ICASSP2017 paper review

Zhehuai Chen
chenzhehuai@foxmail.com

总体感觉（个人）

- 鲁棒、增强、多麦：
  - PIT等；DL/RL+**SP**；联合优化；相位信息
- 自适应：
  - 参数量；在线；非监督
- AM：
  - recurrent；residual；memory
- LM：
  - 更强的结构；对话语义
- KWS、解码：
  - end-to-end；robust

ICASSP2017 paper review （Robust & enhancement）

Zhehuai Chen
chenzhehuai@foxmail.com

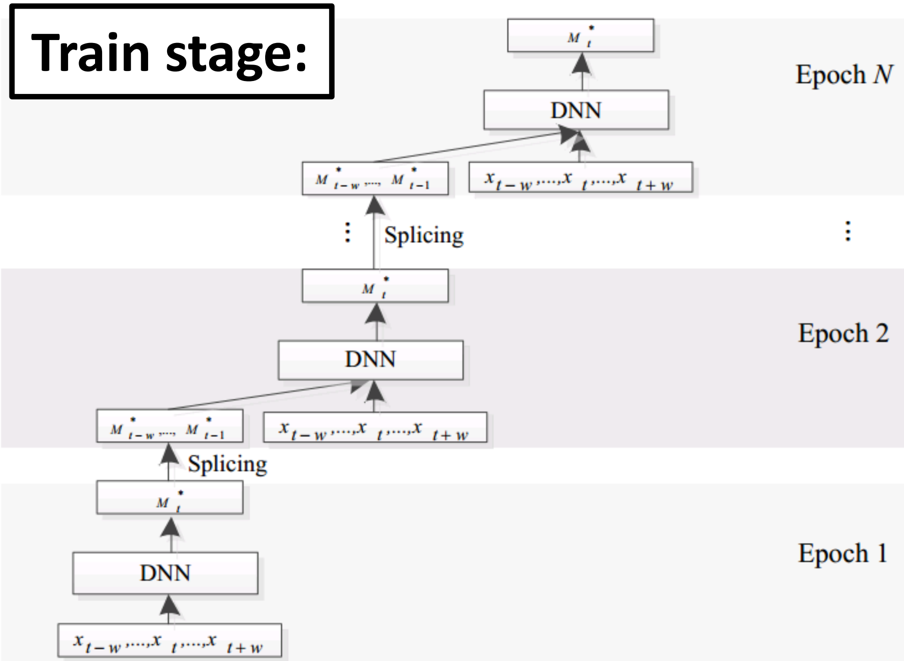
RECURRENT DEEP STACKING NETWORKS FOR SUPERVISED SPEECH SEPARATION

Zhong-Qiu Wang[♣] *and DeLiang Wang*[♣, ♥]

[♣]Department of Computer Science and Engineering, The Ohio State University, USA
[♥]Center for Cognitive and Brain Sciences, The Ohio State University, USA

- Input $< M^*_{t-w}, \ldots, M^*_{t-1}, x_{t-w}, \ldots, x_t, \ldots, x_{t+w} >$ → **explicit context**
- Drawback of recurrent NN → **implicit context**
  - BPTT makes NN "deeper" → more data
  - Gradient vanishing
  - Frame by frame → need shuffle/slower

Train stage:



test stage: Recurrent the output

Result: CHIME 2 T-F mask method even slightly better than LSTM

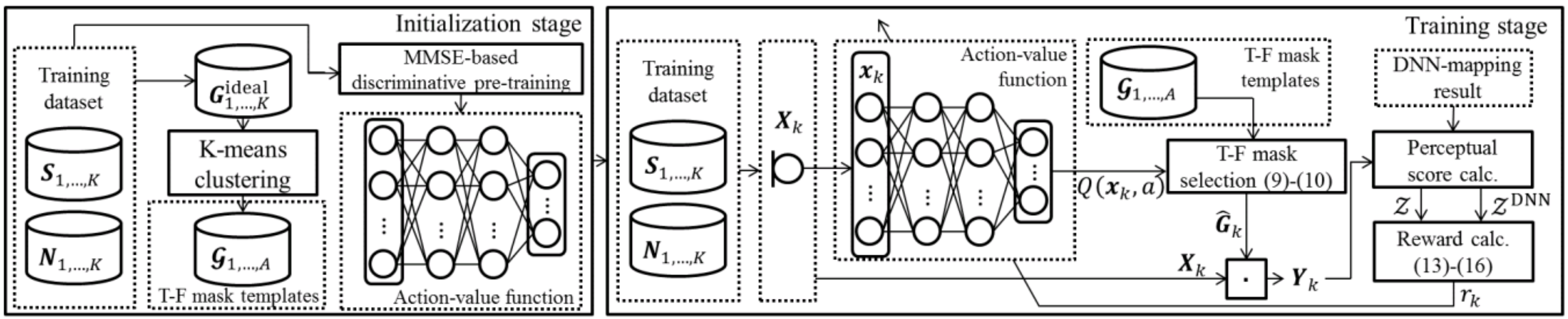
Comment?
if in ASR?

DNN-BASED SOURCE ENHANCEMENT SELF-OPTIMIZED BY REINFORCEMENT LEARNING USING SOUND QUALITY MEASUREMENTS

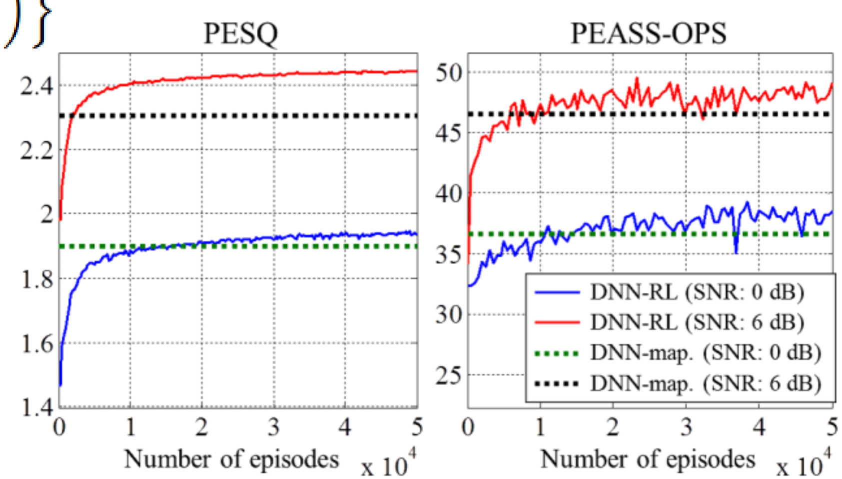
Yuma Koizumi[1,2], *Kenta Niwa*[1], *Yusuke Hioka*[3], *Kazunori Kobayashi*[1] , *and Yoichi Haneda*[2]

[1]: NTT Media Intelligence Laboratories, Tokyo, Japan
[2]: The University of Electro-Communications, Tokyo, Japan
[3]: Department of Mechanical Engineering, University of Auckland, Auckland, New Zealand

- Speech Enhancement ≠ better human perception → reward



- Action: T-F mask template
- Value: PESQ = $\mathcal{R} = \tanh\left\{\alpha\left(\mathcal{Z} - \mathcal{Z}^{\text{DNN}}\right)\right\}$ noise-reduce + perception
- Baseline: DNN-mapping method → who can win the game?
- PESQ need n-frames → designment as a time varying reward
- Q-learning



总结 Robust (不含SP)

session	paper
Deep Learning for Source Separation and Enhancement II	permutation invariant training of deep models for speaker-independent multi-talker speech separation
Deep Learning for Source Separation and Enhancement II	deep attractor network for single-microphone speaker separation
Deep Learning for Source Separation and Enhancement I	dnn-based source enhancement self-optimized by reinforcement learning using sound quality measurements
Deep Learning for Source Separation and Enhancement I	recurrent deep stacking networks for supervised speech separation
Robust Speech Recognition	a network of deep neural networks for distant speech recognition
Deep Learning for Source Separation and Enhancement II	deep mixture density network for statistical model-based feature enhancement
Acoustic Modeling I	student-teacher network learning with enhanced features
Deep Learning for Source Separation and Enhancement II	a speech enhancement algorithm by iterating single- and multi-microphone processing and its application to robust asr
Noise Modelling, Signal Enhancement and Equalization	probabilistic spatial dictionary based online adaptive beamforming for meeting recognition in noisy and reverberant environments
Topics in Speech Recognition	beamnet: end-to-end training of a beamformer-supported multi-channel asr system

