Structured Deep Learning for Context Awareness in Speech and Language Processing

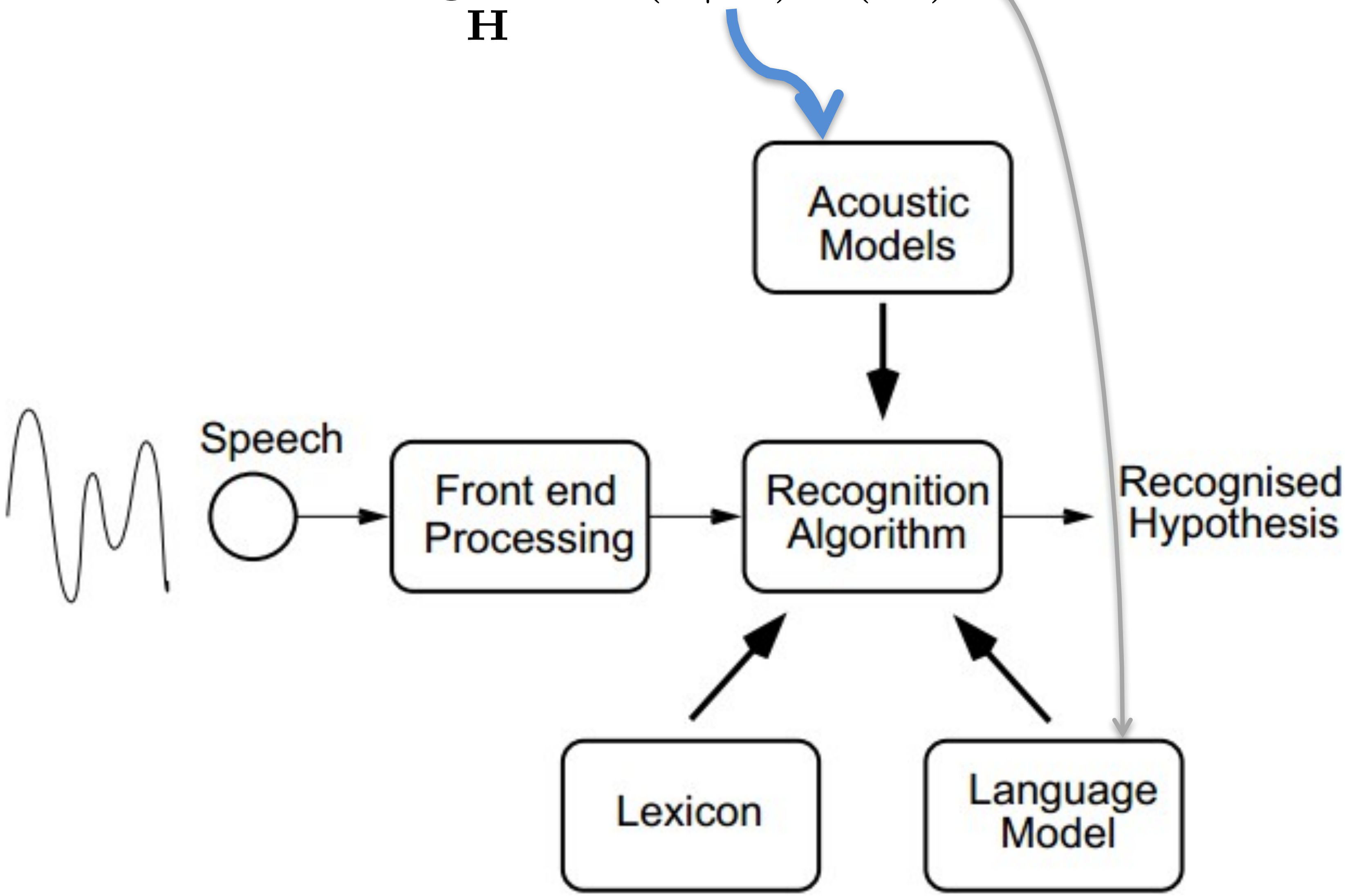
Kai Yu
Shanghai Jiao Tong University

Content

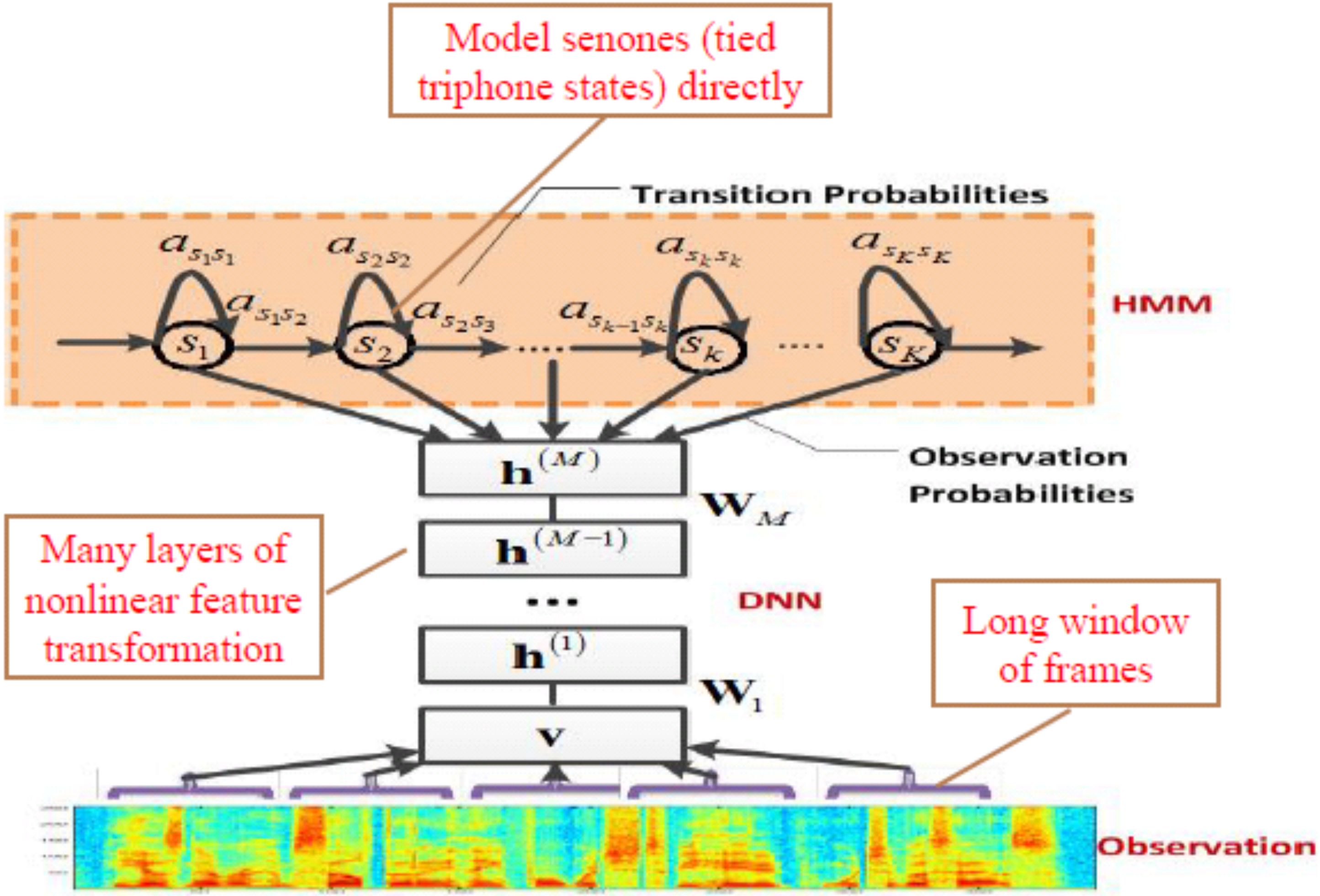
- Part I: Deep learning for Speech Recognition
- Part II: Multi-style and Context-aware Training
- Part III: Structured Deep Learning for Context Awareness

Part I
Deep learning for Speech Recognition

$$\hat{\mathbf{H}} = \arg\max_{\mathbf{H}} p(\mathbf{o}|\mathbf{H})P(\mathbf{H})$$

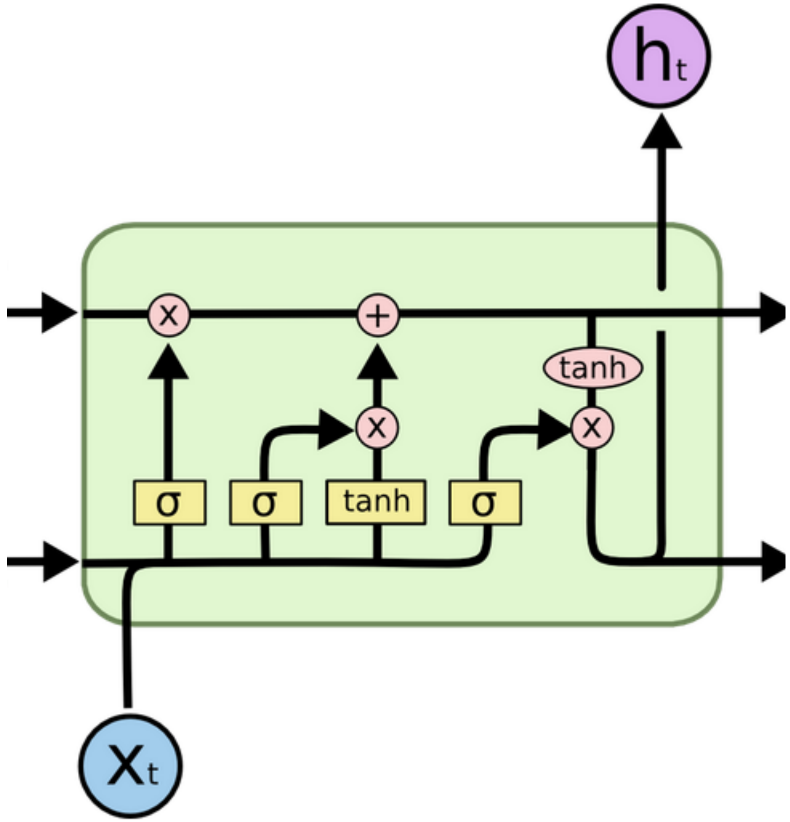


AM: CD-DNN-HMM v.s. GMM-HMM

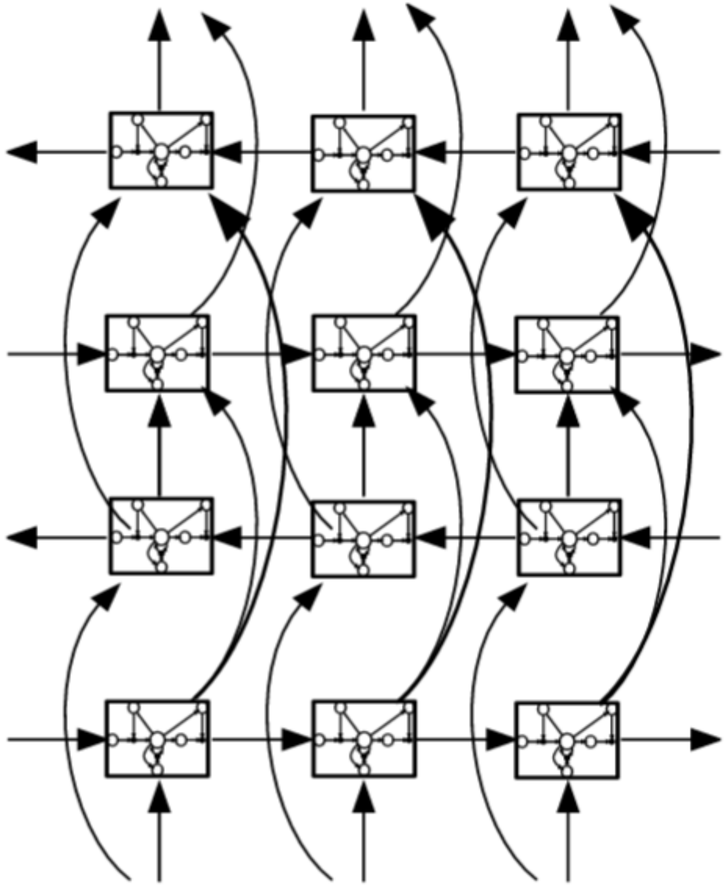


[1]

AM: LSTM



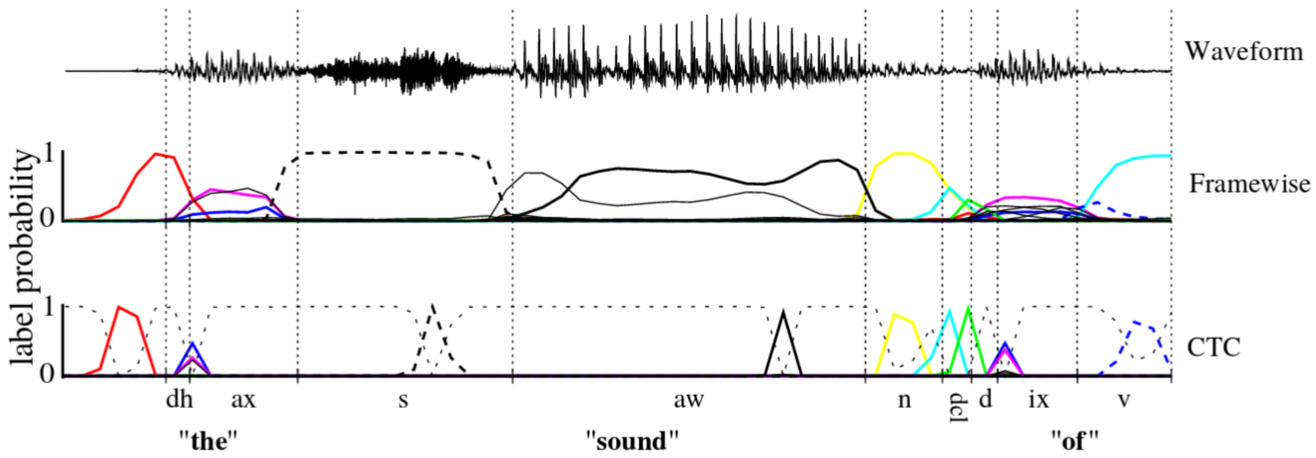
LSTM



[2]

BLSTM

AM: LSTM-CTC



[3]

