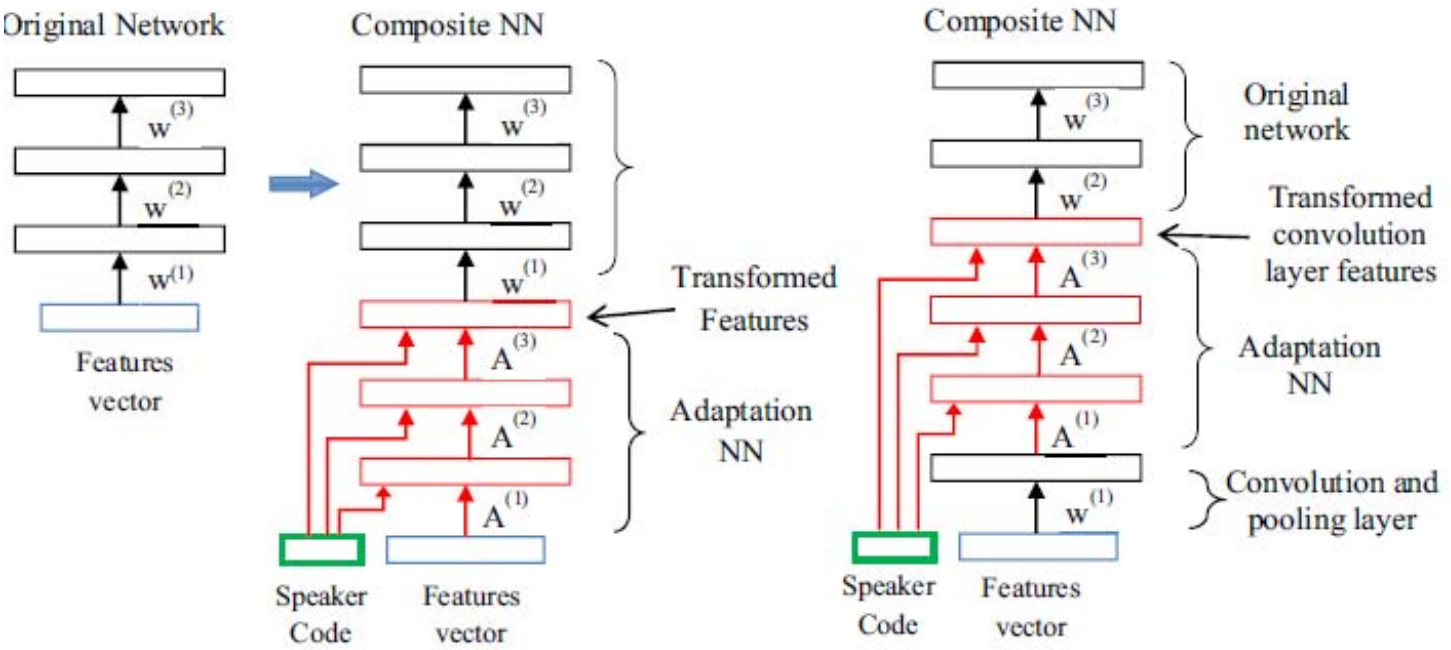
- Affine transform: feature-space discriminative linear regression (fDLR), LDA
- Input transform, output transform, hidden transform
- KL regulation in adaptation

DNN Adaptation

Speech Recognition Systems	WER	WERR (%)
GMM-HMM	43.6%	
DNN	34.1%	-
DNN + AdaptSoftMax (SGD)	29.4%	13.9
DNN + fDLR (SGD)	28.5%	16.8
DNN + AdaptSoftMax (batch)	30.9%	9.3

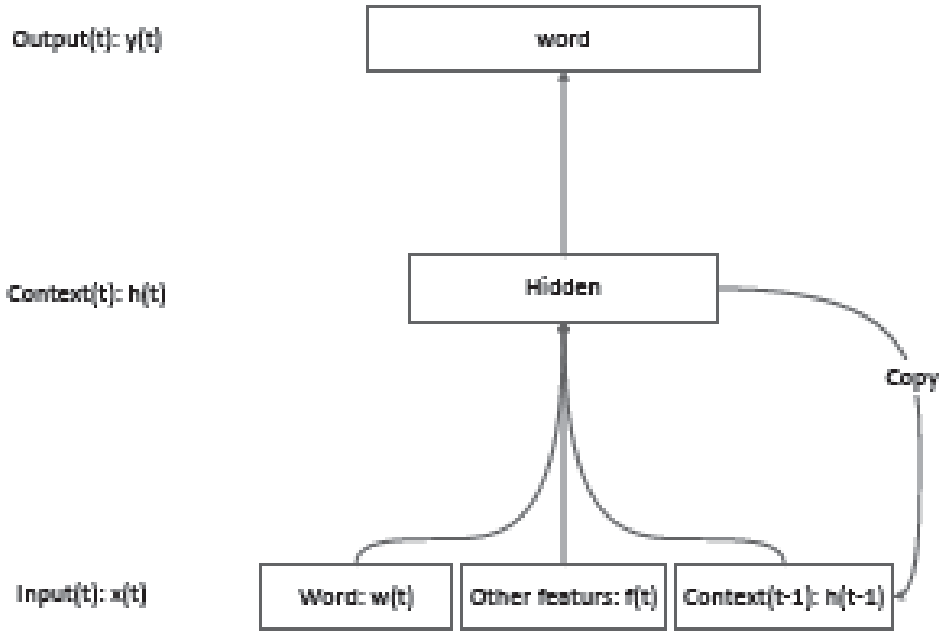
L. Deng, J. Li, J. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero. "Recent advances of deep learning for speech research at Microsoft," ICASSP, 2013.

