IS2016 paper review

Zhehuai Chen
chenzhehuai@foxmail.com

总体感觉（个人）

- AM: CNN, FSMN, highway…
- Robust, far-field: 框架没有改变，结构有研究创新，但不一定可商用
- LM: 同上
- Decoder: 无
- 合成, Speaker, 自适应, SLU: 没仔细看

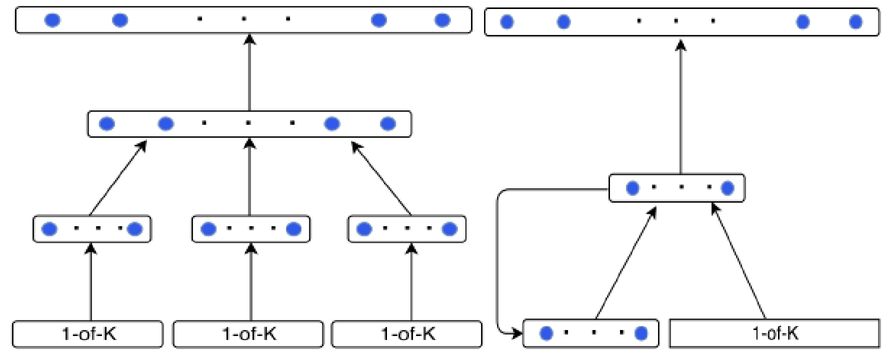
IS2016 paper review （LM & AM）

Zhehuai Chen
chenzhehuai@foxmail.com

Sequential Recurrent Neural Networks for Language Modeling

AUTHORS:
Youssef Oualil, Clayton Greenberg, Mittul Singh, Dietrich Klakow, *Universität des Saarlandes, Germany*

Fnnlm→strong short range
Fsmn→word-dependent weight
Rnnlm→long context
This work → combine all

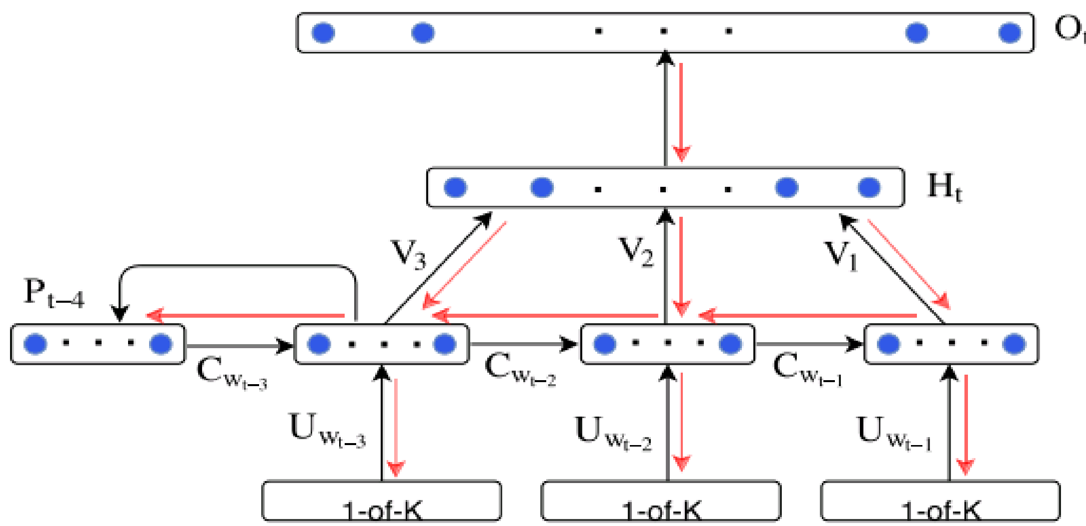


(a) FNN (b) RNN

$$P_{t-i} = f_s(X_{t-i} \cdot U + C_{w_{t-i}} \odot P_{t-i-1}),\ i = N-1, \cdots, 1$$