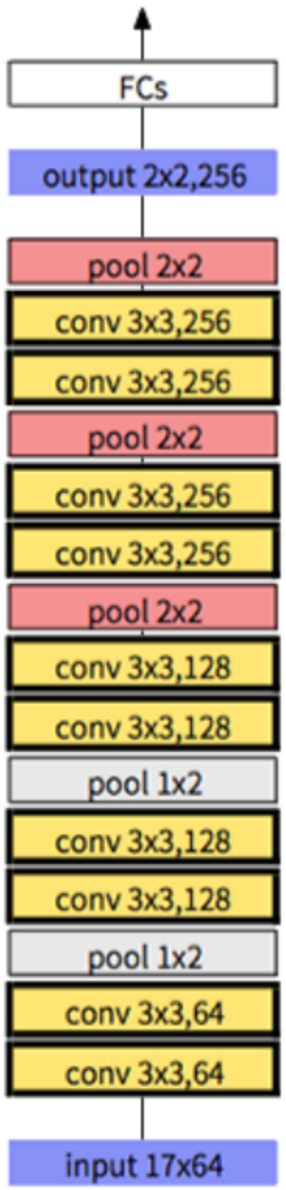
$$\mathcal{L}_{\texttt{ce}} = \sum_{t=1}^{T} \log P(y_t|x_t) \qquad P(\mathbf{y}|\mathbf{x}) = \sum_{\mathbf{a}\in\mathcal{B}^{-1}(\mathbf{y})} P(\mathbf{a}|\mathbf{x})$$

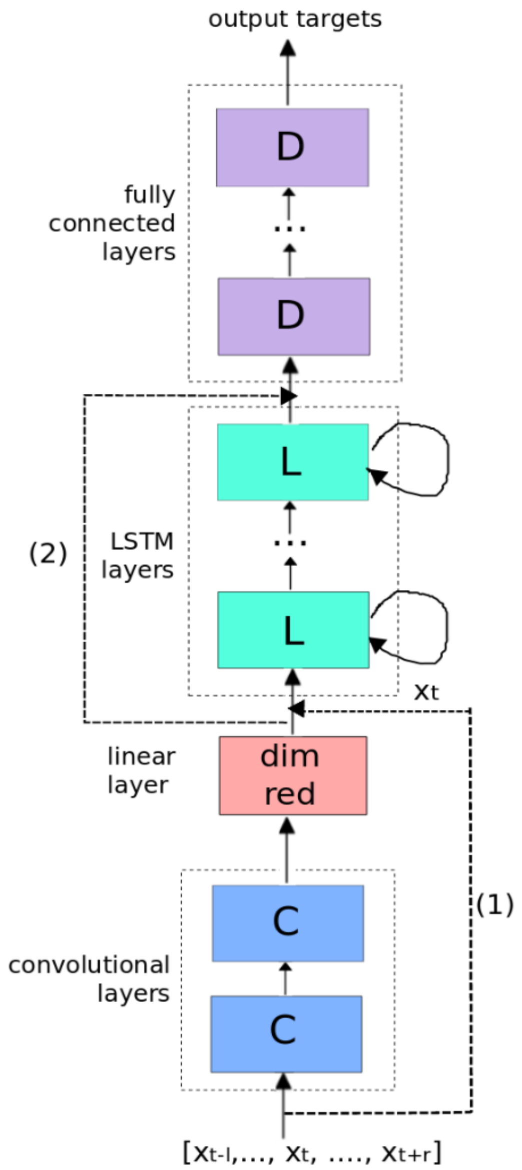
$$\mathcal{L}_{\texttt{ctc}} = \log P(\mathbf{y}|\mathbf{x})$$

$\mathcal{B}$ is an operator that removes first the repeated labels, then the blanks from alignments, for example,

$\mathcal{B}$(a,b,b,b,c,c) = $\mathcal{B}$(a,-,b,-,c,c) = (a,b,c)

AM: Very Deep CNN and CLDNN

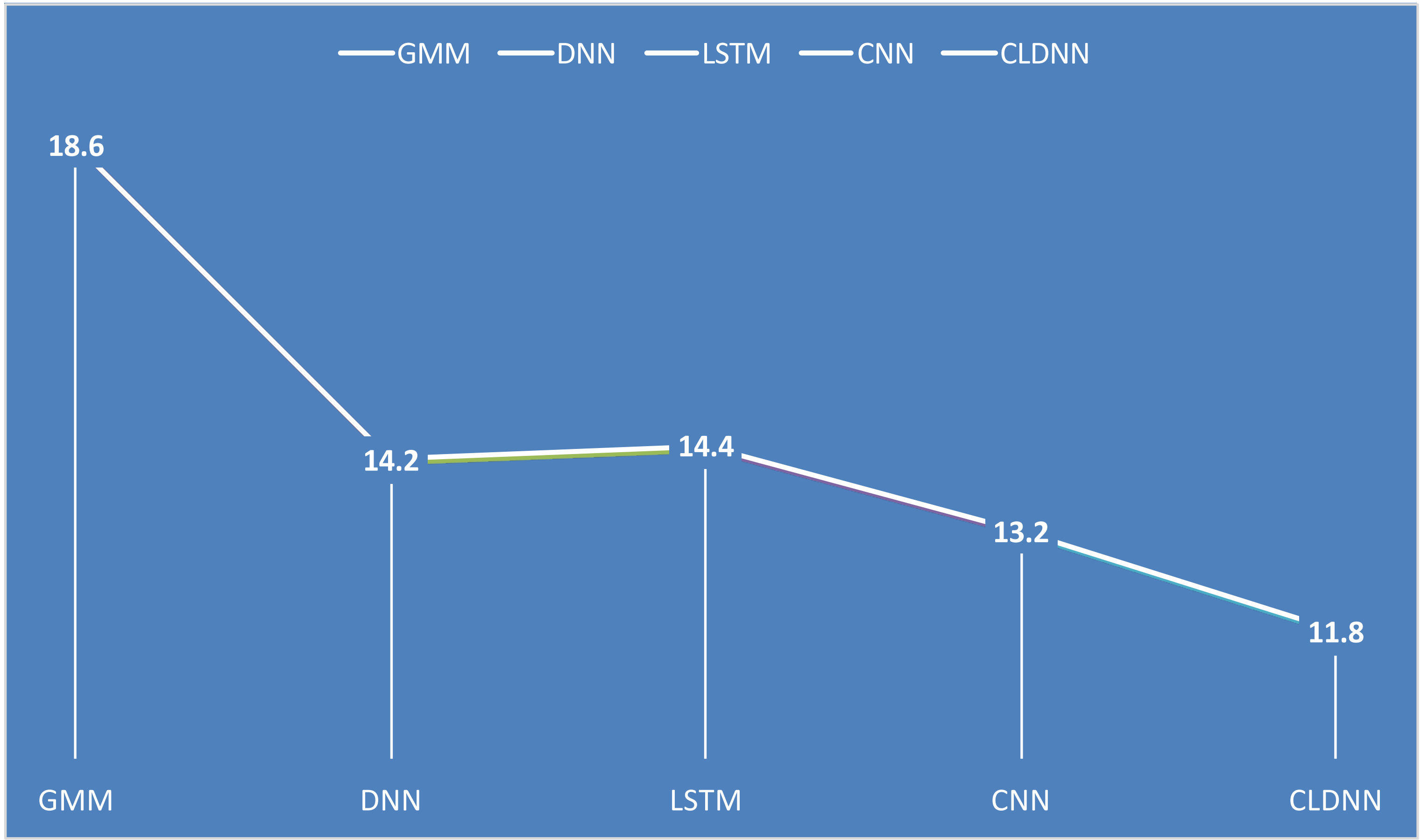




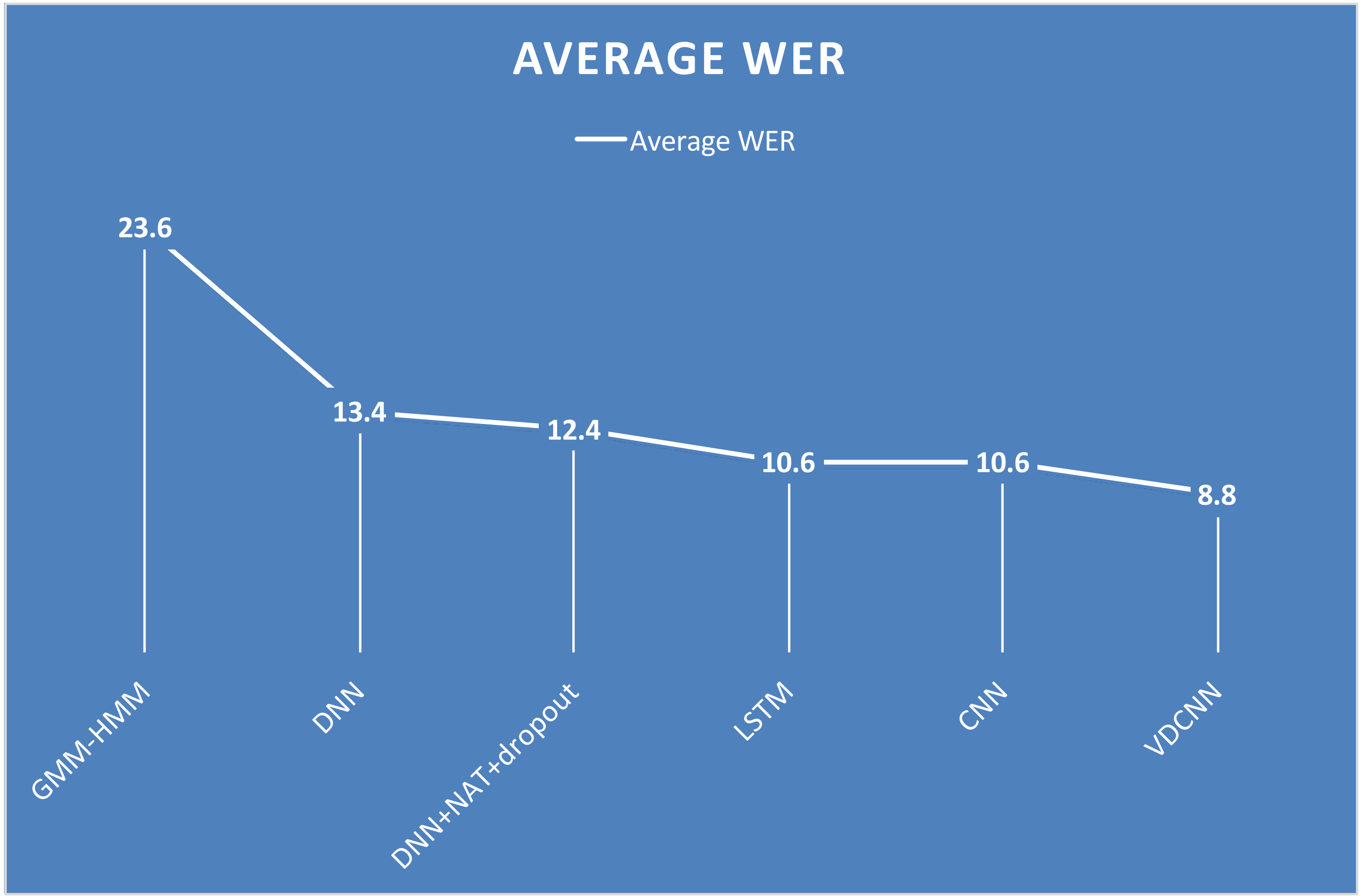
The typical speech inputs, with static, delta and double delta features, can be represented as 3 feature maps and each of them can be viewed as an image-map with a size of #times x #freqs

[4][5]

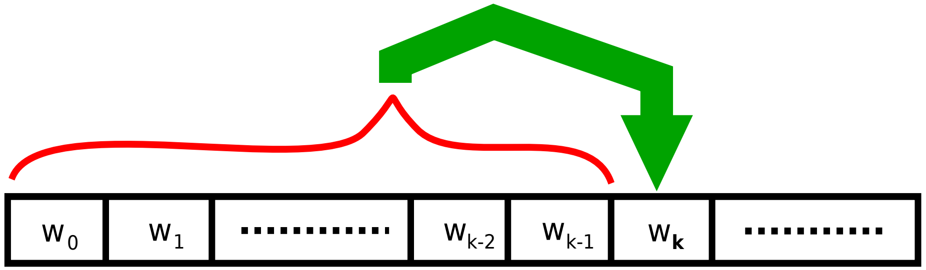
AM: Performance Gain (swbd) [6][7]



AM: Performance Gain (aurora4) [8][9]

