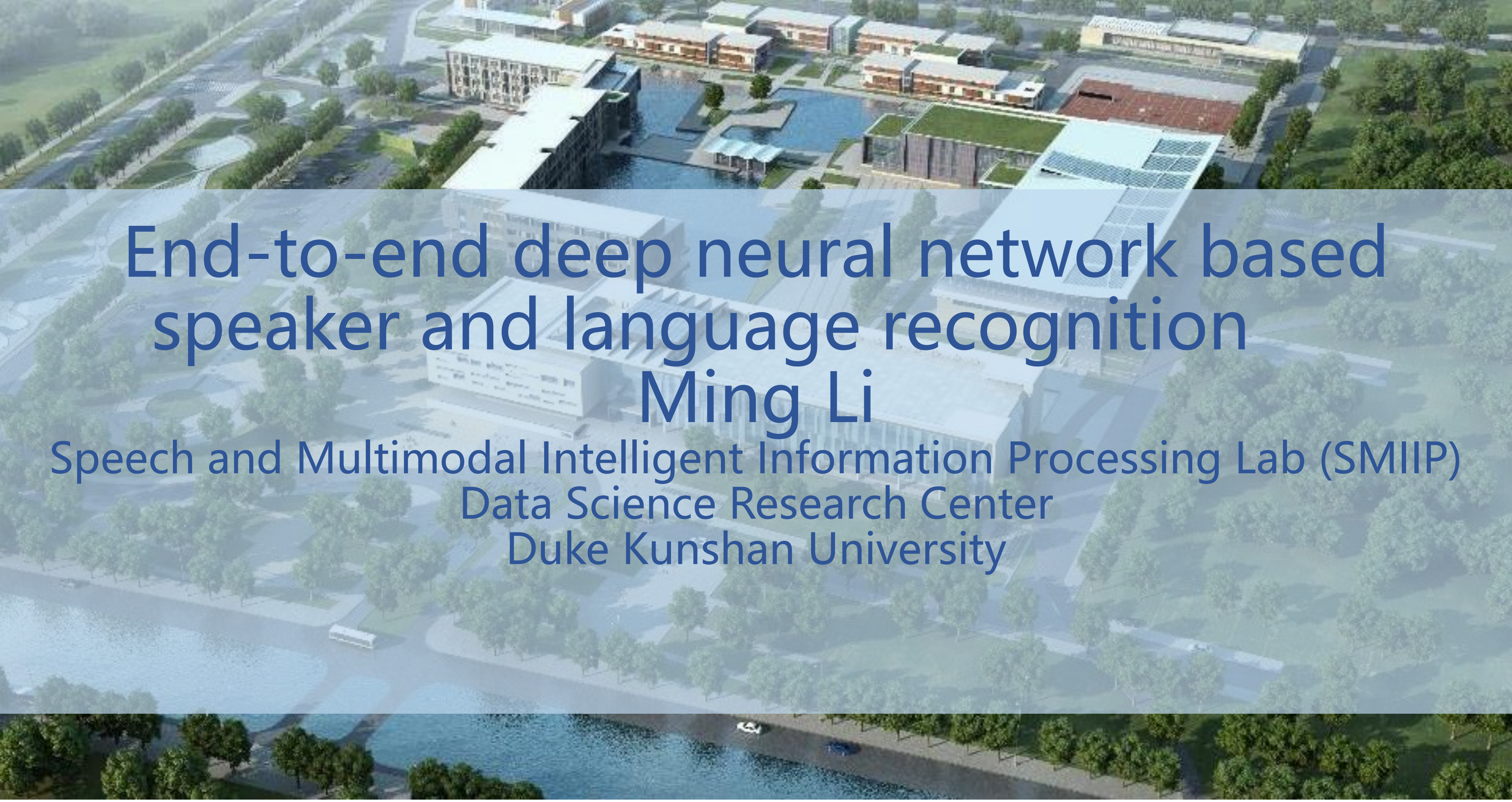
清华大学 | 讲义分享

语言识别技术研讨会

2019.3.25

olr.cslt.org



End-to-end deep neural network based speaker and language recognition

Ming Li
Speech and Multimodal Intelligent Information Processing Lab (SMIIP)
Data Science Research Center
Duke Kunshan University

March 24th 2019

https://scholars.duke.edu/person/MingLi

SMIIP

昆山杜克大学
DUKE KUNSHAN UNIVERSITY

Duke undergraduate experience distinct in the world of elite higher education.