CNN Adaptation



Ossama Abdel-Hamid and Hui Jiang, Rapid and Effective Speaker Adaptation of Convolutional Neural Network Based Models for Speech Recognition, IS 2013

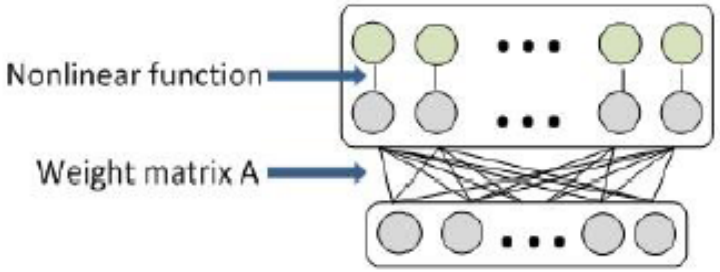
RNN for Language Modeling



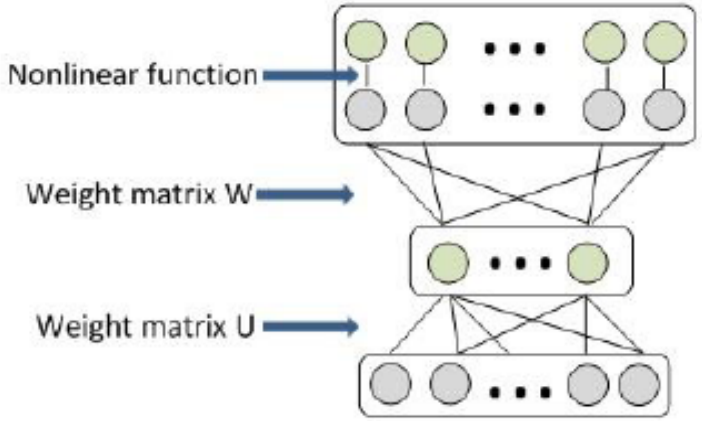
RNN for Language Modeling

Model	Perplexity	WPA
SRILM+KN5	228.82	-
SRILM+class+KN5	227.88	-
RNNLM	114.79	20.60
RNNLM+POS	90.56	22.62
RNNLM+lemma	102.26	21.57
RNNLM+SS	103.99	20.84
RNNLM+T10	102.45	21.58
RNNLM+T20	101.07	21.70
RNNLM+T40	104.48	21.51
RNNLM+T100	106.28	21.38
RNNLM+POS +SS	93.63	22.01
RNNLM+POS +T20	86.43	22.92
RNNLM+SS+T20	94.20	22.08
RNNLM+SS+lemma	93.35	22.03
RNNLM+POS +lemma(no word)	230.75	14.37
RNNLM+POS +lemma	90.49	22.63
RNNLM+POS +SS+T20	87.41	22.60
RNNLM+POS +SS+lemma	83.59	22.93
RNNLM+complete(300H)	85.88	22.85
RNNLM+complete(500H)	84.81	23.11

SVD



(a) One layer in original DNN model



b) Two corresponding layers in new DNN model

Table 1 Results of SVD restructuring on output layer on task 1

Acoustic Model		WER	Number of parameters
Baseline, GMM model		29.1%	11M
Original DNN model		25.6%	29M
SVD (1024)		25.6%	25M
SVD (512)		25.7%	21M
SVD (256)	Before fine-tune	28.6%	19M
	After fine-tune	25.6%	

Jan Xue, et al. Restructuring of Deep Neural Network Acoustic Models with Singular Value Decomposition, IS 2013

Fast I-vector approximation

$$\mathbf{s} = \mathbf{m} + \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{T}\mathbf{w} \qquad \mathbf{w}_{\mathcal{X}} = \mathbf{L}_{\mathcal{X}}^{-1}\mathbf{T}^{*}\mathbf{f}_{\mathcal{X}}$$

$$\mathbf{L}_{\mathcal{X}} = \mathbf{I} + \sum N_{\mathcal{X}}^{(c)}\mathbf{T}^{(c)*}\mathbf{T}^{(c)}$$