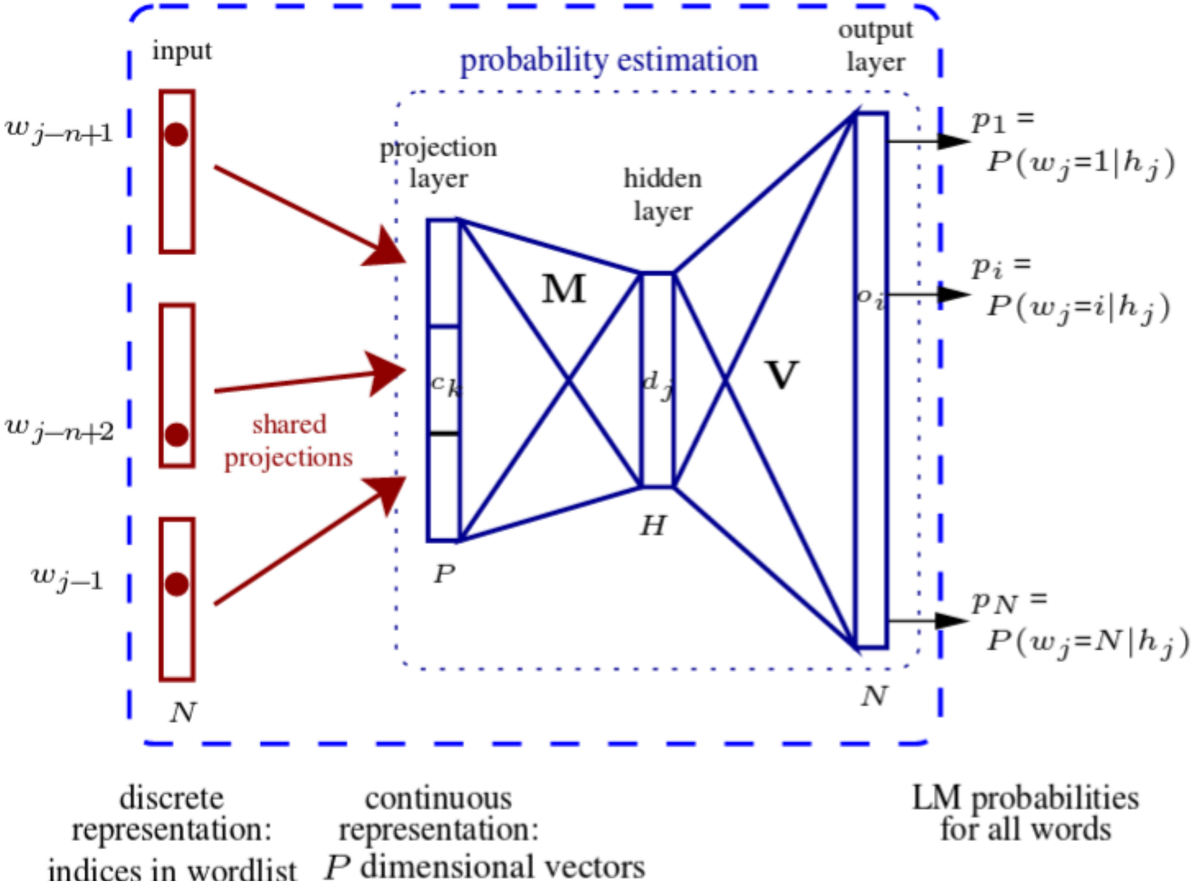


LM: FDNN v.s. N-gram



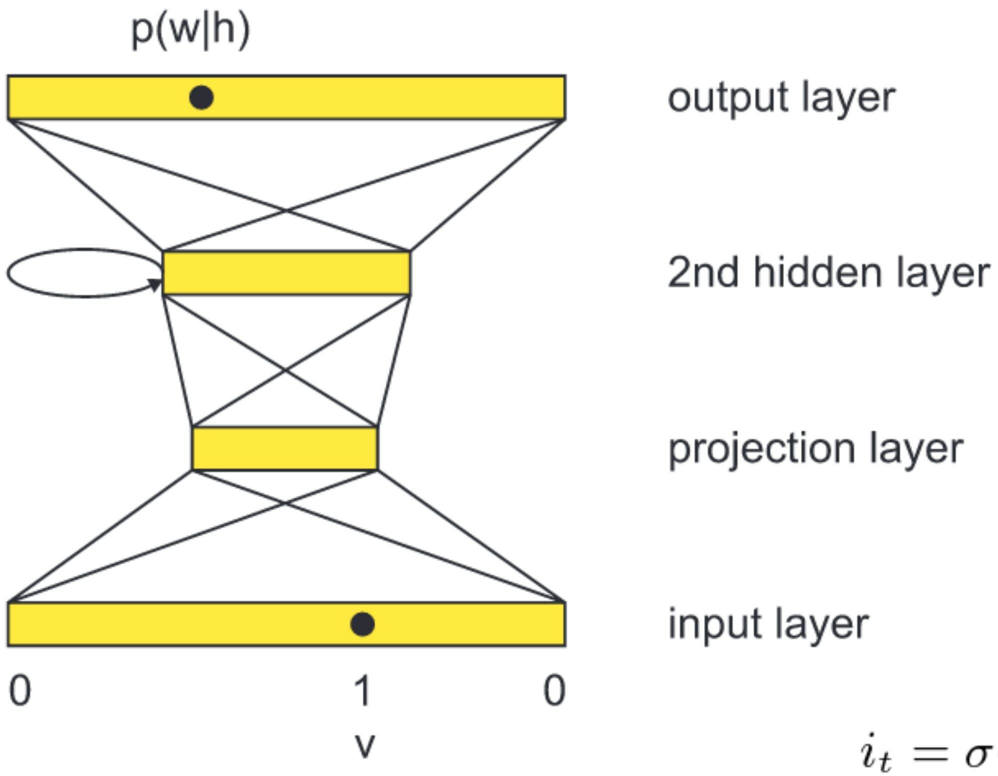
$$P(\mathbf{H}) = \prod_{k=1}^{K+1} P(w_k | w_1, \cdots, w_{k-1})$$

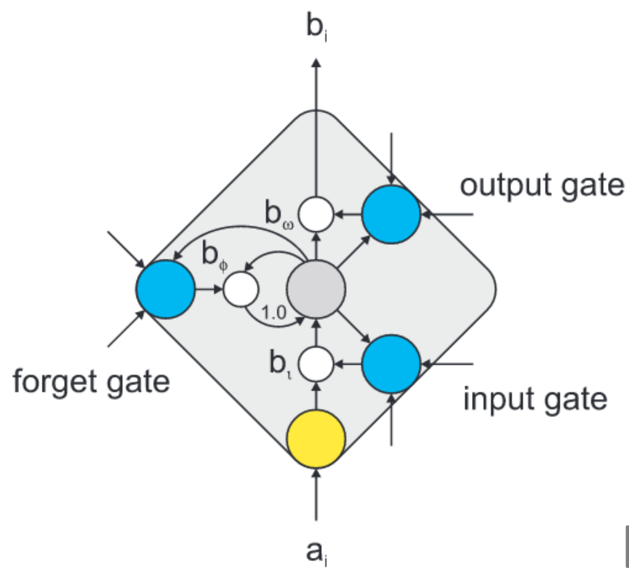
$$P(w_k | w_1, \cdots, w_{k-1}) = P(w_k | w_{k-1}, \cdots, w_{k-n+1})$$



[10]

LM: RNN and LSTM





[11]

$$i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1} + W_{ic}c_{t-1} + b_i)$$
$$f_t = \sigma(W_{fx}x_t + W_{fm}m_{t-1} + W_{fc}c_{t-1} + b_f)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g(W_{cx}x_t + W_{cm}m_{t-1} + b_c)$$
$$o_t = \sigma(W_{ox}x_t + W_{om}m_{t-1} + W_{oc}c_t + b_o)$$
$$m_t = o_t \odot h(c_t)$$
$$y_t = \phi(W_{ym}m_t + b_y)$$

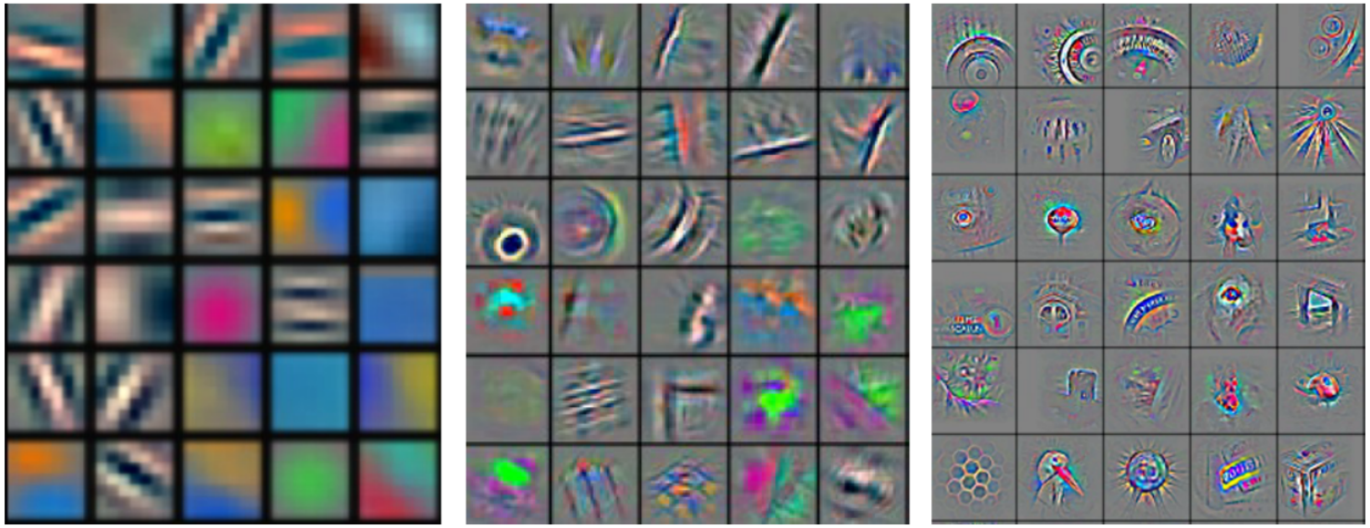
LM: Performance Improvement [12]

Model	Perplexity			Entropy reduction over baseline	
	individual	+KN5	+KN5+cache	KN5	KN5+cache
3-gram, Good-Turing smoothing (GT3)	165.2	-	-	-	-
5-gram, Good-Turing smoothing (GT5)	162.3	-	-	-	-
3-gram, Kneser-Ney smoothing (KN3)	148.3	-	-	-	-
5-gram, Kneser-Ney smoothing (KN5)	**141.2**	-	-	-	-
5-gram, Kneser-Ney smoothing + cache	**125.7**	-	-	-	-
PAQ8o10t	131.1	-	-	-	-
Maximum entropy 5-gram model	142.1	138.7	124.5	0.4%	0.2%
Random clusterings LM	170.1	126.3	115.6	2.3%	1.7%
Random forest LM	131.9	131.3	117.5	1.5%	1.4%
Structured LM	146.1	125.5	114.4	2.4%	1.9%
Within and across sentence boundary LM	116.6	110.0	108.7	5.0%	3.0%
Log-bilinear LM	144.5	115.2	105.8	4.1%	3.6%
Feedforward neural network LM [50]	140.2	116.7	106.6	3.8%	3.4%
Feedforward neural network LM [40]	141.8	114.8	105.2	4.2%	3.7%
Syntactical neural network LM	131.3	110.0	101.5	5.0%	4.4%
Recurrent neural network LM	124.7	105.7	97.5	5.8%	5.3%
Dynamically evaluated RNNLM	123.2	102.7	98.0	6.4%	5.1%
Combination of static RNNLMs	102.1	95.5	**89.4**	7.9%	7.0%
Combination of dynamic RNNLMs	101.0	**92.9**	90.0	8.5%	6.9%

What Issues Have DL Addressed?

- Hierarchical feature representation

Suitable for task



[13]

Low-Level Feature Mid-Level Feature High-Level Feature

Adaptation technique	CD-GMM-HMM (40-mixture)	CD-MLP-HMM (1 × 2,048)	CD-DNN-HMM (7 × 2,048)
Speaker independent	23.6 %	24.2 %	17.1 %
+ VTLN	21.5 % (−9 %)	22.5 % (−7 %)	16.8 % (−2 %)
+ fMLLR/fDLR×4	20.4 % (−5 %)	21.5 % (−4 %)	16.4 % (−2 %)

Deep Learning + Big Data

$\neq$ End_of_SLT_Research

- **Real world data is always non-homogeneous**

Structured Deep Learning

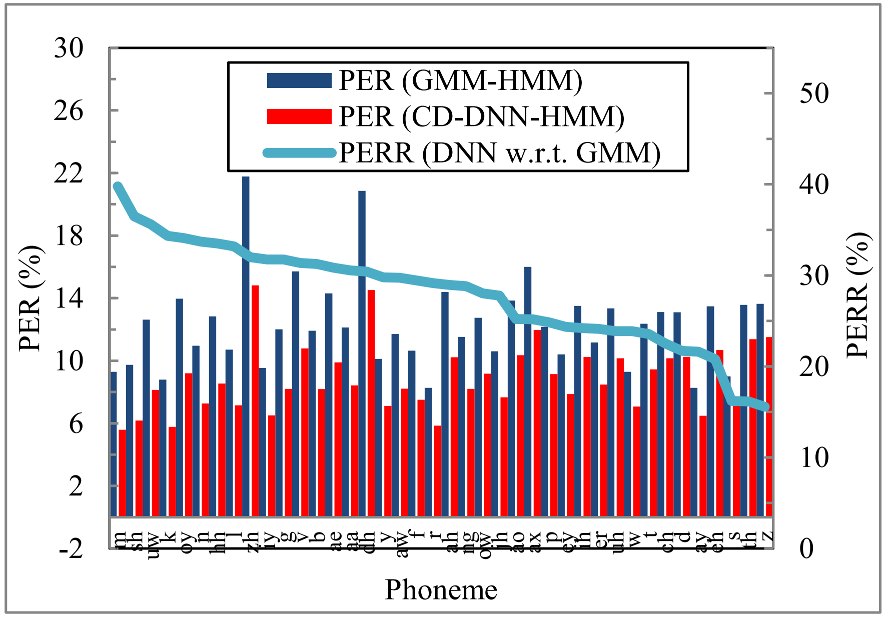
- **From data-driven to data+knowledge driven**

Prior knowledge incorporation

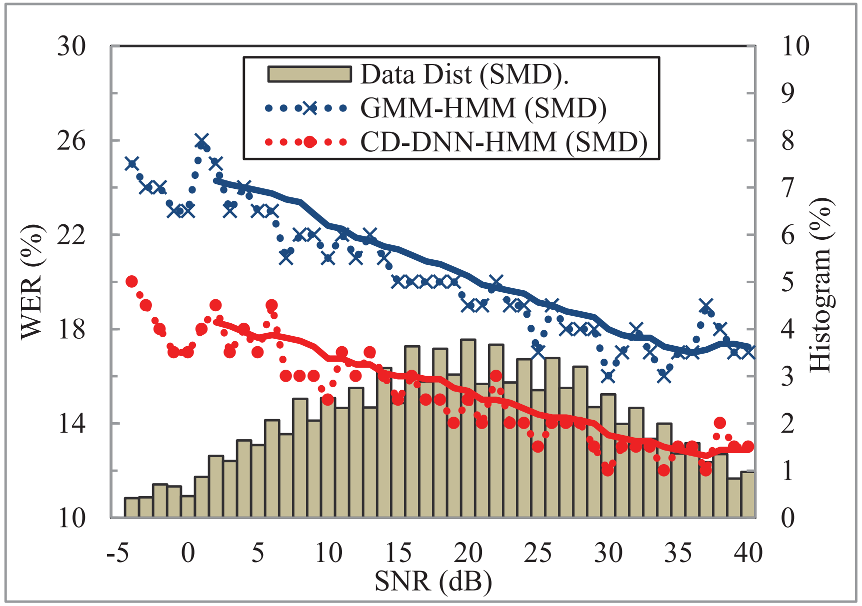
Part II
Multi-style and Context-aware Training

What Issues DL Have Not Addressed?

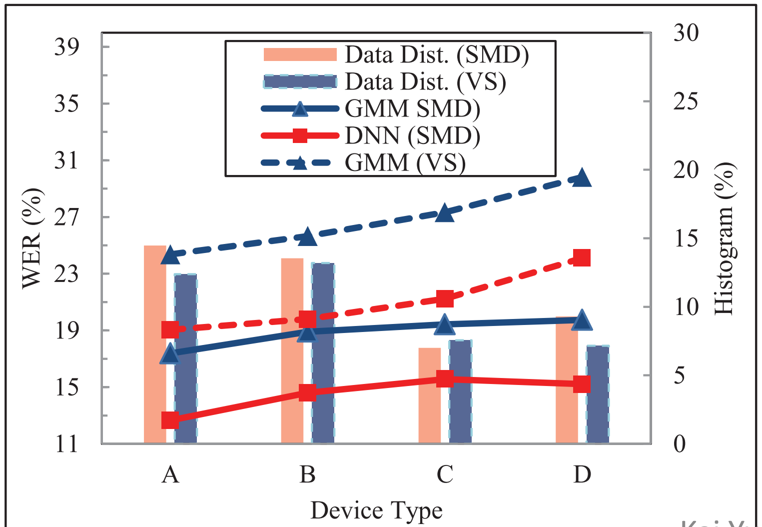
Phone Discrimination



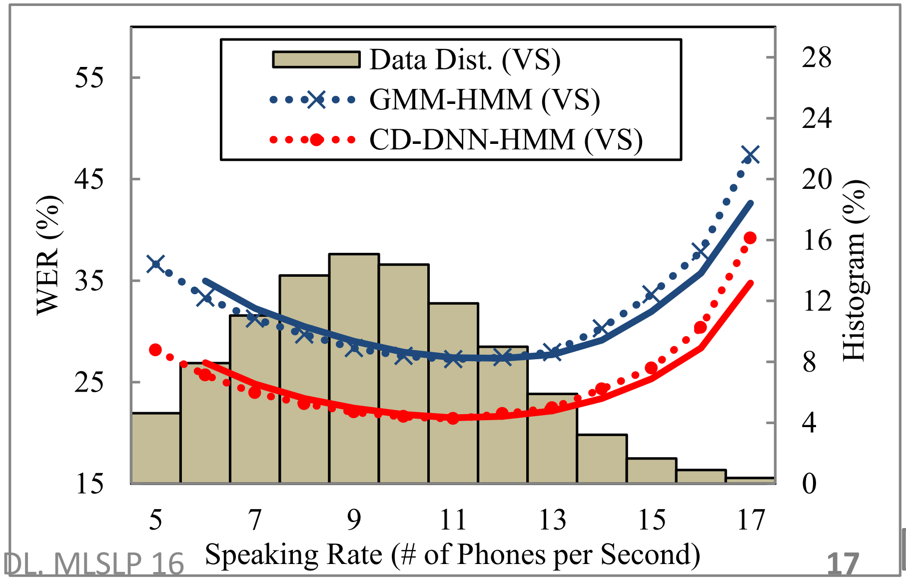
Noise Robustness



Channel Mismatch



Speaking Rate



[14]

Acoustic Variabilities for AM

- **Speech** variability – **desired**
  – Inherent variability related to what a speaker says

- **Acoustic context** variability – **unwanted**
  – Speaker: male/female, accent, speaking rate, etc.
  – Emotion: happy, fear, neutral, etc.
  – Spontaneity: read, natural, spontaneous, etc.
  – Environment: office, car, street, airport, etc.
  – Channel: mobile, microphone, bluetooth, etc.