$$p(w_1^T) \approx \prod_{t=1}^{T} p(w_t | w_{t-N+1}^{t-1}, h_{t-N+1})$$

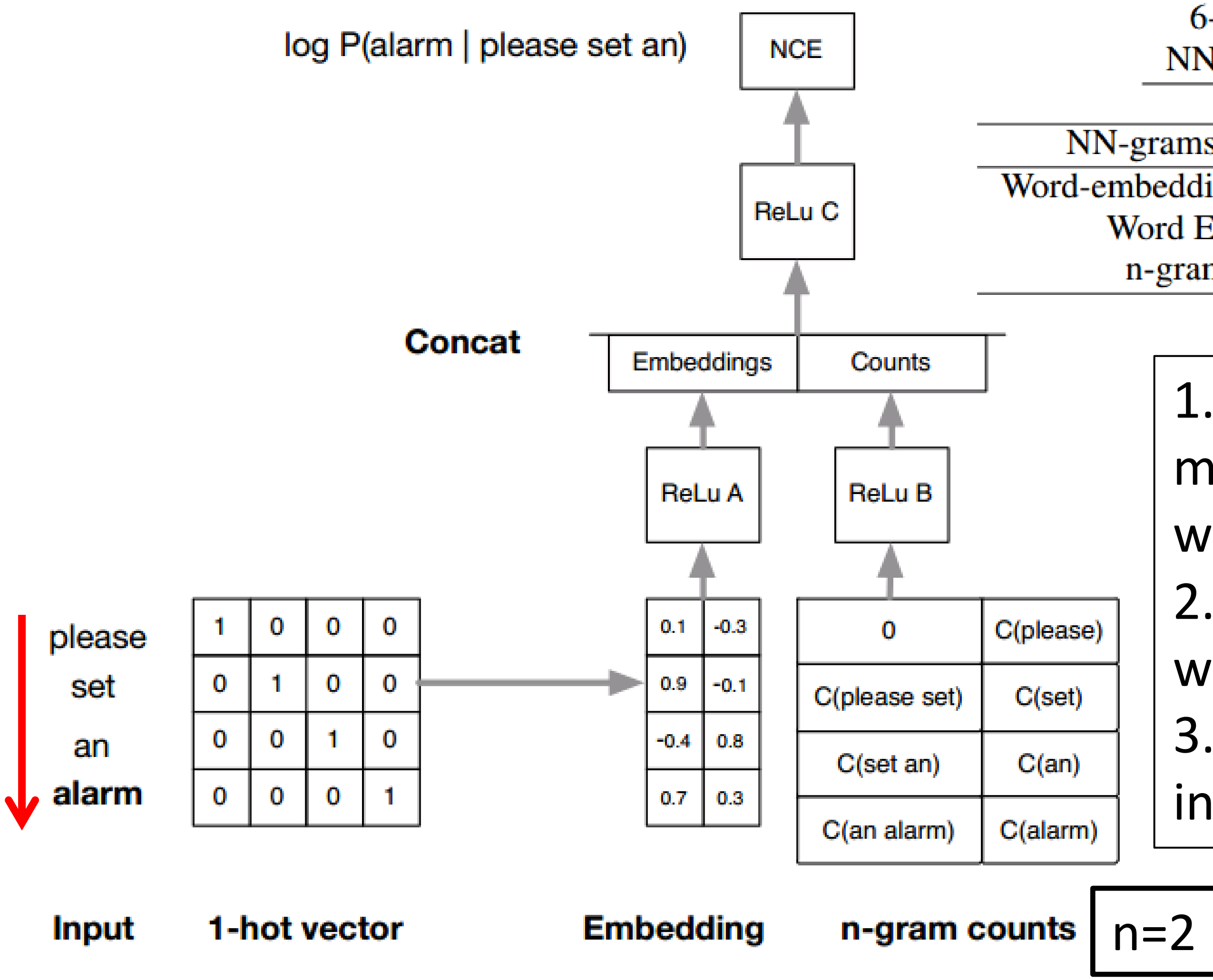
FNN	176	129	114
FOFE	116	108	109
WI-SRNN*	114	108	107
WI-SRNN	109	105	104
WD-SRNN	108	103	104
RNN	123		
Deep RNN	107.5		
LSTM	114		



e.g.: $C = [0, \cdots, 0]$ → FNNLM
N=2 → RNNLM

NN-Grams: Unifying Neural Network and n-Gram Language Models for Speech Recognition

AUTHORS:
Babak Damavandi, Shankar Kumar, Noam Shazeer, Antoine Bruguier, *Google, USA*



	VS	DTN
6-gram	14.9	8.8
NN-grams	14.8	8.2

NN-grams components	VS	DTN
Word-embedding,n-gram counts	14.8	8.2
Word Embedding	15.3	8.8
n-gram Counts	14.9	8.5

1.n-gram counts were more important than word embeddings
2.n-gram → short snt.
word embed→long snt.
3.**Not** compare with interpolated result

Active and Semi-Supervised Learning in ASR: Benefits on the Acoustic and Language Models