Now she makes one step forward..

DUKE KUNSHAN UNIVERSITY provides a historic opportunity to create something truly innovative and world leading for 21st century – even better than the Duke curriculum, and Duke faculty took on this challenge.

- Sino-US Joint Venture University with independent legal status
- Duke-standard education and research
- Comprehensive and small

Duke
UNIVERSITY

+

+

昆山杜克大学
DUKE KUNSHAN
UNIVERSITY

Security & custom care

speaker, language,
gender, age, emotion,
channel, voicing,
psychological states, etc.

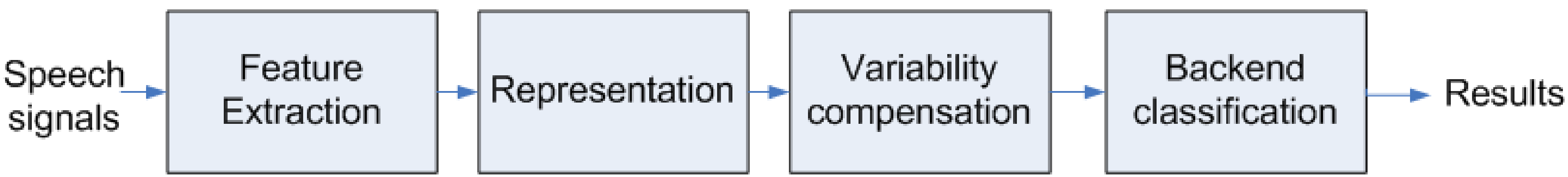
paralinguistic speech attribute recognition

Speech signal not only contains lexicon information, but also deliver various kinds of paralinguistic speech attribute information, such as speaker, language, gender, age, emotion, channel, voicing, psychological states, etc.