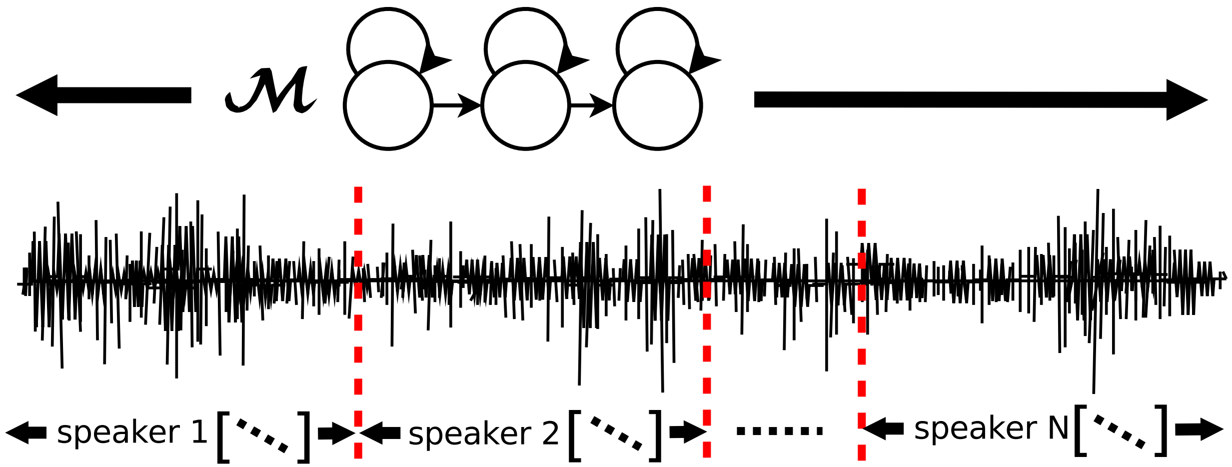
Linguistic Variabilities for LM

- **Word** variability – **desired**
  - Inherent variability related to what a speaker says
- **Linguistic context** variability – **unwanted**
  - Domain: news, science, novel, etc.
  - Topic: politics, sports, family, technology, etc.
  - Speaker role: child, parents, professional, etc.
  - Emotion: sad, happy, disgust, etc.
  - Dialogue: conversation history, search record, etc.

Context and Homogeneity

- **Effect**
  - Non-Targeted but Influential factors
- **Granularity**
  - Different from the primary variability
  - Usually beyond local estimation
- **Prior knowledge is a special kind of context**
  - E.g. telephone number constraint, etc.
- **Context is also structured**
  - E.g. environment, speaker, channel have intersection

Context and **Homogeneity**

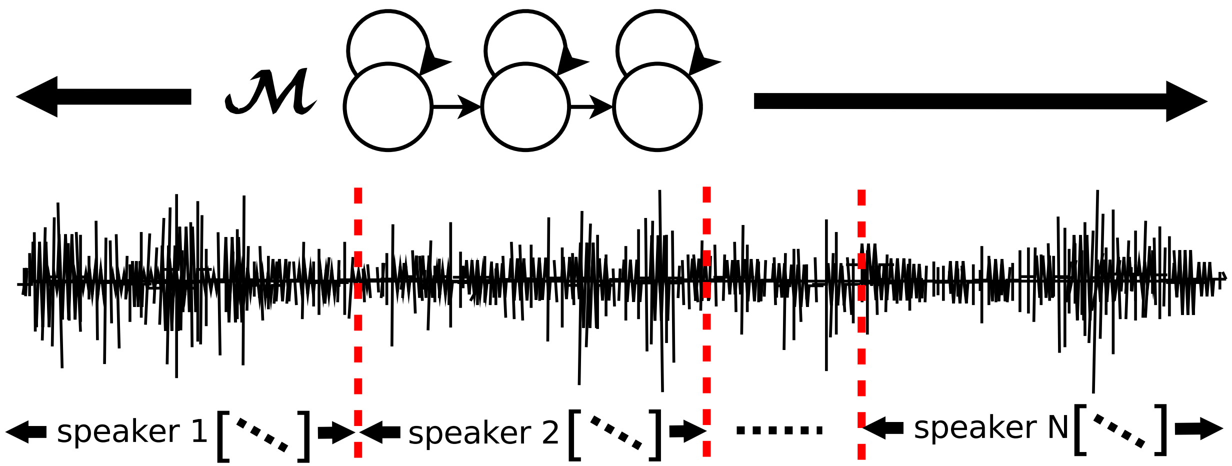


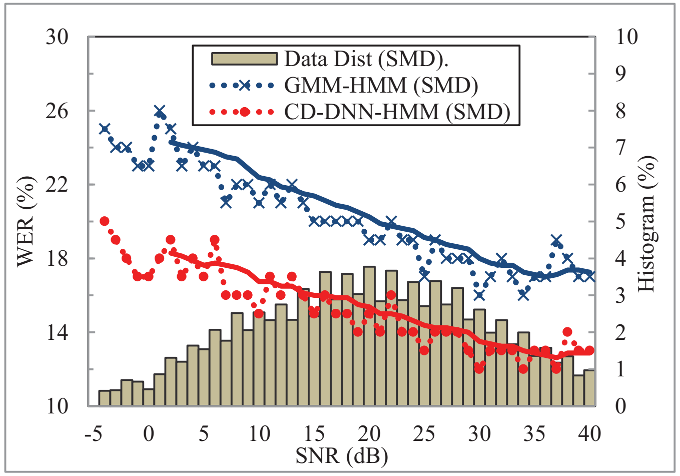
$$\mathcal{O} = \{\mathbf{O}^{(1)}, \cdots, \mathbf{O}^{(S)}\}$$

- Data within a homogeneous block share the same context (concrete specific statistical property)
- Homogeneous block is dependent on context
- **How to deal with non-homogeneity?**
  – Deep learning + big data?

Multi-style Training with Big Data

Multi-style training



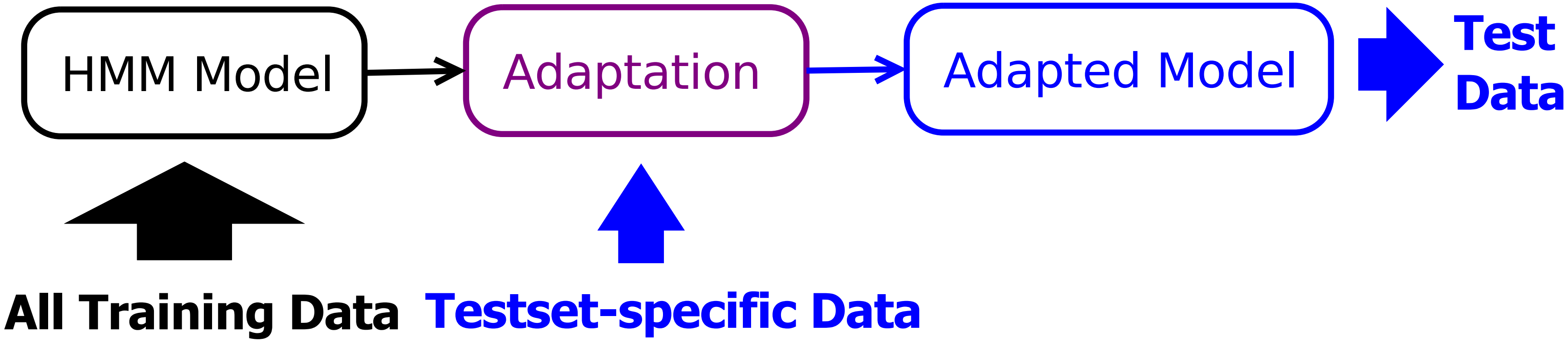


[14]

- Train models on data with large variability as if they are “homogeneous”, i.e. ignore mismatch inside training data
- Rely on good model and big data coverage
- **Big data ≠ rich context**
- **Implicit context modelling has limitation**
- **Explicit modelling: adaptation and adaptive training**

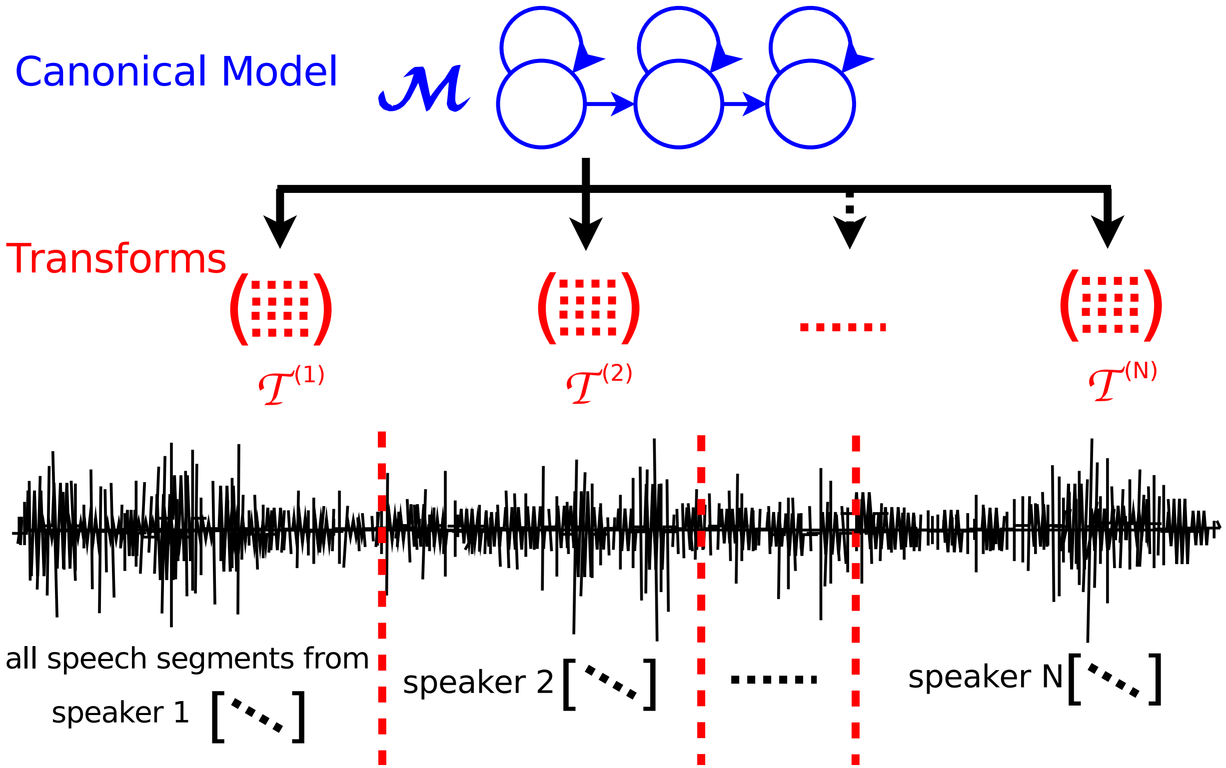
Adaptation – No Change in Training

- Well-built model already exists



- Mode
  - Supervised: annotations are available
  - Unsupervised: only raw data is available

Adaptive Training – Explicit Modelling of Context



Acoustic Modelling as an Example

Training data is split *homogeneous* blocks

$$\mathcal{O} = \left\{ \mathbf{O}^{(1)}, \cdots, \mathbf{O}^{(S)} \right\}$$

- Speech and non-speech variability separately modelled

GMM-HMM Adaptive Training Techniques

- Feature Normalization
  – Cepstral Mean and Variance Normalization
  – Gaussianization
  – Vocal Tract Length Normalization
- Model Adaptation
  – Linear transform based (MLLR, CMLLR etc.)
  – Cluster adaptive training

[15,16,17,18,19,20,21]

Feature Normalization [15][16]

- Simplest form: CMN/CVN

$$\hat{\mathbf{o}}_t^{\texttt{CMN}(s)} = \mathbf{o}_t^{(s)} - \bar{\mathbf{o}}^{(s)} = \mathbf{o}_t^{(s)} - \frac{1}{T}\sum_{i=1}^{T} \mathbf{o}_i^{(s)} \qquad \hat{o}_{t,d}^{\texttt{CVN}(s)} = \hat{o}_{t,d}^{\texttt{CMN}(s)} / \sqrt{\sigma_{dd}^{(s)}}$$

  - Homogeneous block varies from utterance to speaker or corpus
- Comparison to global CMN/CVN
  - Each homogeneous block has ***different*** feature transforms
  - Normalize training and test data ***separately***

Part III
Structured Deep Learning for Context Awareness

Context-aware Deep Learning

- **Re-training under context** – **Implicit** modelling

Basic Idea: Let the updated model be close to the original well-trained model

  – Additional Regularization
    - Conservative training [22]
    - KL-divergence regularization [23]
  – Selective update
    - Only update input/output layer [24]
    - Update weights connected to maximum variance nodes [25]

- **Structured deep learning** – **Explicit** modelling
  – **Multi-view** input with context representation
  – **Multi-task** output with context target or constraint
  – **Structured model** parameter to reflect context

Structured Input – Multi-view Techniques

- External context embedding as input

Basic Idea: Estimate context representation using an external model and augment feature with the context representation