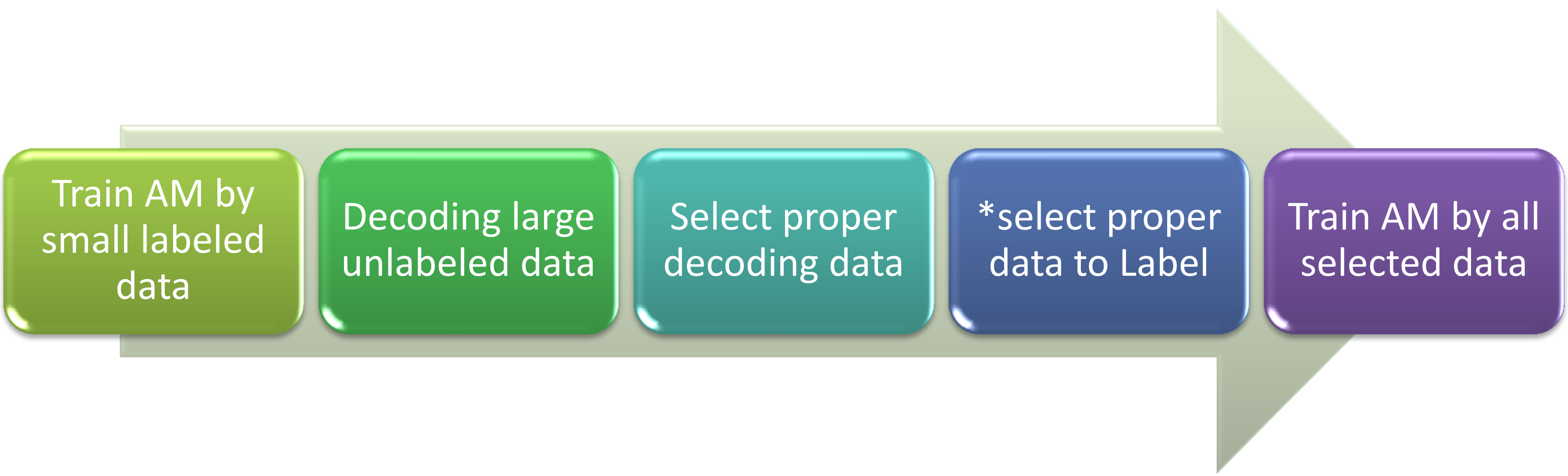
AUTHORS:
Thomas Drugman [1], Janne Pylkkönen [2], Reinhard Kneser [1]
[1] *Amazon.com, Germany;* [2] *Amazon.com, Finland*

Semi-Supervised Training in Deep Learning Acoustic Model

AUTHORS:
Yan Huang, Yongqiang Wang, Yifan Gong, *Microsoft, USA*

Investigation of Semi-Supervised Acoustic Model Training Based on the Committee of Heterogeneous Neural Networks

AUTHORS:
Naoyuki Kanda, Shoji Harada, Xugang Lu, Hisashi Kawai, *NICT, Japan*



- **Select proper data**
  - **Confidence**
    - **High→well train; low→bad labeling**
    - **Word level: Lattice posterior, avg acoustic score, ROVER (system combine)**
    - **Frame level: lattice arc posterior in frame, recalibration (system combine)**
  - **Committee**
    - **Vote from AMs of different archetecture**
- **Integrate data quality metric into training**
  - **Weighted error signal (frame level)**
  - **Weighted gate parameter in LSTM (frame level)**
  - **Importance sample (data value, quality, prior distribution)**
- **How to define well-trained data & label quality respectively?**

- **Some useful conclusion**
  - **Imperfect labeling is also useful**
  - **Data of high confidence is useless**
  - **LSTM is more sensitive to wrong labeling**
  - **Sequence training is more sensitive to wrong labeling (except LFMMI?)**

补充

topic	paper
NNLM summary	LSTM, GRU, Highway and a Bit of Attention: An Empirical Overview for Language Modeling in Speech Recognition
student-teacher	Sequence Student-Teacher Training of Deep Neural Networks
student-teacher	Robust Speech Recognition Using Generalized Distillation Framework
student-teacher	Model Compression Applied to Small-Footprint Keyword Spotting
student-teacher	Distilling Knowledge from Ensembles of Neural Networks for Speech Recognition