ICASSP2017 paper review （Adaptation）

Zhehuai Chen
chenzhehuai@foxmail.com

UNSUPERVISED SPEAKER ADAPTATION OF BATCH NORMALIZED ACOUSTIC MODELS FOR ROBUST ASR

Zhong-Qiu Wang[♣] *and DeLiang Wang*[♣,♥]

[♣]Department of Computer Science and Engineering, The Ohio State University, USA
[♥]Center for Cognitive and Brain Sciences, The Ohio State University, USA

$$h^{(m)} = \delta\left(\gamma^{(m)} \frac{W^{(m)} h^{(m-1)} - \mu^{(m)}}{\sigma^{(m)}} + \beta^{(m)}\right)$$

$$\hat{x}_{t,f} = w_f \frac{x_{t,f} - \mu_f}{\sigma_f} + b_f$$

$$\hat{h}^{(m)} = \delta\left(\gamma^{(m)} \frac{W^{(m)} \hat{h}^{(m-1)} - \mu_{train}^{(m)}}{\sigma_{train}^{(m)}} + \beta^{(m)}\right)$$

- Batch norm
- Linear input network
  - Only change input as CMVN
  - How to get distribution in each layer?
- Proposed method
  - Get distribution from batch norm
  - Only do scaling & shifting
- LHUC
  - After the non-linear (before is better?)

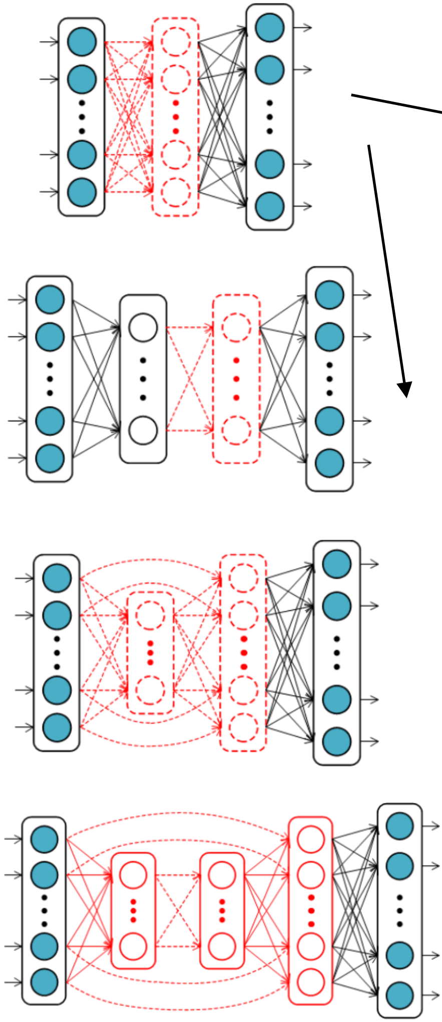
Table 1. *ASR performance (%WER) using first-pass decoding results of a tri-gram language model for adaptation.*

Approaches	LMs for decoding	Dev. set			Test set	
		SIMU	REAL	AVG	SIMU	REAL
Baseline acoustic model	Tri-gram	7.22	6.87	7.05	7.86	10.40
Batch normalized acoustic model	Tri-gram	6.85	6.47	6.66	7.17	9.51
LIN adaptation	Tri-gram	5.60	5.79	5.69	6.37	8.34
Scaling and shifting factors adaptation (proposed)	Tri-gram	4.98	**4.92**	4.95	**5.05**	**7.24**
Scaling and shifting factors adaptation (proposed) + LIN adaptation	Tri-gram	**4.93**	4.96	**4.94**	5.10	7.28
LHUC [19]	Tri-gram	5.18	5.36	5.27	5.58	7.78