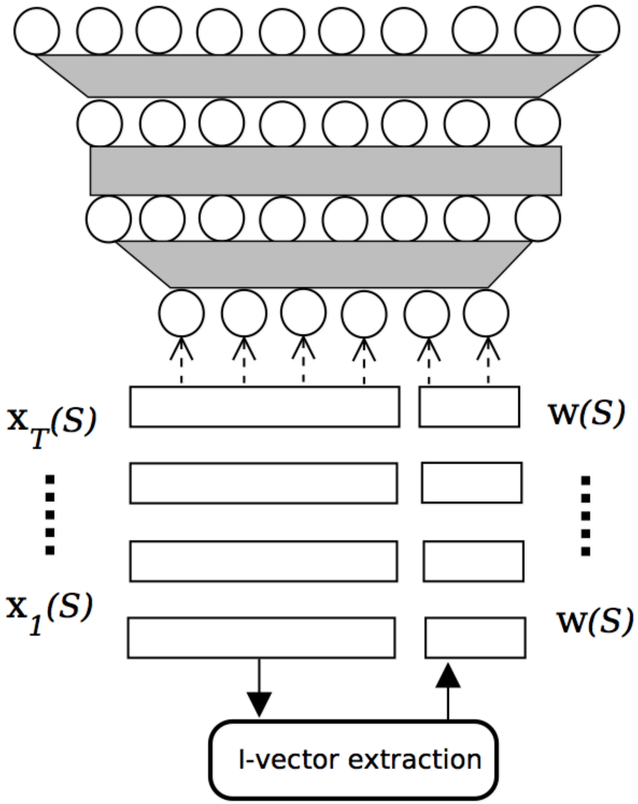
  - Speaker feature: i-vector [26], i-vector for LSTM [31]
  - Environment-feature or combined:noise-energy [27], combined [28]
  - Structured VAD [32]

- Internal trainable context embedding

Basic Idea: Context representation is estimated using the same model for speech recognition

  - Speaker-code [29]
  - Paragraph-vector [30]

i-Vector-based DNN Adaptation[26,33]

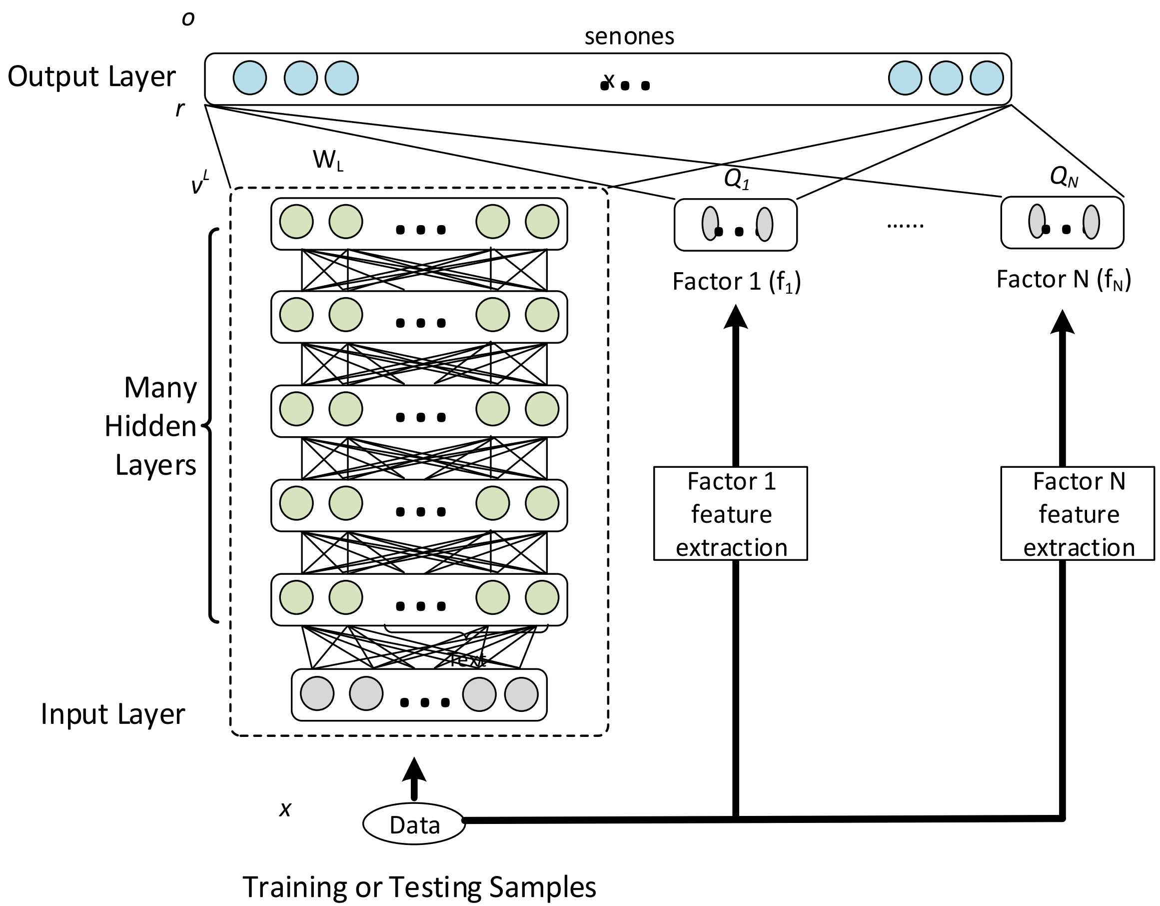


Crit.	Sys	Hub5	RT03-Fsh	RT-03-Swb
CE	DNN	16.1	18.9	29.0
	+iVec	**13.9**	**16.7**	**25.8**
Sequence	DNN	14.1	16.9	26.5
	+iVec	**12.4**	**15.0**	**24.0**

Conversational Telephone Speech (English)

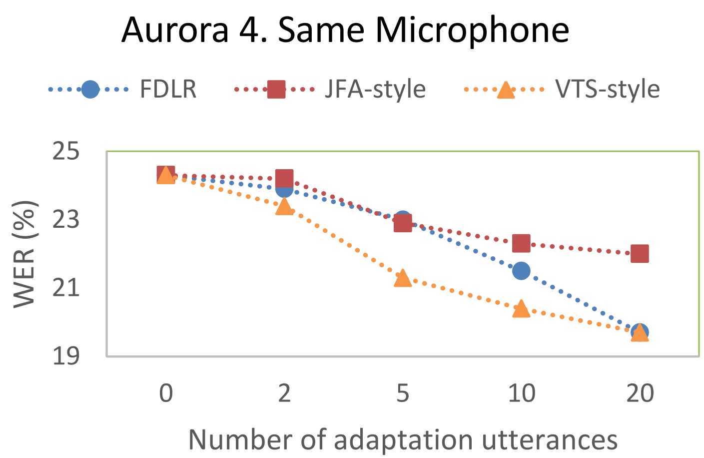
- i-Vector encapsulates all relevant information about a speaker's identity in a low-dimensional vector
- i-Vector extraction is independent of DNN training
- Speaker-level i-Vector is fed into DNN together with frame-level features as augmented input features
- Noise-vector (energy) can also be used in the framework [27]

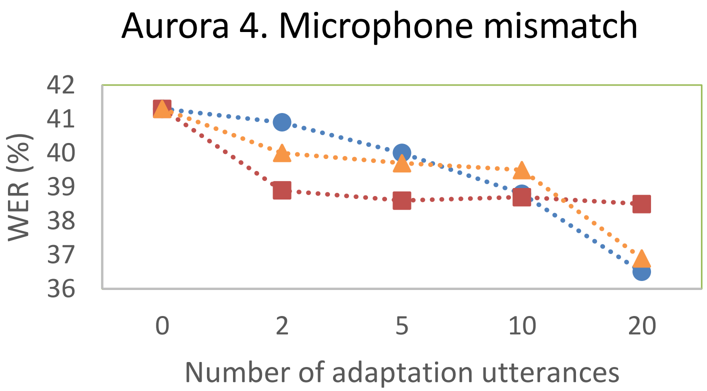
Factored DNN Adaptation [28]



Factor vector estimation

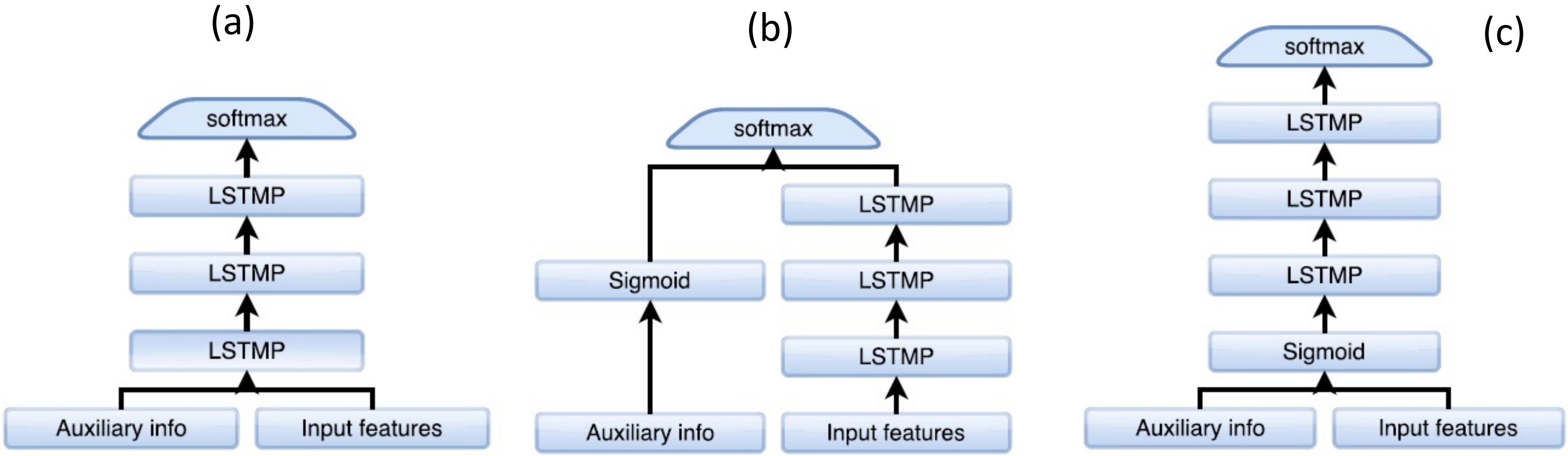
- Joint factor analysis (JFA) [34]
- Vector Taylor Series (VTS) [35,36]





- Factors are estimated separately from DNN
- Factor vectors fed to the softmax output layer
- Loading matrix (weights connected to acoustic factor vectors) need to be estimated using standard BP

Speaker-aware training on LSTM-RNNs [31]