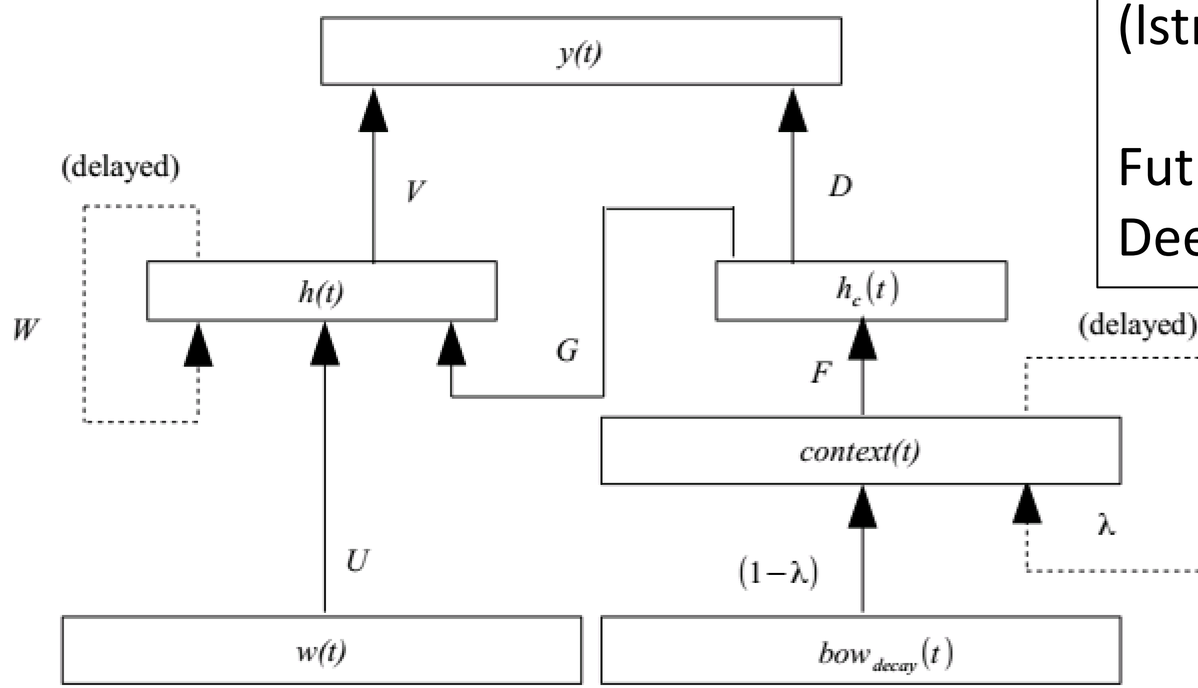
Recurrent Neural Network Language Model with Incremental Updated Context Information Generated Using Bag-of-Words Representation

AUTHORS:
Md. Akmal Haidar, Mikko Kurimo, *Aalto University, Finland*

$$context(t) = \lambda context(t-1) + (1-\lambda) bow_{decay}(t)$$

$$h(t) = f(Uw(t) + Wh(t-1) + Gh_c(t))$$

$$h_c(t) = f(Fcontext(t))$$



Direct concatenate → no use

This work:
increment+delay+nonlinear

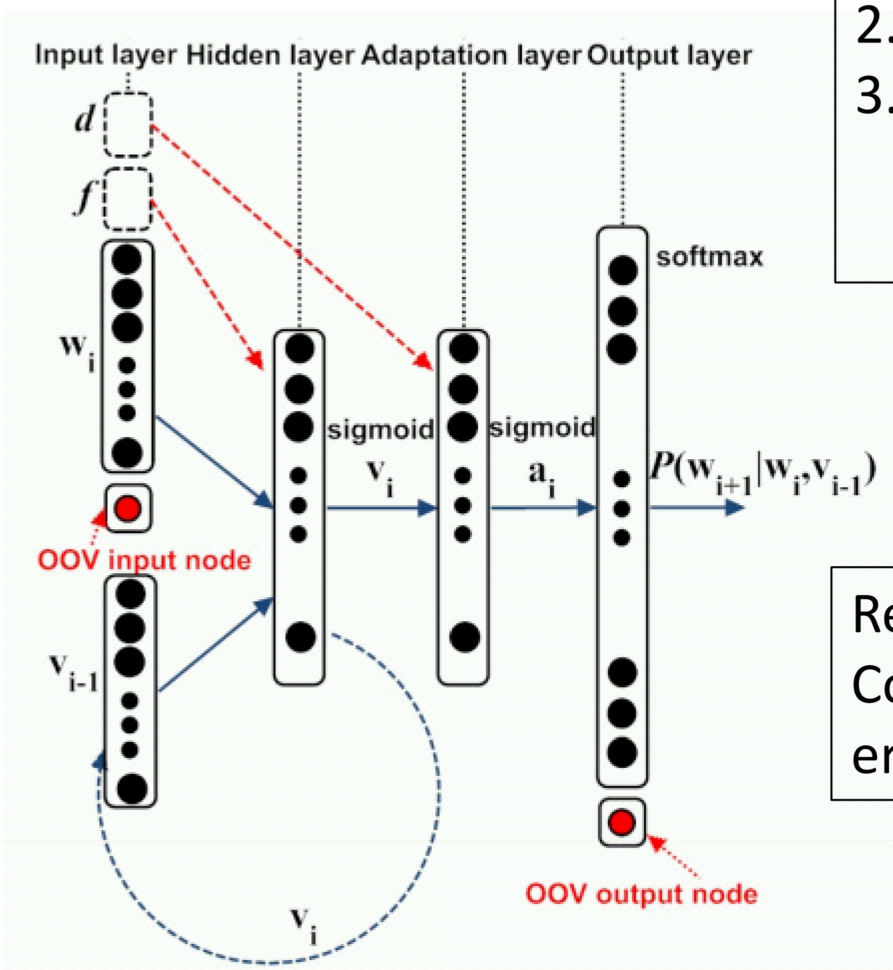
ppl: 126 → 115
(lstmlm:115)