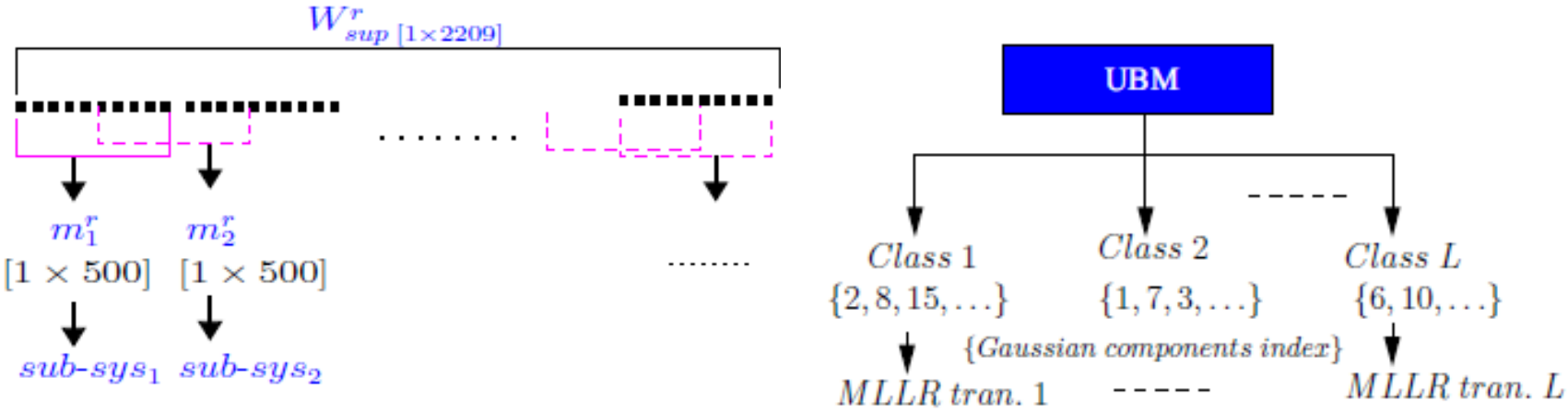
$$\mathbf{f}_{\mathcal{X}}^{(c)} = \mathbf{\Sigma}^{(c)^{-\frac{1}{2}}}\left[\sum_{t}\left(\gamma_t^{(c)}\mathbf{x}_t\right) - N_{\mathcal{X}}^{(c)}\mathbf{m}^{(c)}\right]$$

$$\mathbf{T}^{(c)}\mathbf{G}^{(c)} = \mathbf{O}^{(c)}\mathbf{\Pi}_{M}^{(c)}$$

$$\hat{\mathbf{T}}^{(c)} \approx \mathbf{O}^{(c)}\mathbf{\Pi}^{(c)}\mathbf{Q}\,.$$

Sandro Cumani, **Fast and Memory Effective I-Vector Extraction using a Factorized Sub–space, IS2013**

M-vectors



Anchor and UBM-Based Multi-Class MLLR *M-Vector* System for Speaker Verification, *A.K. Sarkar, Claude Barras, IS2013*

I-vector based short SV

Approach	Interview-interview		Interview-telephone		Telephone-microphone		Telephone-telephone	
	EER	DCF	EER	DCF	EER	DCF	EER	DCF
WCCN	19.81%	0.0798	24.33%	0.0896	19.76%	0.0821	17.87%	0.0695
LDA_WCCN	18.10%	**0.0767**	22.67%	0.0861	19.03%	0.0817	**16.46%**	**0.0679**
SN-LDA_WCCN	**18.01%**	0.0771	**21.57%**	**0.0858**	**18.94%**	**0.0813**	16.56%	0.0683

WCCN training	Interview-interview		Interview-telephone		Telephone-microphone		Telephone-telephone	
	EER	DCF	EER	DCF	EER	DCF	EER	DCF
Full-length	**18.10%**	**0.0767**	**22.67%**	**0.0861**	**19.03%**	**0.0817**	**16.46%**	**0.0679**
Matched-length	20.39%	0.0822	25.34%	0.0894	21.05%	0.0842	17.87%	0.0720
Mixed-length	19.57%	0.0804	24.44%	0.0879	20.38%	0.0823	17.30%	0.0708

$$(\mathbf{S}_b^{\mathrm{tel_{utt}}} + \mathbf{S}_b^{\mathrm{mic_{utt}}})\mathbf{v} = \lambda \mathbf{S}_w \mathbf{v}$$

$$\mathbf{S}_b^{\mathrm{tel_{utt}}}\mathbf{v} = \lambda \mathbf{S}_w \mathbf{v}$$

$$\mathbf{S}_b^{\mathrm{mic_{utt}}}\mathbf{v} = \lambda \mathbf{S}_w \mathbf{v}$$

$$\mathbf{A} = [\mathbf{A}_{tel}\mathbf{A}_{mic}]$$

Improving Short Utterance Based I-Vector Speaker Recognition Using Source and Utterance-Duration Normalization Techniques, IS2013