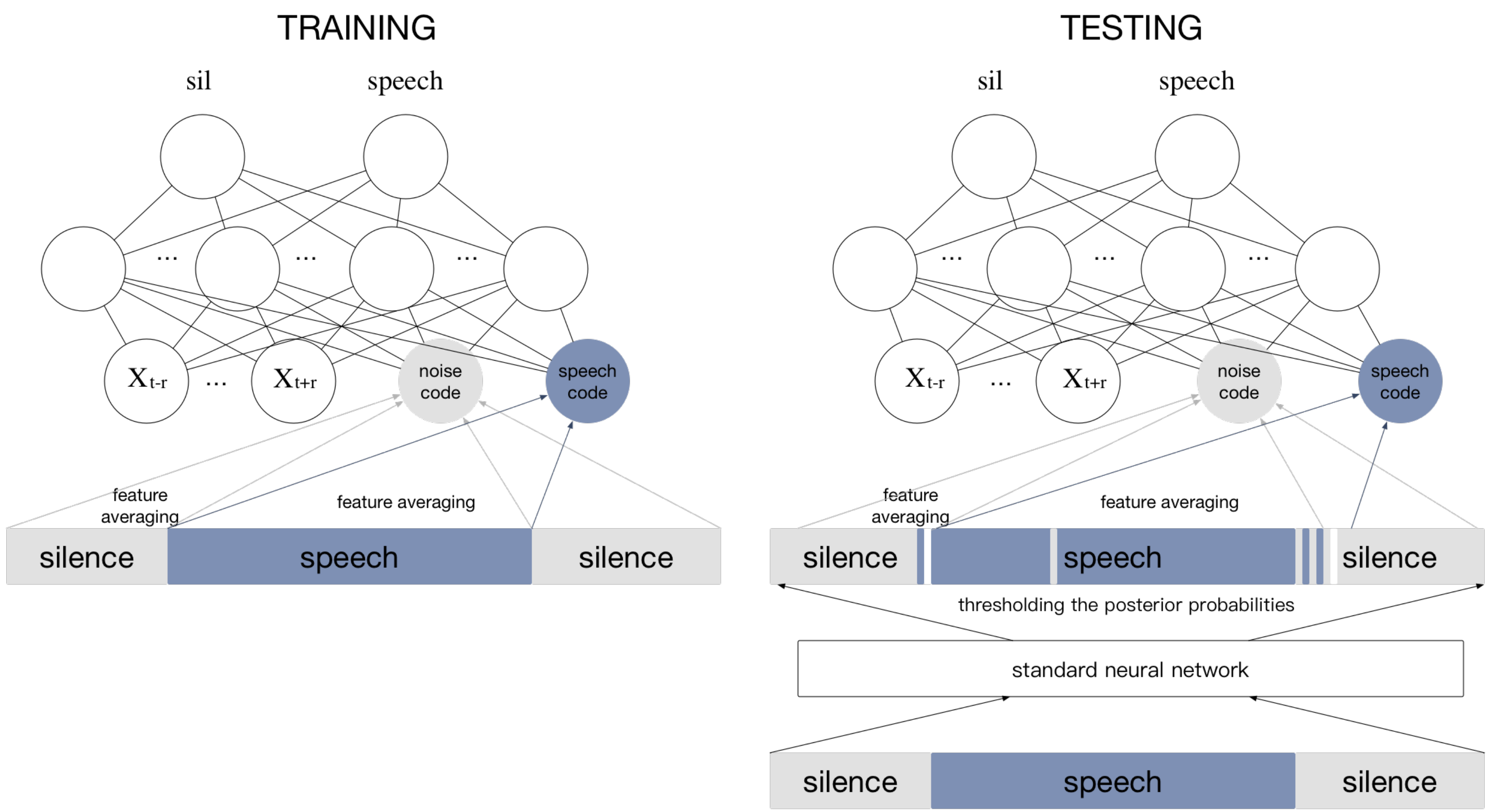
AMI ihm subset

Model	Structure	WER
LSTMP	-	32.4
+ i-vector	(a)	**31.1**
	(b)	33.2
	(c)	34.5

AMI ihm full set

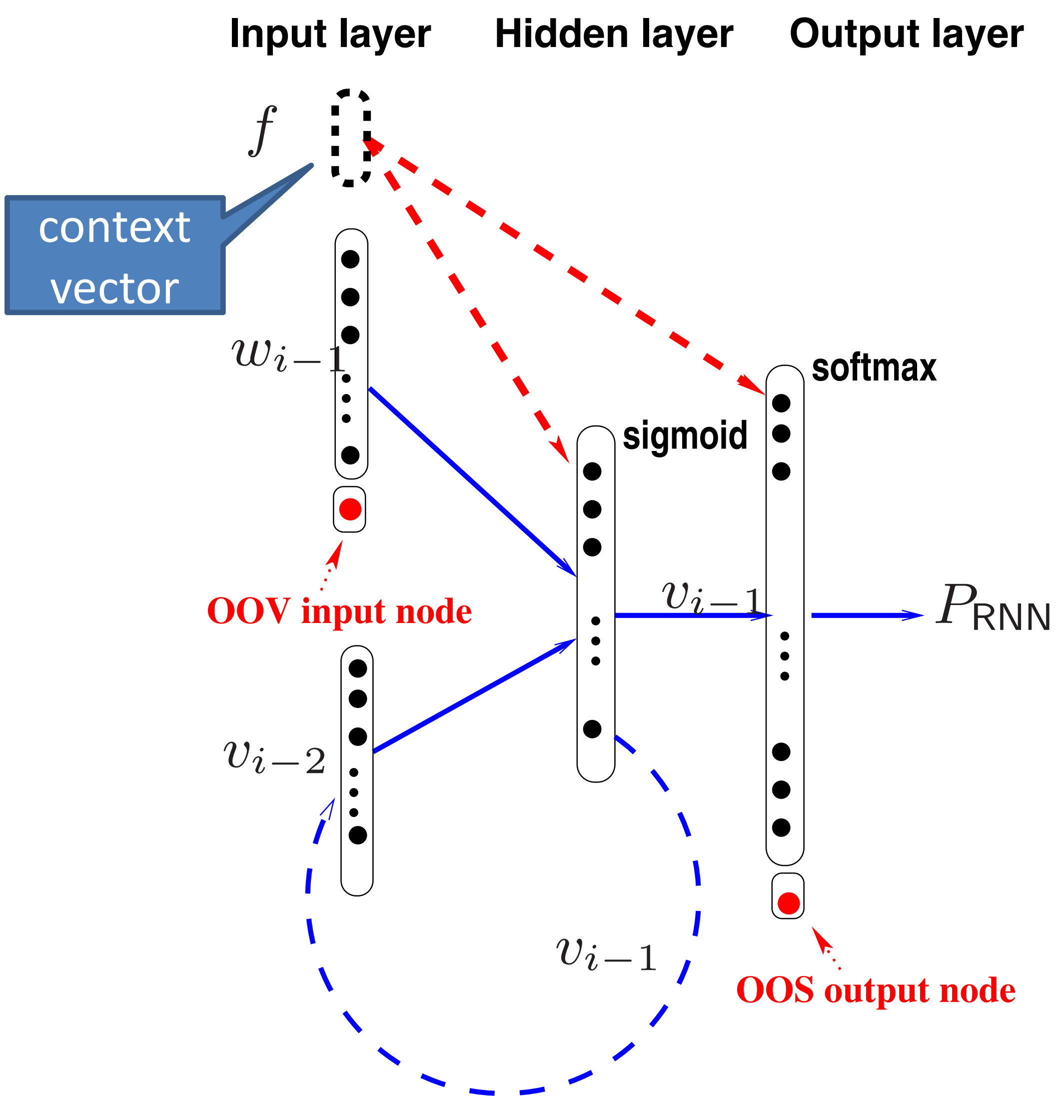
Feature	WER
FMLLR	26.0
+ i-vector + BSV + speaking rate	**24.3** 25.0 25.7

Noise-aware VAD [32]



TEST SET	NAT-DNN	NAT-LSTM	NAT-CNN
SEEN NOISE	(3.52)->3.14	(3.15)->2.82	(5.05)->3.30
UNSEEN NOISE	(11.19)->8.58	(9.24)->6.72	(9.76)->7.14

RNN Language Model Adaptation[37]

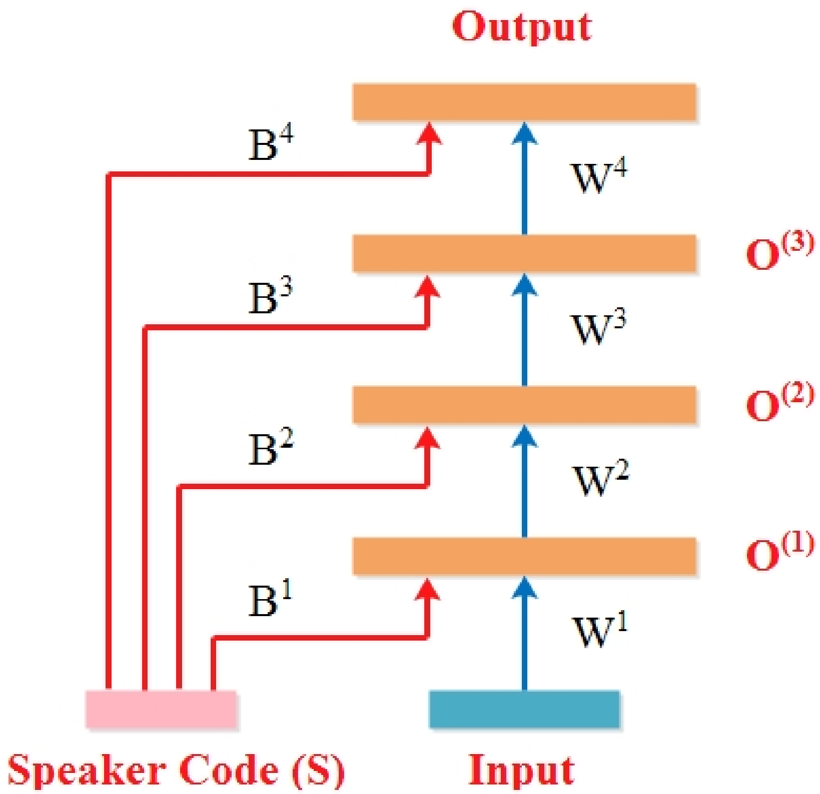


Unsupervised Topic *f* extraction:

- **LDA**
- **PLSA**
- **HDP**

Topic M	Sup	PPL		WER
		RNN	+4g	
-	-	152.5	113.5	31.38
PLSA	hyp	137.8	106.3	31.16
	ref	137.3	105.1	31.08
LDA	hyp	133.7	105.0	31.14
	ref	134.1	104.2	31.07
HDP	hyp	138.9	106.6	31.19
	ref	138.0	105.2	31.10

Speaker Code for DNN Adaptation [29]



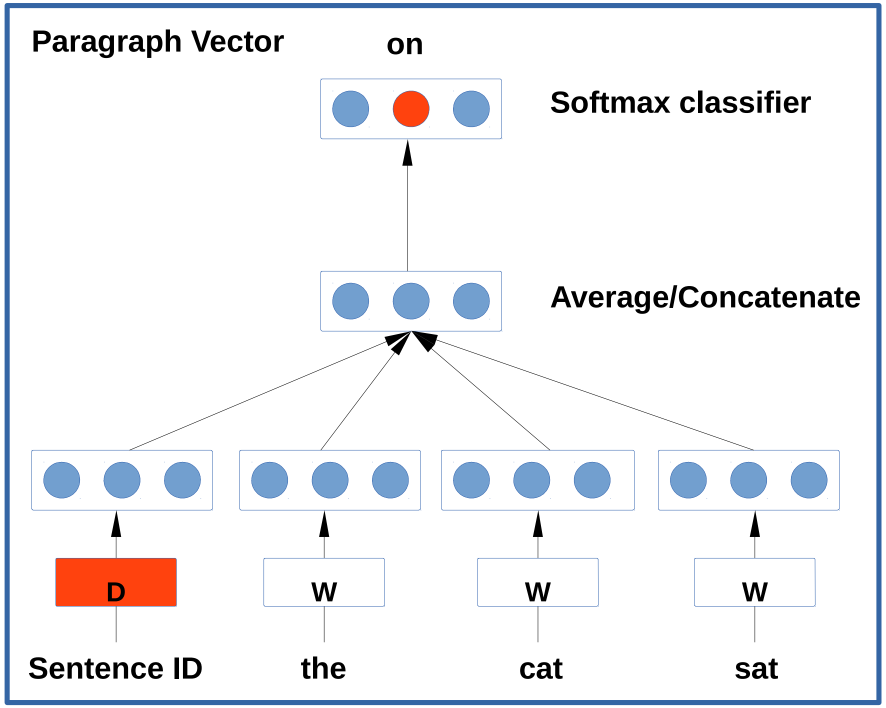
Use 10 adaptation utterances for speaker code estimation

Crit.	Additional Training Data for Speaker Code Connection	WER
CE	0	16.2
	10%	15.8
	100%	15.2
Sequence	0	14.0
	10%	13.7
	100%	13.4

- Speaker code is a vector embedding speaker identity using DNN
- Speaker code and connection weights are randomly intialized and updated using standard BP during training
- Smaller set of training data may be used to train B
- During adaptation, only the speaker code is updated on adaptation data and re-decoding is then performed

Paragraph Vector [30]

Distributed Memory Model



Model	Error rate (Positive/ Negative)	Error rate (Fine-grained)
Naïve Bayes (Socher et al., 2013b)	18.2 %	59.0%
SVMs (Socher et al., 2013b)	20.6%	59.3%
Bigram Naïve Bayes (Socher et al., 2013b)	16.9%	58.1%
Word Vector Averaging (Socher et al., 2013b)	19.9%	67.3%
Recursive Neural Network (Socher et al., 2013b)	17.6%	56.8%
Matrix Vector-RNN (Socher et al., 2013b)	17.1%	55.6%
Recursive Neural Tensor Network (Socher et al., 2013b)	14.6%	54.3%
Paragraph Vector	**12.2%**	**51.3%**

- Paragraph matrix is shared for all words within the paragraph
- Standard BP training can be used to estimate paragraph vector
- Not used for language model adaptation yet.

Structured Output – Multi-task Training

- Multi-task training

Basic Idea: Jointly estimate the target-of-interest and context-related task, expecting context is embedded during deep feature extraction

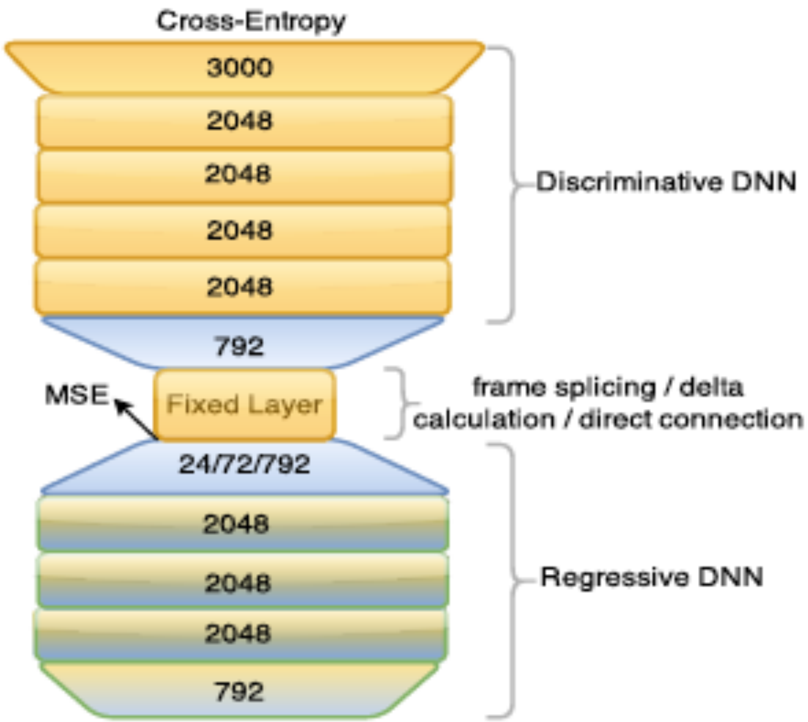
  – Multi-task training for text-dependent speaker verification
  – Multi-task joint training for robust ASR

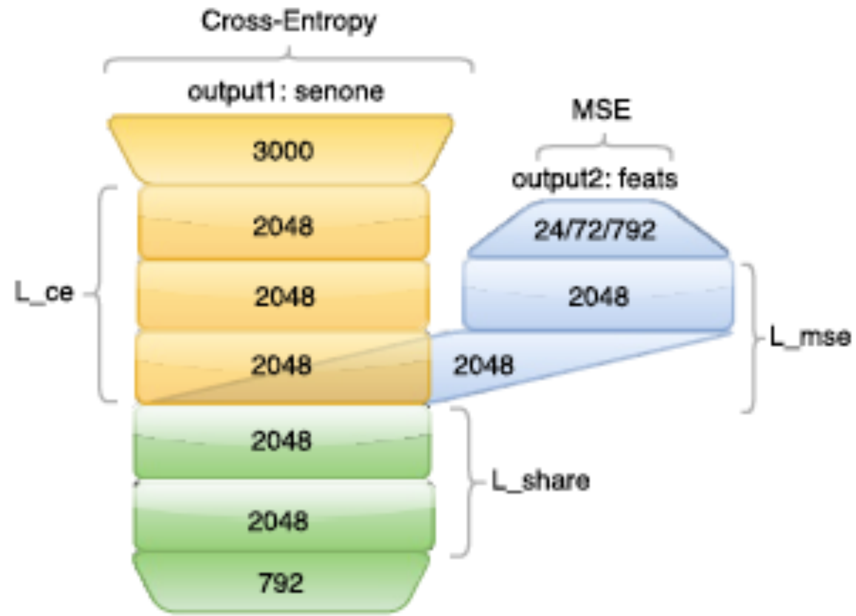
- Multi-view and multi-task combination

Basic Idea: Reinforce context modelling by combining both input and output context representation

  – Multi-factor training for robust ASR

Multi-task Joint Learning for Robust ASR [40]