Future work:
Deep context layer & lstmlm

Combining Feature and Model-Based Adaptation of RNNLMs for Multi-Genre Broadcast Speech Recognition

AUTHORS:
Salil Deena, Madina Hasan, Mortaza Doulaty, Oscar Saz, Thomas Hain, *University of Sheffield, UK*

1. 1-of-K encoding of domain
2. LDA feature
3. Linear hidden network (LHN) adaptation layer fine-tune by adaptation data



Result:
Consistent 10% perplexity and 2% word error rate improvements

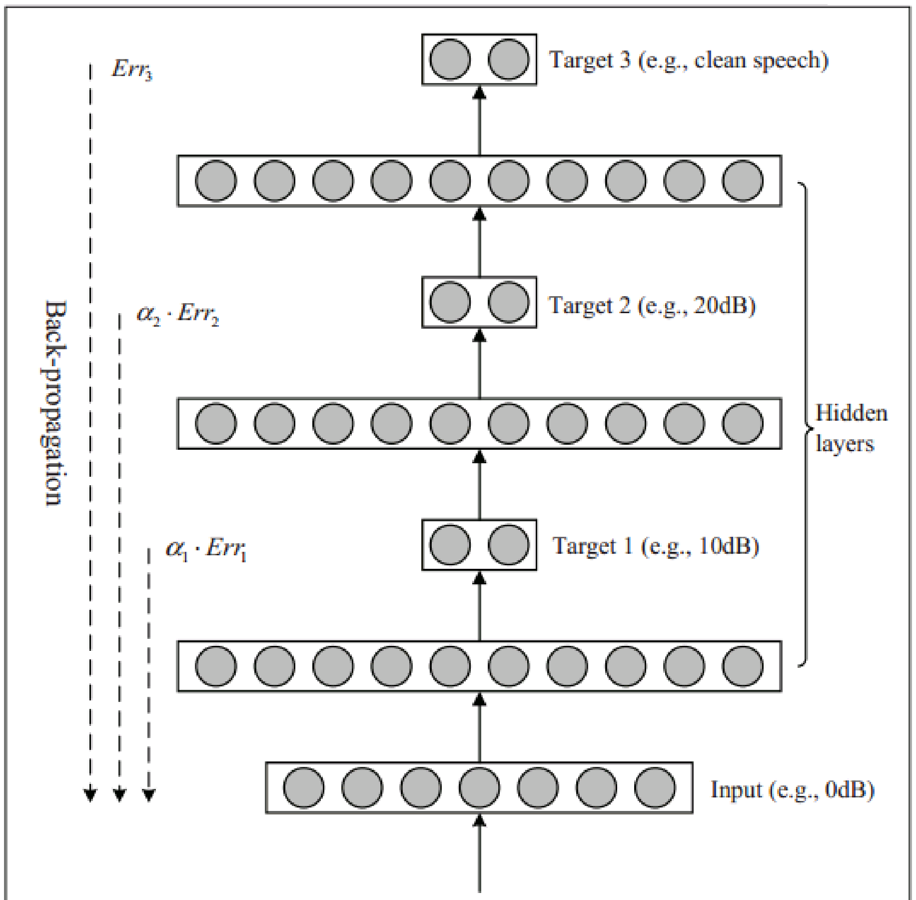
IS2016 paper review （robust ASR & far field）

Zhehuai Chen
chenzhehuai@foxmail.com

SNR-Based Progressive Learning of Deep Neural Network for Speech Enhancement

AUTHORS:
Tian Gao [1], Jun Du [1], Li-Rong Dai [1], Chin-Hui Lee [2]
[1] USTC, China; [2] Georgia Institute of Technology, USA

$$\epsilon = \frac{\partial(Err_3)}{\underset{1\leq\ell\leq L_3+1}{\partial(\boldsymbol{W}^\ell, \boldsymbol{b}^\ell)}} + \alpha_2 \frac{\partial(Err_2)}{\underset{1\leq\ell\leq L_2+1}{\partial(\boldsymbol{W}^\ell, \boldsymbol{b}^\ell)}} + \alpha_1 \frac{\partial(Err_1)}{\underset{1\leq\ell\leq L_1+1}{\partial(\boldsymbol{W}^\ell, \boldsymbol{b}^\ell)}} \tag{2}$$