EXTENDED LOW-RANK PLUS DIAGONAL ADAPTATION FOR DEEP AND RECURRENT NEURAL NETWORKS

Yong Zhao, Jinyu Li, Kshitiz Kumar, and Yifan Gong

Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA

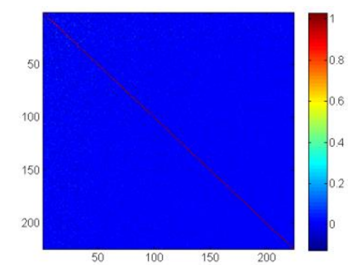


mn

because the adapted model should not deviate too far from the SI model given the limited number of adaptation data.

$$W_{s,m\times n} = U_{m\times k}\underline{S_{s,k\times k}}V_{k\times n}$$

BUT:



(a) An adaptation matrix

k^2

$$W_{s,k\times k} \approx D_{s,k\times k} + P_{s,k\times c}Q_{s,c\times k}.$$

k(2c+1)

c=0 → LHUC

$$W_{s,k\times k} \approx D_{s,k\times k} + P_{k\times c}T_{s,c\times c}Q_{c\times k}$$

c^2+k (k>>c)

PQ trained from small subset of training data
Without DPTQ to do unsupervised adaptation
→ Get SD parameter DT

总结Adaptation

session	paper
Acoustic Modeling and Adaptation	extended low-rank plus diagonal adaptation for deep and recurrent neural networks
Robust Speech Recognition	unsupervised speaker adaptation of batch normalized acoustic models for robust asr
Acoustic Modeling and Adaptation	joint optimisation of tandem systems using gaussian mixture density neural network discriminative sequence training
Acoustic Modeling II	unsupervised adaptation for deep neural networks using alternating direction method of multipliers
Acoustic Modeling II	cumulative moving averaged bottleneck speaker vectors for online speaker adaptation of cnn-based acoustic models
Acoustic Modeling II	personalized acoustic modeling by weakly supervised multi-task deep learning using acoustic tokens discovered from unlabeled data

ICASSP2017 paper review （KWS & Search）

Zhehuai Chen
chenzhehuai@foxmail.com

END-TO-END ASR-FREE KEYWORD SEARCH FROM SPEECH

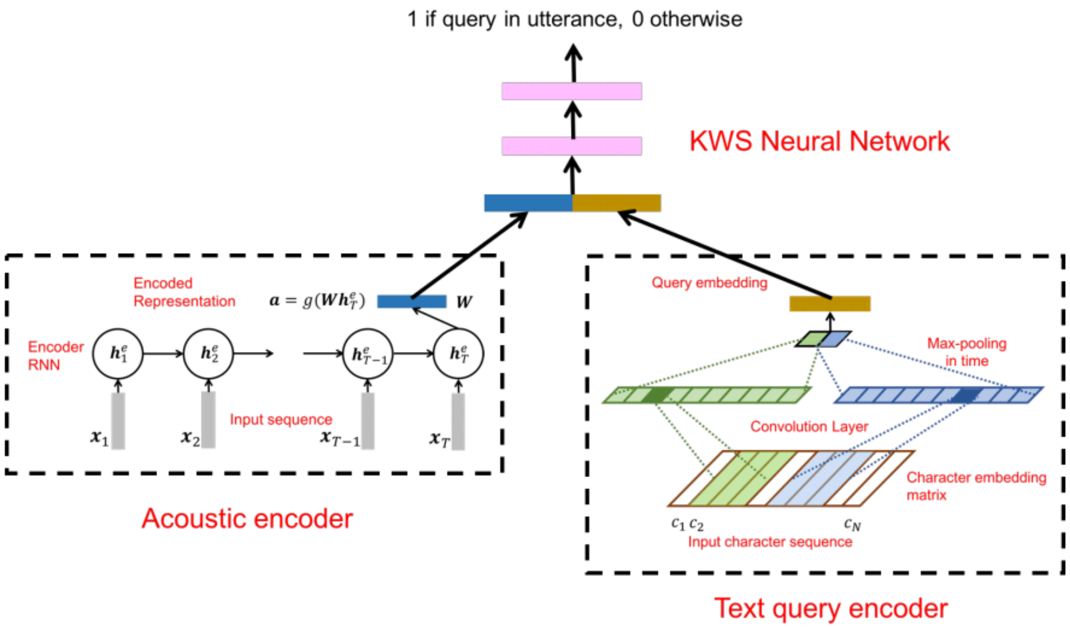
Kartik Audhkhasi, Andrew Rosenberg, Abhinav Sethy, Bhuvana Ramabhadran, Brian Kingsbury