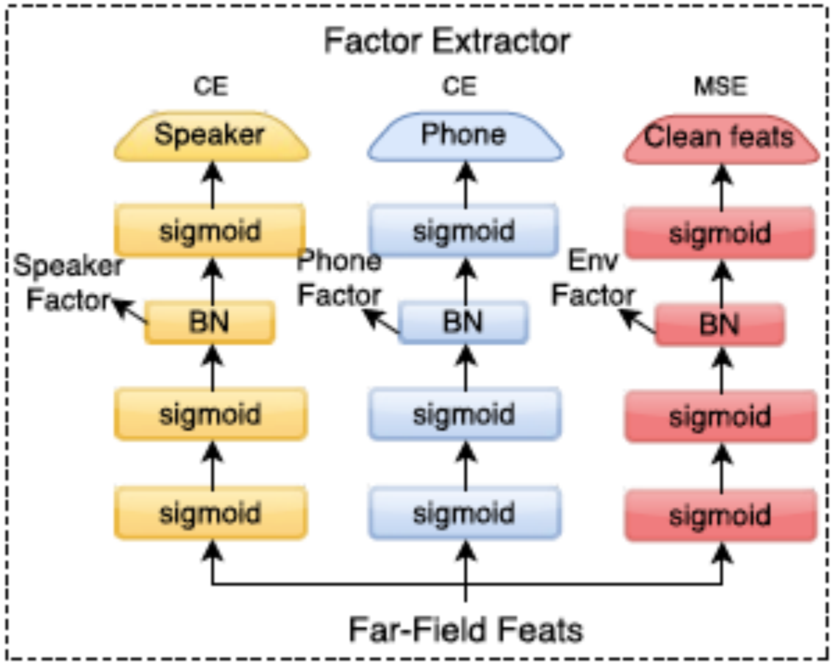


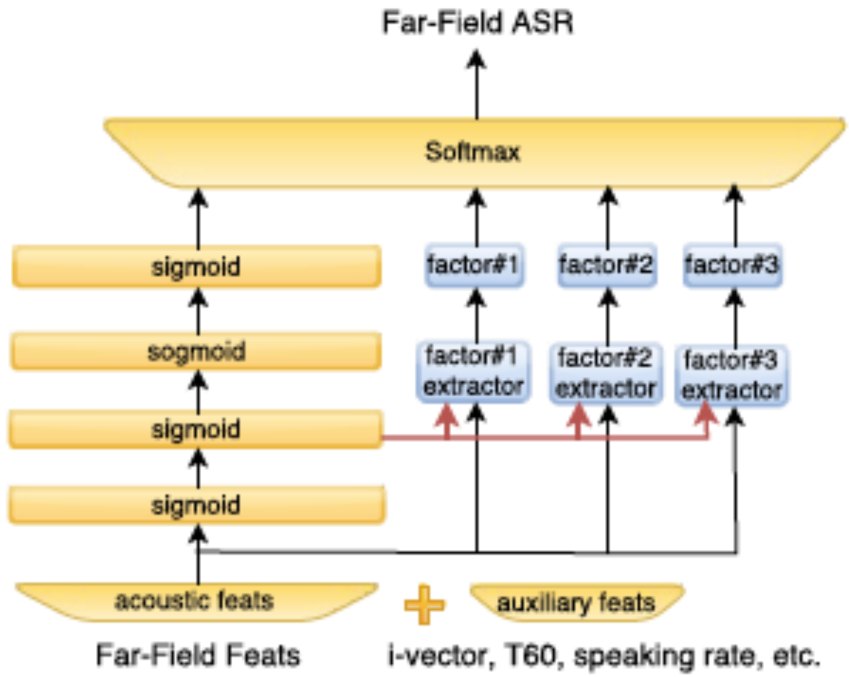
Aurora 4 Result

System	A	B	C	D	AVG
Baseline	4.6	8.2	8.8	18.5	12.4
MT Joint Learning	3.9	6.7	6.3	15.4	10.2
+ NAT	3.8	6.5	6.0	14.5	9.7

System	A	B	C	D	AVG
Best GMM-HMM [9]	5.6	11.0	8.8	17.8	13.4
DNN NAT DP [10]	5.4	8.3	7.6	18.5	12.4
DNN PP [15]	4.5	7.5	7.4	19.3	12.3
Spectral Mask [27]	4.5	7.9	7.5	17.7	11.4
JNAT [14]	4.5	7.4	8.1	16.5	11.1
TVWR Adap [6]	4.4	7.5	7.1	15.6	10.7
Joint FE BE [18]	4.4	6.8	6.4	15.4	10.3
AD OSN LRF [19]	4.0	7.2	6.4	14.5	10.0
MT Joint-learning	3.8	6.5	6.0	14.5	9.7

Multi-factor Joint Training [41]





System	Factor	Integration	WER(%)
DNN	—	—	65.2
MF-DNN	Speaker	Output	61.6
		+X-connection	**61.0**
	Phone	Output	61.4
		+X-connection	**60.8**
	Env	Output	61.2
		+X-connection	**60.7**
	Spk+Phn+Env	Output	60.4
		+X-connection	**60.1**

AMI Far-field Dataset

System	Sub Set	Full Set
DNN	65.2	55.9
DNN+i-vector	62.5	52.0
MF-DNN	60.1	53.5
MF-DNN+i-vector	**57.1**	**50.0**

Structured Model – Context-Specific Structure

- Context-specific linear transform

Basic Idea: Apply speaker-specific linear transform to normalize hierarchical deep features

  - Input/output feature transform [42][43][46][47][49]
  - Hidden layer feature transform [44][45]

- Explicit context-specific structure

Basic Idea: Construct context specific subspace within deep learning models

  - Context-dependent layer [48]
  - Additional or factorized structures [49][50][51][52][38]

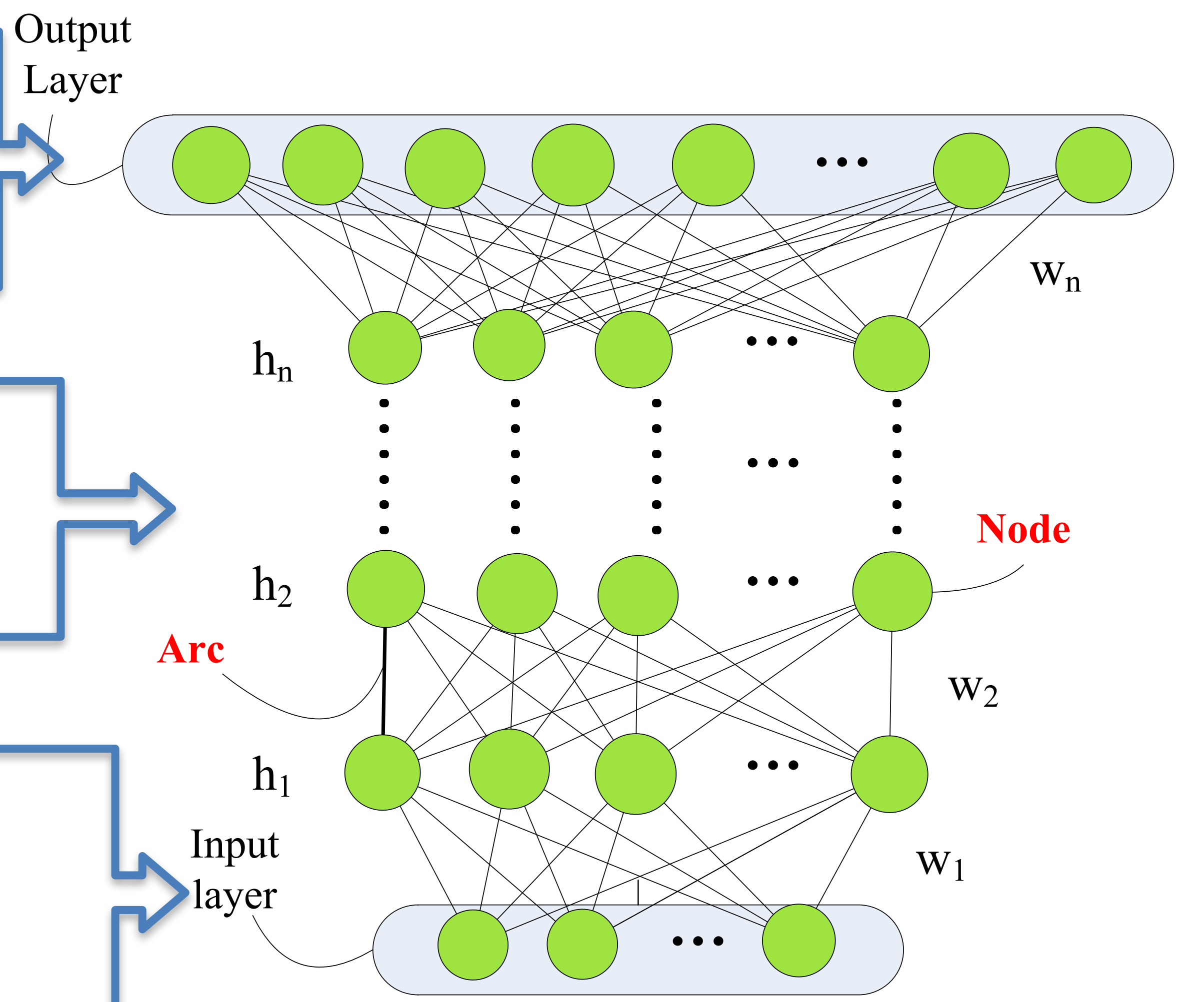
Transfer Linear Transform Adaptation to DNN

Basic Idea: Apply context-specific linear transform to normalize features of DNN

- Output feature discriminative linear transform (oDLR) [42]
- Linear output network (LON) [43]

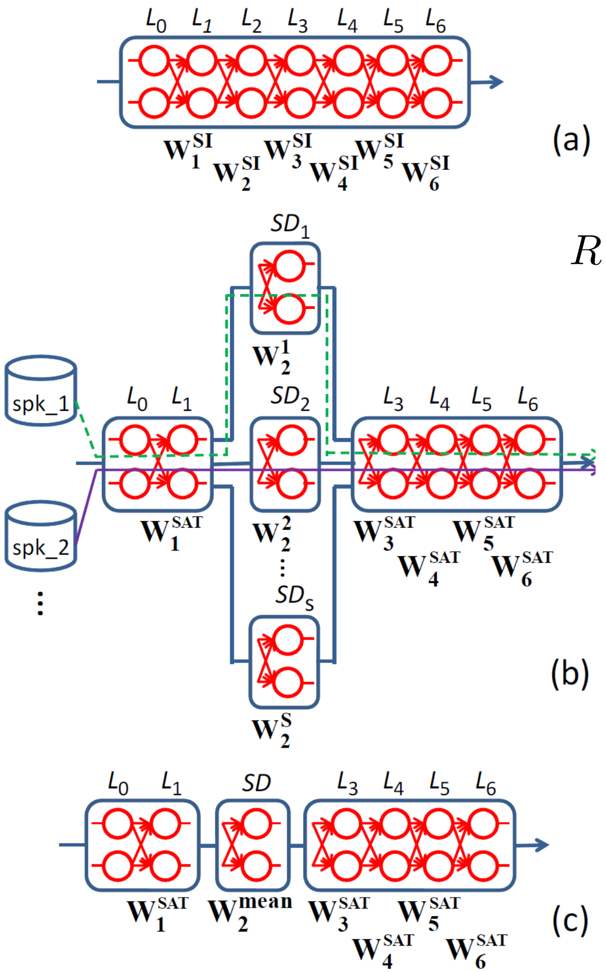
- SVD bottleneck adaptation [44]
- Linear hidden network [45]

- Feature discriminative linear transform (fDLR) [46]
- Linear input network (LIN) [47,49]



Adaptive Training with Context-specific Layer [48]

Basic Idea: Split DNN into context-dependent and context-independent layers and interleavingly update them



- Regularization is required for speaker-dependent layer update

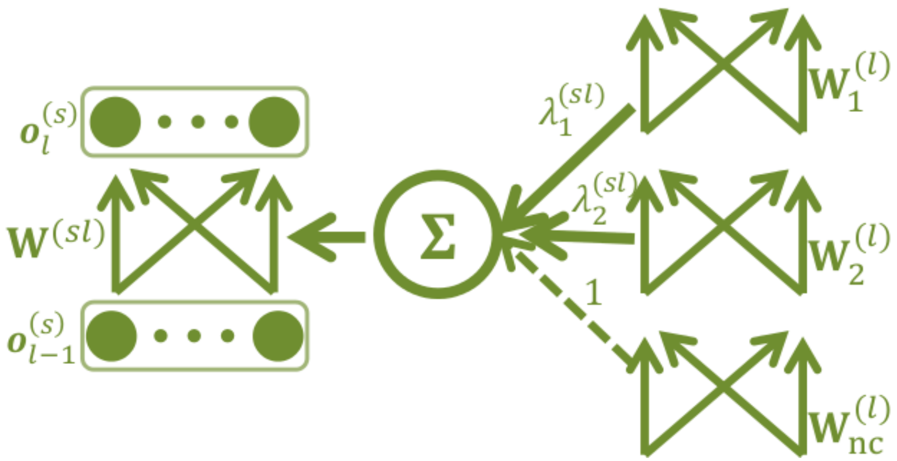
$$R(\Lambda) = \frac{1}{2}\left\|\mathbf{W}_{l_{\text{SD}}}^{t} - \mathbf{W}_{l_{\text{SD}}}^{\texttt{mean}}\right\|_2^2 + \frac{1}{2}\left\|\mathbf{b}_{l_{\text{SD}}}^{t} - \mathbf{b}_{l_{\text{SD}}}^{\texttt{mean}}\right\|_2^2 \quad (t = 1, 2, 3, \cdots, T)$$

- Performance

l_{SD}	SI	Adaptation	SAT
1	26.4	20.0	18.9
2	26.4	19.0	18.2
3	26.4	18.7	18.0
4	26.4	19.0	18.4
5	26.4	19.5	19.0

TED Talks corpus, supervised adaptation

Cluster Adaptive Training [49]



DNN

$$\boldsymbol{y_l} = \boldsymbol{W}^{(l)}\boldsymbol{o_{l-1}} + \boldsymbol{b}^{(l)}$$

$$\boldsymbol{o_l} = \boldsymbol{\sigma}(\boldsymbol{y_l})$$

CAT-DNN

$$\boldsymbol{y}_l^{(s)} = \boldsymbol{W}^{(sl)}\boldsymbol{o}_{l-1}^{(s)} + \boldsymbol{b}^{(l)}$$

$$\boldsymbol{o_l} = \boldsymbol{\sigma}(\boldsymbol{y_l})$$

$$\mathbf{M}^{(l)} = \left[\boldsymbol{W}_1^{(l)} \dots \boldsymbol{W}_P^{(l)}\right]$$

$$\boldsymbol{\lambda}^{(sl)} = \left[\boldsymbol{\lambda}_1^{(sl)} \dots \boldsymbol{\lambda}_P^{(sl)}\right]^{\top}$$