Better quality + better **intelligibility**

Data Selection by Sequence Summarizing Neural Network in Mismatch Condition Training

AUTHORS:
Kateřina Žmolíková [1], Martin Karafiát [1], Karel Veselý [1], Marc Delcroix [2], Shinji Watanabe [3], Lukáš Burget [1], Jan Černocký [1]
[1]Brno University of Technology, Czech Republic; [2]NTT, Japan; [3]MERL, USA

selecting **subset of training data** with respect to **similarity** of acoustic conditions to test data

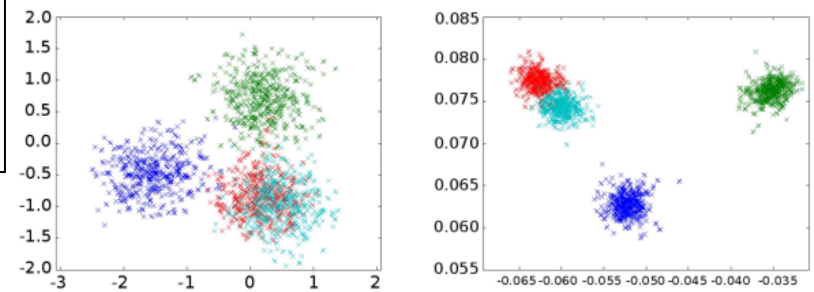
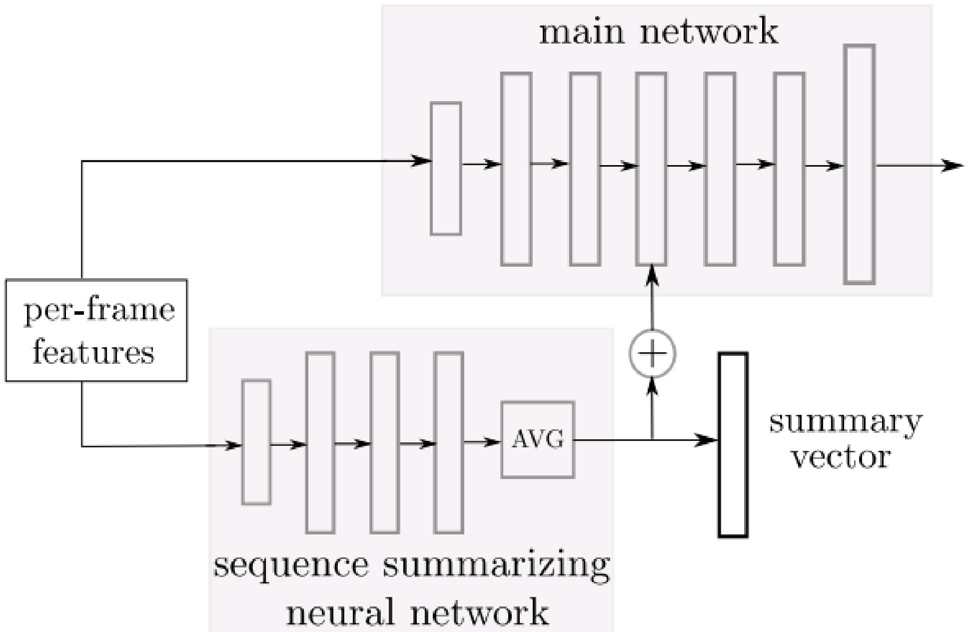


Figure 3: Plot of the first and second LDA basis on CHiME3 data for i-vectors (left) and summary-vectors (right).

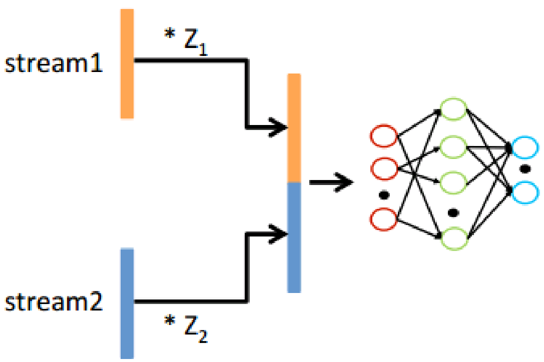


Dataset	Selection [%WER]		
	Random	i-vector	summary-vector
dev	25.8	25.61	**24.72**
eval	45.58	44.02	**43.23**