IBM Watson, IBM T. J. Watson Research Center, Yorktown Heights, New York

END-TO-END SPEECH RECOGNITION AND KEYWORD SEARCH ON LOW-RESOURCE LANGUAGES

Andrew Rosenberg, Kartik Audhkhasi, Abhinav Sethy, Bhuvana Ramabhadran, Michael Picheny

IBM TJ Watson Research Center
Yorktown Heights, NY, USA



#1 difficult to derive a reliable representation for short queries due to the lack of context
#2 If without text encoder

Query Type →	IV	OOV
DNN-HMM (2gm word LM)	76.7	50.0 (chance)
DNN-HMM (4gm grapheme LM)	70.7	55.5
E2E ASR-free	55.6	57.7

End2End KWS (Con.)

- LVCSR framework (lattice method)

Language	ID	HMM-DNN	CTC	Attn
Pashto	104	52.7	52.8	55.5
Guarani	305	50.5	51.7	53.8

ID	HMM-DNN		Hyb-CTC	
	WER	MTWV	WER	MTWV
104	52.7	0.3853	51.0	0.3447
305	50.5	0.5345	47.7	0.5171

#1 1-best is good

#2 HMM > CTC feature > CTC > ETE
the ”peakiness” may not only impact posteriors, but aspects of the encoded features as well

#3 entropy: HMM > CTC > ETE

#4 final result

ID	ML24+RWTH		CTC+RWTH		Attn+RWTH	
	WER	MTWV	WER	MTWV	WER	MTWV
104	47.9	0.4088	49.3	0.3775	50.5	0.3528

AN LSTM-CTC BASED VERIFICATION SYSTEM FOR PROXY-WORD BASED OOV KEYWORD SEARCH

Zhiqiang Lv, Jian Kang, Wei-Qiang Zhang, Jia Liu

Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

- Proxy-word:
  - Watermelon (OOV) -> water merry (proxy)
  - Define by phone confusion
- If $P(\text{water merry}|X) < \text{thres}$:
  - Replace “water merry” by “watermelon”
- How to detect proxy-word -> $P(W|X)$
- $P(W|X)$ is a CTC trained in LVCSR (without OOV)
- Get better result compared with $P(X|W)P(W)$
  - Hasn’t compared with $P(X|W)P(W)/P(X)$

总结KWS & Search

session	paper
Keyword Search	an lstm-ctc based verification system for proxy-word based oov keyword search
End to End Speech Processing	end-to-end asr-free keyword search from speech
Acoustic Modeling I	end-to-end speech recognition and keyword search on low-resource languages
Keyword Search	distance metric learning for posteriorgram based keyword search
Spoken Term Detection	morph-to-word transduction for accurate and efficient automatic speech recognition and keyword search

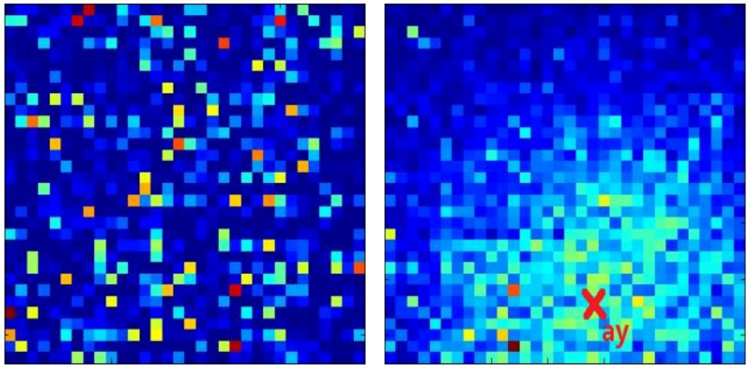
ICASSP2017 paper review （AM）

Zhehuai Chen
chenzhehuai@foxmail.com

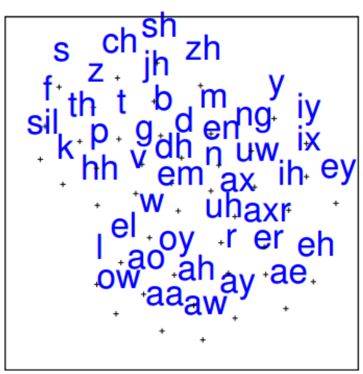
STIMULATED TRAINING FOR AUTOMATIC SPEECH RECOGNITION AND KEYWORD SEARCH IN LIMITED RESOURCE CONDITIONS