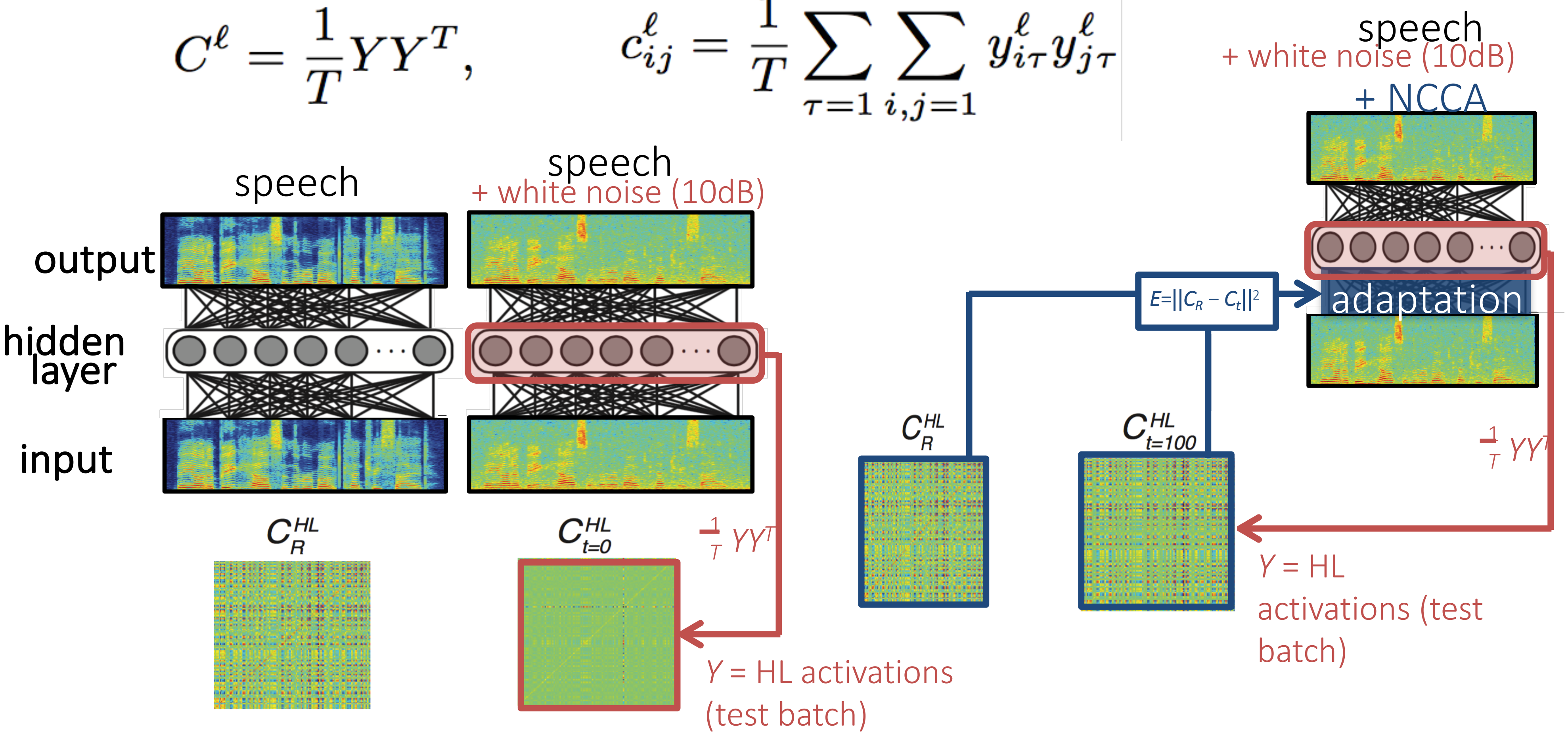
$$\mathbf{W}^{(sl)} = \sum_{c=1}^{P} \lambda_c^{(sl)} \boldsymbol{W}_c^{(l)}$$

System	Cluster	#Adapt Para	swb	fsh
SI	-	0	15.8	19.9
+ i-vector	-	100	14.8	18.3
H1	2	2	15.2	18.8
	5	5	15.0	18.7
	10	10	**14.6**	**17.8**

Unsupervised Adaptation with Node Co-activation Prior [38]

Node co-activation matrix: captures the correlational structure of nodes over time

$$C^{\ell} = \frac{1}{T} Y Y^{T}, \qquad c_{ij}^{\ell} = \frac{1}{T} \sum_{\tau=1}^{T} \sum_{i,j=1}^{N} y_{i\tau}^{\ell} y_{j\tau}^{\ell}$$



Sequence summarizing [50]

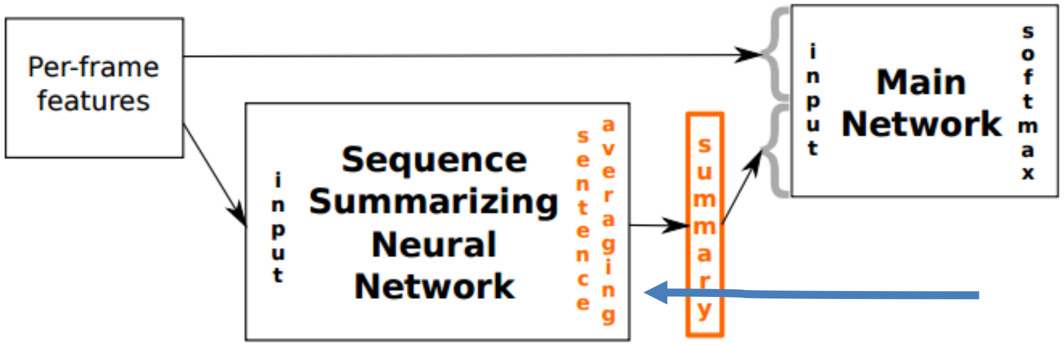
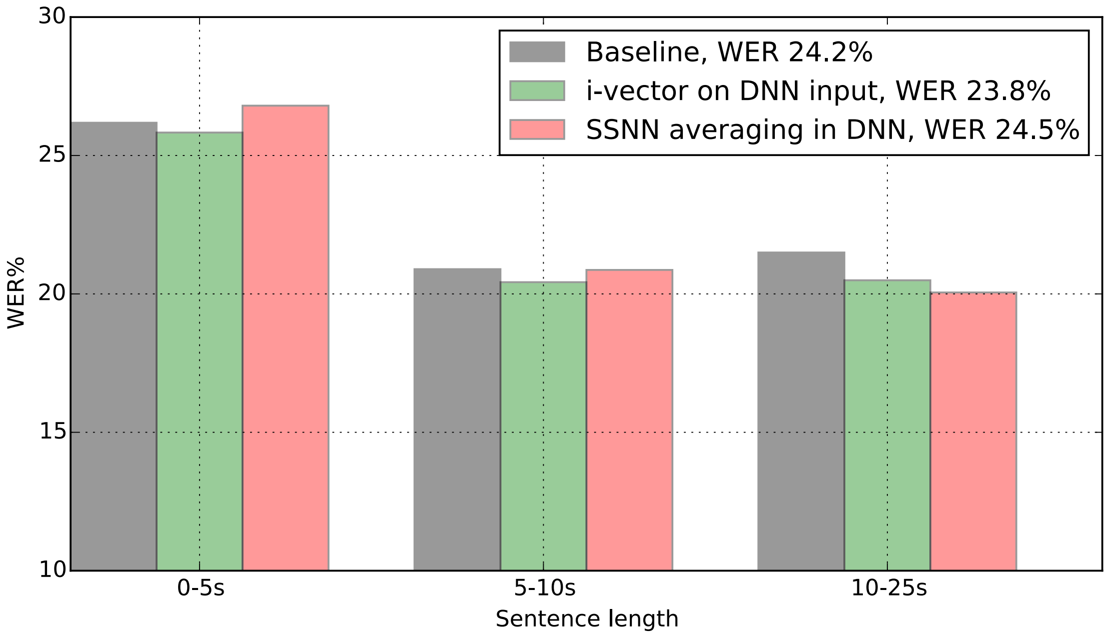


Fig. 1. *Topology of main-network with "sequence summary" input. The summary is computed by Sequence Summarizing Neural Network with sentence-averaging on the output.*

$$\bar{\mathbf{o}}_{u,t} = [\mathbf{o}_{u,t} \quad \mathbf{x}_u]^\top$$

$$\mathbf{x}_u(\mathbf{o}_u;\theta_x) = \frac{1}{T_u}\sum_{t=1}^{T_u}\mathbf{x}_u(\mathbf{o}_u;\theta_x)$$



- Sequence sum network estimates utterance level context representation
- Joint training of summarizing NN and main NN
- Sequence level BP
- Better for long sentence

Context Modelling with Structured Learning

- Information rate modelling
- Easy prior knowledge incorporation
- Explicit structure related to context effect
- Unsupervised on-line context learning
- Text based context modelling

Summary

- Context is the non-targeted but influential factors, which may have different information rates
- Context modelling is an unsolved issue for DL
- **Re-training under context** – **Implicit** modelling
- **Structured deep learning** – **Explicit** modelling

Multi-view input with context representation

Multi-task output with context target or constraint

Structured model parameters to incorporate context

Acknowledgement

- The speaker would like to thank the students of the SJTU speech lab, Yimeng Zhuang, Tian Tan and Tianxing He, who helped to prepare the slides.

- Some images/tables in this slides are directly taken from the cited research papers. The authors of these papers are appreciated.

- Some figures are redrawn based on lecture notes/slides from the web.The original authors are also appreciated.

References

[1] Yu, Dong, and Li Deng. "Deep learning and its applications to signal and information processing [exploratory dsp]." IEEE Signal Processing Magazine 28.1 (2011): 145-154.

[2] Graves, Alex, Navdeep Jaitly, and Abdel-rahman Mohamed. "Hybrid speech recognition with deep bidirectional LSTM." Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE, 2013.

[3] Graves, Alex, and Navdeep Jaitly. "Towards End-To-End Speech Recognition with Recurrent Neural Networks." ICML. Vol. 14. 2014.

[4] Yanmin Qian, Mengxiao Bi , Tian Tan, Kai Yu. "Very Deep Convolutional Neural Networks for Noise Robust Speech Recognition“IEEE/ACM Transactions on Audio, Speech, and Language Processing . IEEE, 2016.

[5] Sainath, Tara N., et al. "Convolutional, long short-term memory, fully connected deep neural networks." 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015.

[6] Hannun, Awni, et al. "Deep speech: Scaling up end-to-end speech recognition." arXiv preprint arXiv:1412.5567 (2014).

[7] Sercu, Tom, et al. "Very deep multilingual convolutional neural networks for LVCSR." 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.

[8] Seltzer, Michael L., Dong Yu, and Yongqiang Wang. "An investigation of deep neural networks for noise robust speech recognition." 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013.

[9] Yu, Dong, et al. "Feature learning in deep neural networks-studies on speech recognition tasks." arXiv preprint arXiv:1301.3605 (2013).

[10] Schwenk, Holger, Anthony Rousseau, and Mohammed Attik. "Large, pruned or continuous space language models on a gpu for statistical machine translation." Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT. Association for Computational Linguistics, 2012.

[11] Sundermeyer, Martin, Ralf Schlüter, and Hermann Ney. "LSTM Neural Networks for Language Modeling." Interspeech. 2012.

[12] Mikolov, Tomáš. "Statistical language models based on neural networks." Presentation at Google, Mountain View, 2nd April (2012).

[13] Dong Yu. "Deep Learning and Its Applications in Signal Processing." ICASSP Tutorial. 2012

[14] Huang, Yan, et al. "A comparative analytic study on the Gaussian mixture and context dependent deep neural network hidden Markov models." INTERSPEECH. 2014.

[15] B. S. Atal. "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identication and verication." Journal of the acoustical society of America, 55(6):1304.1312, 1974.

[16] P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young. "The development of the 1994 HTK large vocabulary speech recognition system." In ARPA Workshop on Spoken Language Systems Technology, pages 104.109, 1995.

[17] G. Saon, A. Dharanipragada, and D. Povey. "Feature space Gaussianization." In Proc. ICASSP, 2004.

[18] L. Lee and R. C. Rose. "Speaker normalization using efcient frequency warping procedures." Proc. ICASSP, 1:353.356, 1996.

[19] M. J. F. Gales. "Maximum likelihood linear transformations for HMM-based speech recognition." Computer Speech and Language, 12:75.98, 1998.

[20] M. J. F. Gales. Cluster adaptive training for speech recognition. In Proc. ICSLP, pages 1783.1786, 1998.

[21] T. Anastasakos, J. Mcdonough, R. Schwartz, and J. Makhoul. A compact model for speaker adaptive training. In Proc. ICSLP, pages 1137–1140, 1996.

[22] Yu D, Yao K, Su H, et al. KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition[C]Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013: 7893-7897.

[23] Liao H. Speaker adaptation of context dependent deep neural networks[C]Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013: 7947-7951.

[24] D. Albesano, R. Gemello, P. Laface, F. Mana, and S. Scanzio, "Adaptation of artificial neural networks avoiding catastrophic forgetting," in Proc. Int. Jnt. Conference on Neural Networks 2006, pp. 2863-2870, 2006.

[25] J. Stadermann and G. Rigoll, "Two-stage speaker adaptation of hybrid tied-posterior acoustic models," in Proc. ICASSP'05, vol. I, pp. 997-1000, 2005.

[26] G. Saon, H. Soltau, D. Nahamoo, et al. Speaker adaptation of neural network acoustic models using i-vectors. ASRU, 2013.

[27] Michael Seltzer, Dong Yu, and Yongqiang Wang. An Investigation Of Deep Neural Networks For Noise Robust Speech Recognition. ICASSP, 2013

[28] J. Li, J. T. Huang, Y. Gong. Factorized adaptation for deep neural network. ICASSP, 2014.

[29] O. Abdel-Hamid, H. Jiang. Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code. ICASSP, 2013.

[30] Quoc Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. ICML 2015.

[31] Tan, Tian, et al. "Speaker-aware training of LSTM-RNNS for acoustic modelling." 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.

[32] Tong, Sibo, Hao Gu, and Kai Yu. "A comparative study of robustness of deep learning approaches for VAD." 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.

[33] Vishwa Gupta, Patrick Kenny, Pierre Ouellet, Themos Stafylakis. I-vector-based speaker adaptation of deep neural networks for french broadcast audio transcription. ICASSP, 2014

[34] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 4, pp. 1435-1447, 2007

[35] P. J. Moreno, "Speech recognition in noisy environments," Ph.D. thesis, Carnegie Mellon University, 1996.

[36] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions," Computer, Speech and Language, vol. 23, no. 3, pp. 389–405, 2009.

[37] Chen, Xie, et al. "Recurrent neural network language model adaptation for multi-genre broadcast speech recognition." Proceedings of InterSpeech. 2015.

[38] T. Nagamine et al. "Adaptation of neural networks constrained by prior statistics of node co-activations.", INTERSPEECH, 2016.

[39] Yu, Nanxin Chen Yanmin Qian Kai. "Multi-task learning for text-dependent speaker verification." (2015).

[40] Qian, Yanmin, et al. "Multi-task joint-learning of deep neural networks for robust speech recognition." 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2015.

[41] Qian, Yanmin, et al. "Integrated adaptation with multi-factor joint-learning for far-field speech recognition." 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.

[42] K. Yao, D. Yu, F. Seide, et al. Adaptation of context-dependent deep neural networks for automatic speech recognition. IEEE SLT. 2012: 366-369.

[43] B. Li and K. C. Sim, "Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems," in Proc. InterSpeech, 2010

[44] Xue, Jian, et al. "Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network." 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014.

[45] R. Gemello, F. Mana, S. Scanzio, et al. Linear hidden transformations for adaptation of hybrid ANN/HMM models. Speech Communication, 2007, 49(10): 827-835.

[46] Seide, Frank, et al. "Feature engineering in context-dependent deep neural networks for conversational speech transcription." Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on. IEEE, 2011.

[47] J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals, and T. Robinson, "Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system," EuroSpeech, 1995.

[48] T. Ochiai, S. Matsuda, X. Lu, et al. Speaker Adaptive Training using Deep Neural Networks. ICASSP, 2014.

[49] Tan, Tian, Yanmin Qian, and Kai Yu. "Cluster adaptive training for deep neural network based acoustic model." IEEE/ACM Transactions on Audio, Speech, and Language Processing 24.3 (2016): 459-468.

[50] Karel Vesely, et. Al. Sequence Summarizing Neural Network for Speaker Adaptation. ICASSP 2016.

[51] He, Tianxing, et al. "Recurrent neural network language model with structured word embeddings for speech recognition." 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015.

[52] Swietojanski, Pawel, and Steve Renais. "SAT-LHUC: Speaker adaptive training for learning hidden unit contributions." 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.