I-vector based short SV

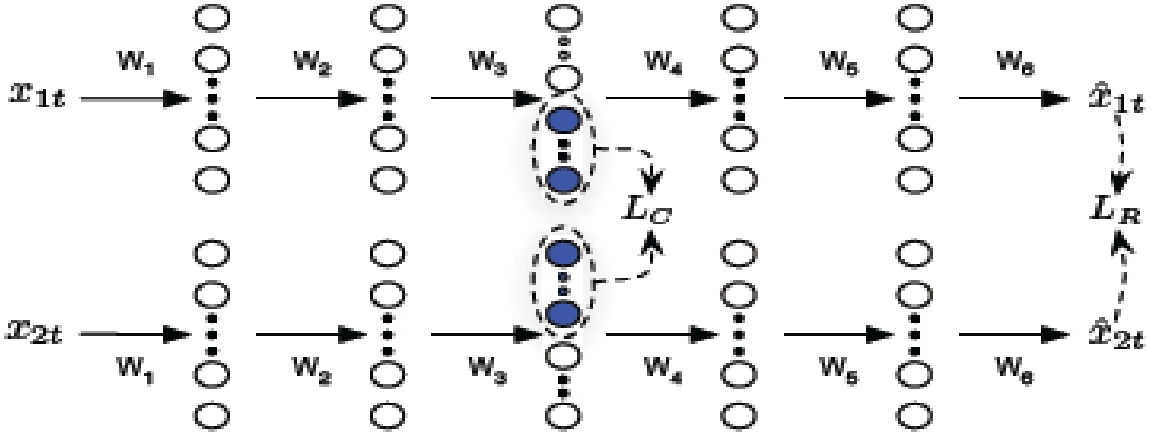
(a) *SUN-LDA-pooled vs LDA*

System				Interview-interview		Interview-telephone		Telephone-microphone		Telephone-telephone	
α_{tf}	α_{mf}	α_{ts}	α_{ms}	EER	DCF	EER	DCF	EER	DCF	EER	DCF
Baseline approach (LDA_WCCN)											
-	-	-	-	18.10%	0.0767	22.67%	0.0861	19.03%	0.0817	16.46%	0.0679
Source and utterance-duration normalized approach (SUN-LDA-pooled_WCCN])											
1.0	1.0	1.0	1.0	18.04%	0.0764	21.37%	0.0857	18.75%	0.0790	16.56%	0.0670
1.0	1.0	0.0	1.0	18.34%	0.0766	21.47%	0.0859	18.68%	0.0811	16.56%	0.0681
1.0	1.0	1.0	0.0	17.98%	**0.0759**	21.22%	0.0858	18.60%	0.0783	**16.31%**	0.0667
1.0	1.0	0.0	0.0	**17.97%**	0.0761	21.47%	0.0859	18.87%	0.0807	16.48%	0.0674
1.0	0.0	1.0	0.0	18.06%	0.0774	**21.21%**	**0.0856**	**17.66%**	**0.0776**	**16.31%**	**0.0665**

(b) *SUN-LDA-concat vs LDA*

System				Interview-interview		Interview-telephone		Telephone-microphone		Telephone-telephone	
α_{tf}	α_{mf}	α_{ts}	α_{ms}	EER	DCF	EER	DCF	EER	DCF	EER	DCF
Baseline approach (LDA_WCCN)											
-	-	-	-	18.10%	0.0767	22.67%	0.0861	19.03%	0.0817	16.46%	0.0679
Source and utterance-duration normalized approach (SUN-LDA-concat_WCCN)											
1.0	1.0	1.0	1.0	17.64%	**0.0760**	20.19%	0.0852	18.06%	0.0852	16.14%	0.0706
1.0	1.0	0.0	1.0	17.54%	0.0767	20.36%	0.0846	17.59%	0.0831	16.06%	0.0694
1.0	1.0	1.0	0.0	**17.29%**	0.0764	20.39%	0.0848	17.79%	0.0832	16.31%	0.0692
1.0	1.0	0.0	0.0	17.64%	**0.0760**	**19.91%**	**0.0841**	**17.39%**	0.0814	**15.82%**	0.0677
1.0	0.0	1.0	0.0	17.98%	0.0773	21.84%	0.0848	18.13%	**0.0798**	16.64%	**0.0654**

Deep learning for SID



Combining Deep Speaker Specific Representations with GMM-SVM for Speaker Verification, IS2013.