A Framework for Practical Multistream ASR

AUTHORS:
Sri Harish Mallidi, Hynek Hermansky, *Johns Hopkins University, USA*

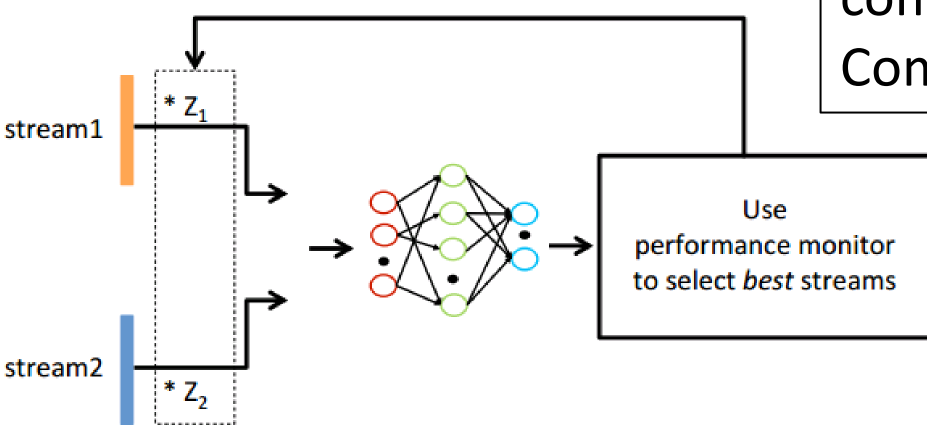
(a) Training stage

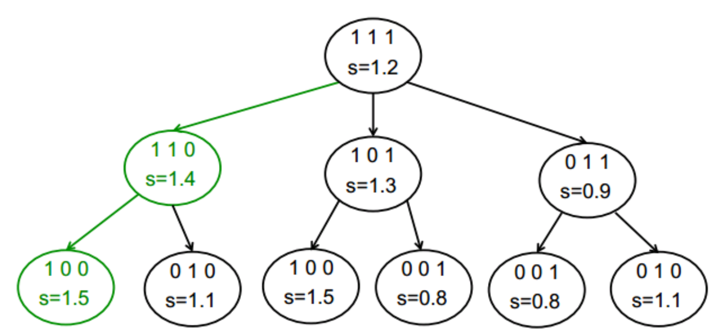


- Z_i is a binary switch
- Block Dropout features of a stream
- DNN randomly sees one of the following input patterns:
  - [stream1; **0**]
  - [**0**; stream2]
  - [stream1; stream2]

Compared to multistream: reduce complexity (**but** not compare CER)
Compared to dnn baseline: better CER

(b) Testing stage





补充

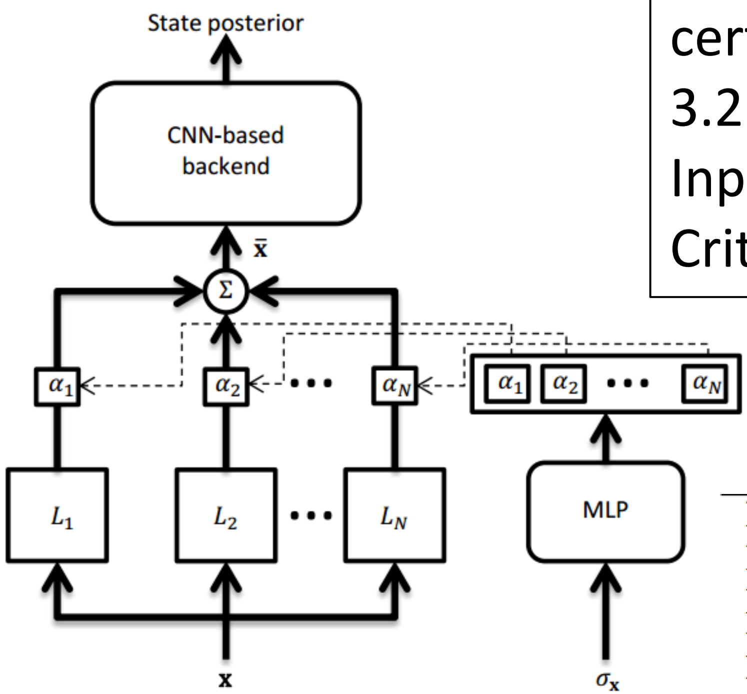
topic	paper
NMF	A DNN-HMM Approach to Non-Negative Matrix Factorization Based Speech Enhancement
enhance	Robust Example Search Using Bottleneck Features for Example-Based Speech Enhancement
enhance	Optimization of Speech Enhancement Front-End with Speech Recognition-Level Criterion
data augmentation	Realistic Multi-Microphone Data Simulation for Distant Speech Recognition
data augmentation	Synthesis of Device-Independent Noise Corpora for Realistic ASR Evaluation
data augmentation	Data Augmentation Using Multi-Input Multi-Output Source Separation for Deep Neural Network Based Acoustic Modeling
others	Adversarial Multi-Task Learning of Deep Neural Networks for Robust Speech Recognition
others	Reducing the Computational Complexity of Multimicrophone Acoustic Models with Integrated Feature Extraction
others	Far-Field ASR Without Parallel Data