A. Ragni, C. Wu, M. J. F. Gales, J. Vasilakes, K. M. Knill

Department of Engineering, University of Cambridge
Trumpington Street, Cambridge CB2 1PZ, UK



(a) Unstimulated Activations (b) Stimulated Activations



Language	Stimulated	TER (%)	MTWV		
			IV	OOV	Total
Pashto	✗	44.6	0.4720	0.3986	0.4644
	✓	44.4	0.4752	0.4032	0.4672

Pashto	32 × 32	44.4	0.4752	0.4032	0.4672
	45 × 45	43.8	0.4828	0.4083	0.4750

- Pool interpretability
  - Feature space transformation
  - Encourage NN group together
- Make activation corresponding to:
  - Predefined phone similarity **prior** (data driven by t-SNE)

$$\mathcal{F}(\boldsymbol{\lambda}) = \mathcal{L}(\boldsymbol{\lambda}) + \alpha\mathcal{R}(\boldsymbol{\lambda})$$

- Improve interpretability & discrimination
- Experiment: low-resource KWS
  - With generalization problem
  - Prior includes: position, tone, stress, diacritic, etc.
  - Improve lattice → better generalization

Residual Memory Networks: Feed-forward approach to learn long-term temporal dependencies

Murali Karthick Baskar, Martin Karafiát, Lukáš Burget, Karel Veselý, František Grézl and Jan "Honza" Černocký

Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Brno, Czech Republic

RECURRENT CONVOLUTIONAL NEURAL NETWORK FOR SPEECH PROCESSING

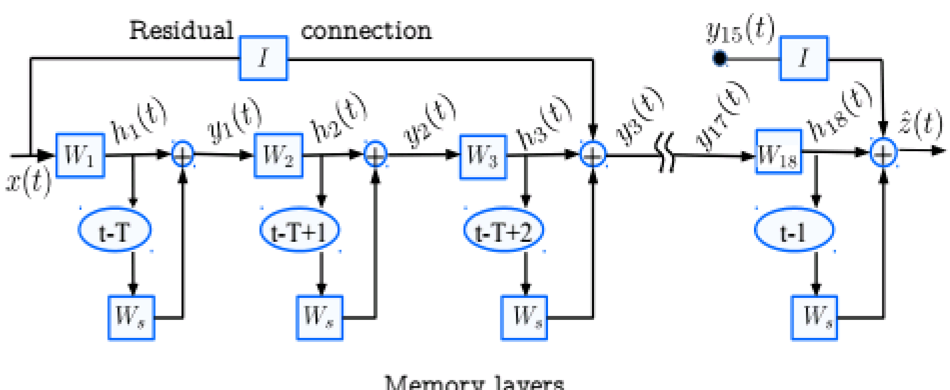
Yue Zhao, Xingyu Jin

Department of Electronic Engineering, TNList, Tsinghua University, Beijing 100084, China

Xiaolin Hu

Department of Computer Science and Technology, TNList, Tsinghua University, Beijing 100084, China

- TDNN, FSMN model longer context, but fail to model temporal order
  - TDNN < LSTM(RNN)



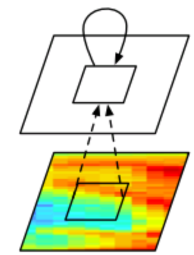
$$y_l(t) = \phi\left(x(t)W_l + h_l(t-m)W_s\right),\ l = 1, 2, ..L \quad (1)$$

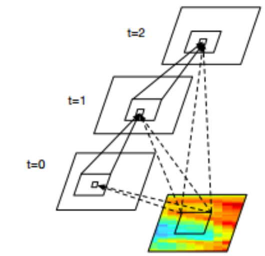
where $h_l(t) = x(t)W_l$ and ϕ is the relu activation output.