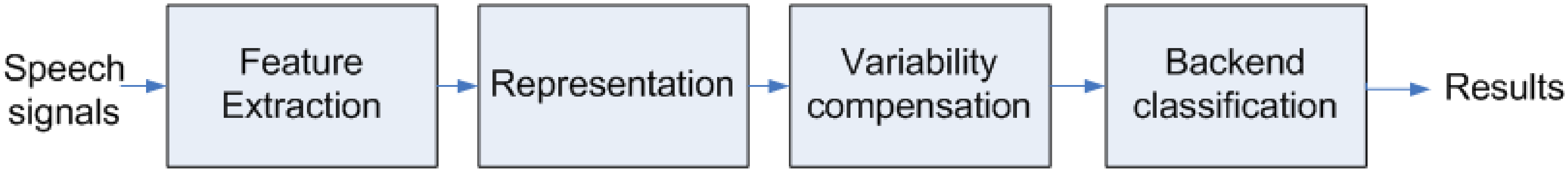
http://compare.openaudio.eu/

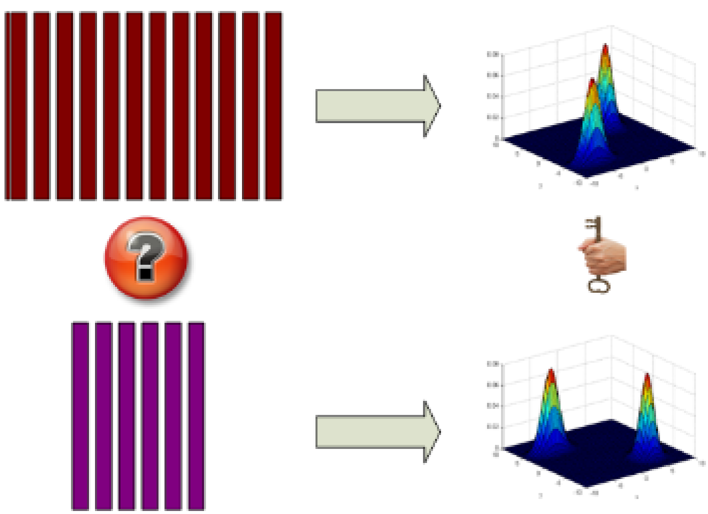
The core technique question behind it is utterance level supervised learning based on text independent speech signal with flexible duration

General framework



General framework





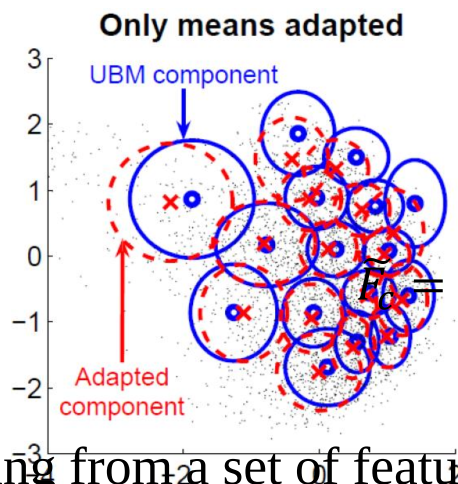
- time varying property

  -> short time frame level features

- generative model for data description -> features (supervectors) in model parameters' space for classification

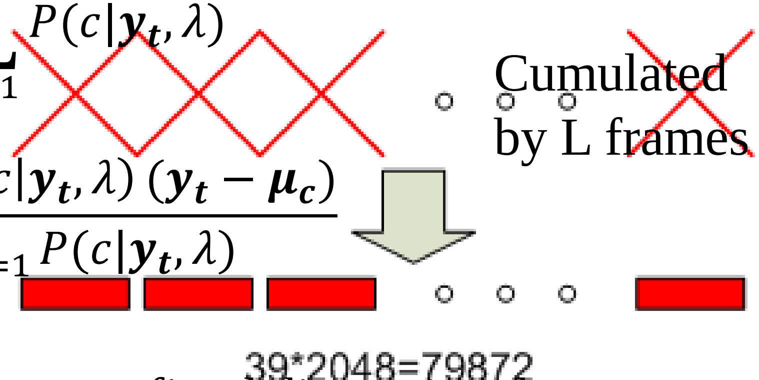
Generative model, adaptation, supervectors

- Gaussian Mixture Model (GMM) serves as the generative model
  - model adaptation from universal background model (UBM)
    - MAP adaptation, large dimensional GMM mean supervector
    - Maximum Likelihood Linear Regression (MLLR) adaptation
  - The statistics vector for a set of features on UBM
    - 0th order statistics vector N,centered 1st order statistics vector F



$$N_c = \sum_{t=1}^{L} P(c|\boldsymbol{y_t}, \lambda)$$

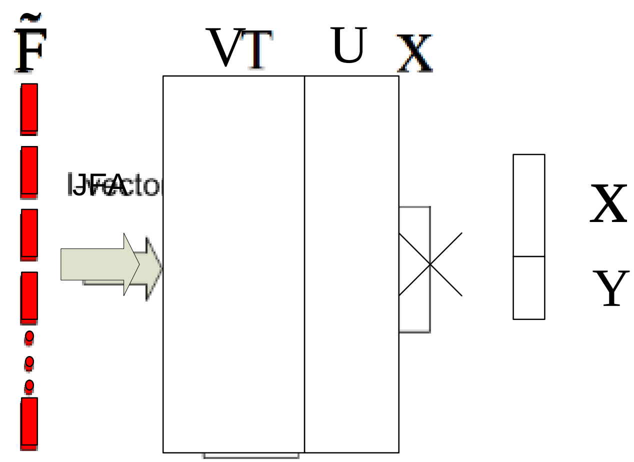
$$\boldsymbol{F_c} = \frac{\sum_{t=1}^{L} P(c|\boldsymbol{y_t}, \lambda)\,(\boldsymbol{y_t} - \boldsymbol{\mu_c})}{\sum_{t=1}^{L} P(c|\boldsymbol{y_t}, \lambda)}$$