Factorized Linear Input Network for Acoustic Model Adaptation in Noisy Conditions

AUTHORS:
Dung T. Tran, Marc Delroix, Atsunori Ogawa, Tomohiro Nakatani, *NTT, Japan*

1.Training CNN: noisy data
2.enhance: WPE+MVDR
3.adaptation:
3.1 LIN: bp of certain adaptation data of certain noise type
3.2 MLP of FLIN:
Input: enhanced data
Criteria: | noisy - enhanced | after FLIN

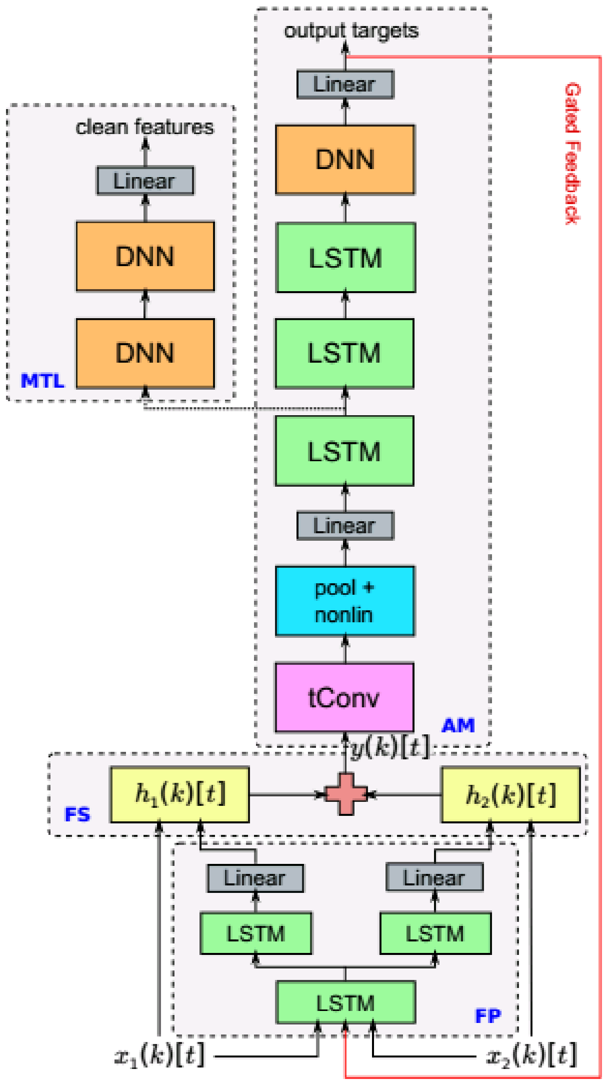
$$\hat{\mathbf{x}}_{n,t} = \mathbf{L}_n(\mathbf{x}_t) = \mathbf{W}_n \mathbf{x}_t + \mathbf{b}_n,$$



	BUS	CAF	PED	STR	Ave
Baseline	11.89	8.12	8.95	8.90	9.32
Retrain	10.12	6.57	7.54	7.53	7.94
LIN adaptation	10.57	7.36	8.18	7.90	8.50
FLIN (utt)	10.80	7.09	7.90	7.76	8.38
FLIN + retrain(utt)	9.71	6.69	7.73	7.25	7.82
FLIN + retrain(frame)	**9.37**	**6.40**	**7.33**	**7.25**	**7.58**

Neural Network Adaptive Beamforming for Robust Multichannel Speech Recognition

AUTHORS:
Bo Li, Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, Michiel Bacchiani, *Google, USA*



In the NAB model, we estimate the filter coefficients jointly with the AM parameters by directly minimizing a cross-entropy or sequence loss function.

An LSTM to predict N filter coefficients per channel.

Compared to fixed factor filters (Tara. ICASSP2016): less computation
Compared to single chan.: better WER

Model	WER (%)	
	CE	Seq.
unfactored [2]	21.7	17.5
factored [3]	20.4	17.1
NAB	20.5	17.2