Novel NN (Con.)

$$\mathbf{h}(t) = \sigma(W_{xh}\mathbf{x}(t) + W_{hh}\mathbf{h}(t-1) + b_h)$$

$$\mathbf{h}^{(t)}(i,j) = \sigma\Big(\sum_{i'=-s}^{s}\sum_{j'=-s}^{s}\mathbf{w}_k^f(i',j')\mathbf{x}^{(t)}(i-i',i-j') + \sum_{i'=-s}^{s}\sum_{j'=-s}^{s}\mathbf{w}_k^r(i',j')\mathbf{h}^{(t-1)}(i-i',j-j') + b\Big)$$





TIMIT

RCL(2)+CL+3-layer MLP	**17.0%**	**18.0%**
DBN [20]	-	20.7%
CNN (limited weight sharing) [1]	-	20.5%
bottleneck CNN [27]	16.1%	18.6%
3-layer LSTM + HMM [30][4]	17.7%	18.8%
3-layer LSTM + pre-trained transducers [10]	-	17.7%
Attention model [6]	15.8%	17.6%
time- and frequency- domain convolution [28]	14.2%	17.6%
time- and frequency- domain convolution (with dropout) [28]	**13.9%**	**16.7%**

Table 5: Comparison of RMN with existing methods in literature trained using 300 hours of Switchboard corpus and tested with Hub5-00 eval set. In this table 3g is trigram, 4g is meant as 4-gram, bn-fMLLR is bottleneck features with fMLLR and ivec represents 100 dimensional ivectors built using section 3.2.

% WER	Model Type	SWB (% CE WER)		
		3g	4g	4g+ivec
Proposed Models	RMN	13.0	12.0	10.9
	BRMN	11.8	10.8	9.9
State-of-the-art results in literature	TDNN [19]			12.5
	Unfolded RNN + fMLLR [18]			12.7
	LSTM + bn-fMLLR [20]			10.8
	LSTM [19]			11.6
	BLSTM [19]			10.3

In CNTK

	train	decode
RCNN	2012 samples per second	1.721 utterances per second
LSTM	275 samples per second	0.944 utterances per second

总结 AM

topic	paper
Acoustic Modeling I	recurrent convolutional neural network for speech processing
Neural Network Trends in Speech Recognition	residual memory networks: feed-forward approach to learn long-term temporal dependencies
Neural Network Trends in Speech Recognition	stimulated training for automatic speech recognition and keyword search in limited resource conditions
Neural Network Trends in Speech Recognition	advances in all-neural speech recognition
Acoustic Modeling I	the microsoft 2016 conversational speech recognition system

ICASSP2017 paper review （Others）

Zhehuai Chen
chenzhehuai@foxmail.com

总结 LM

topic	paper
Language Modeling	recurrent neural network based language modeling with controllable external memory
Language Modeling	character-level language modeling with hierarchical recurrent neural networks
Language Modeling	learning concepts through conversations in spoken dialogue systems
Language Modeling	a neural network approach for mixing language models
Language Modeling	dialog context language modeling with recurrent neural networks

总结 engineering

topic	paper
Deep Learning for Source Separation and Enhancement II	impact of low-precision deep regression networks on single-channel source separation
Deep Learning for Source Separation and Enhancement II	improving music source separation based on deep neural networks through data augmentation and network blending
Deep Learning I	selecting optimal layer reduction factors for model reduction of deep neural networks
Acoustic Modeling I	semi-supervised ensemble dnn acoustic model training
Noise Robust Speech Recognition	a study on data augmentation of reverberant speech for robust speech recognition
Robust Speech Recognition	discriminative importance weighting of augmented training data for acoustic model training
Topics in Speech Recognition	improving latency-controlled blstm acoustic models for online speech recognition
Topics in Speech Recognition	predicting error rates for unknown data in automatic speech recognition
Topics in Speech Recognition	speeding up softmax computations in dnn-based large vocabulary speech recognition by senone weight vector selection
Keyword Search	trainable frontend for robust and far-field keyword spotting