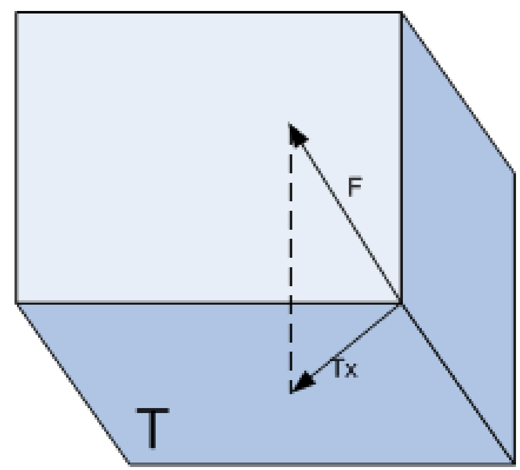


Mapping from a set of feature vectors to a fixed dimensional supervector

Factor analysis based dimension reduction

- Factor analysis on the concatenated 1st order statistics vector
  - Total variability i-vector, $\tilde{\boldsymbol{F}} \rightarrow \mathbf{Tx}$ (Dehak et.al, IEEE TASLP, 2011)
    T: factor loading matrix; **x**: i-vector
  - Joint factor analysis (JFA), $\tilde{\boldsymbol{F}} \rightarrow \mathbf{Vx+Uy}$ (Kenny et.al, IEEE TASLP, 2007)
    V: Eigenvoices, **U**: Eigenchannels, **x**: speaker factor, **y**: channel factor



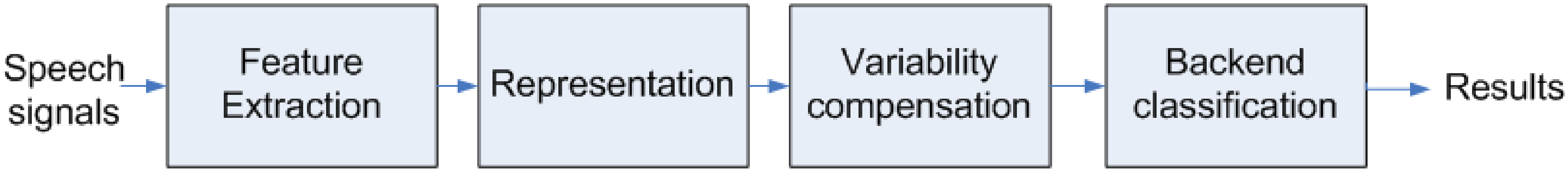


Variability compensation, modeling

- Variability compensation
  - Linear discriminant analysis (LDA)
  - Within-class covariance normalization (WCCN) (Hatch et.al, Interspeech, 2006)
  - Whitening and length normalization (Garcia-Romero, Interspeech, 2011)
- Verification modeling
  - Probabilistic linear discriminant analysis (PLDA) (Prince et.al, ICCV, 2007)
    - $\boldsymbol{x_{ij}}=\mathbf{m}+\mathbf{\Phi}\beta_{\boldsymbol{i}}+\epsilon_{ij}$
    - ϵ follows $N(0, \Sigma)$, β follows $N(0,1)$ (Garcia-Romero, Interspeech, 2011)
    - ϵ follows $N(0, \Sigma/T_j)$, (Ming Li, Interspeech 2015)
    - Hypothesis testing based scoring

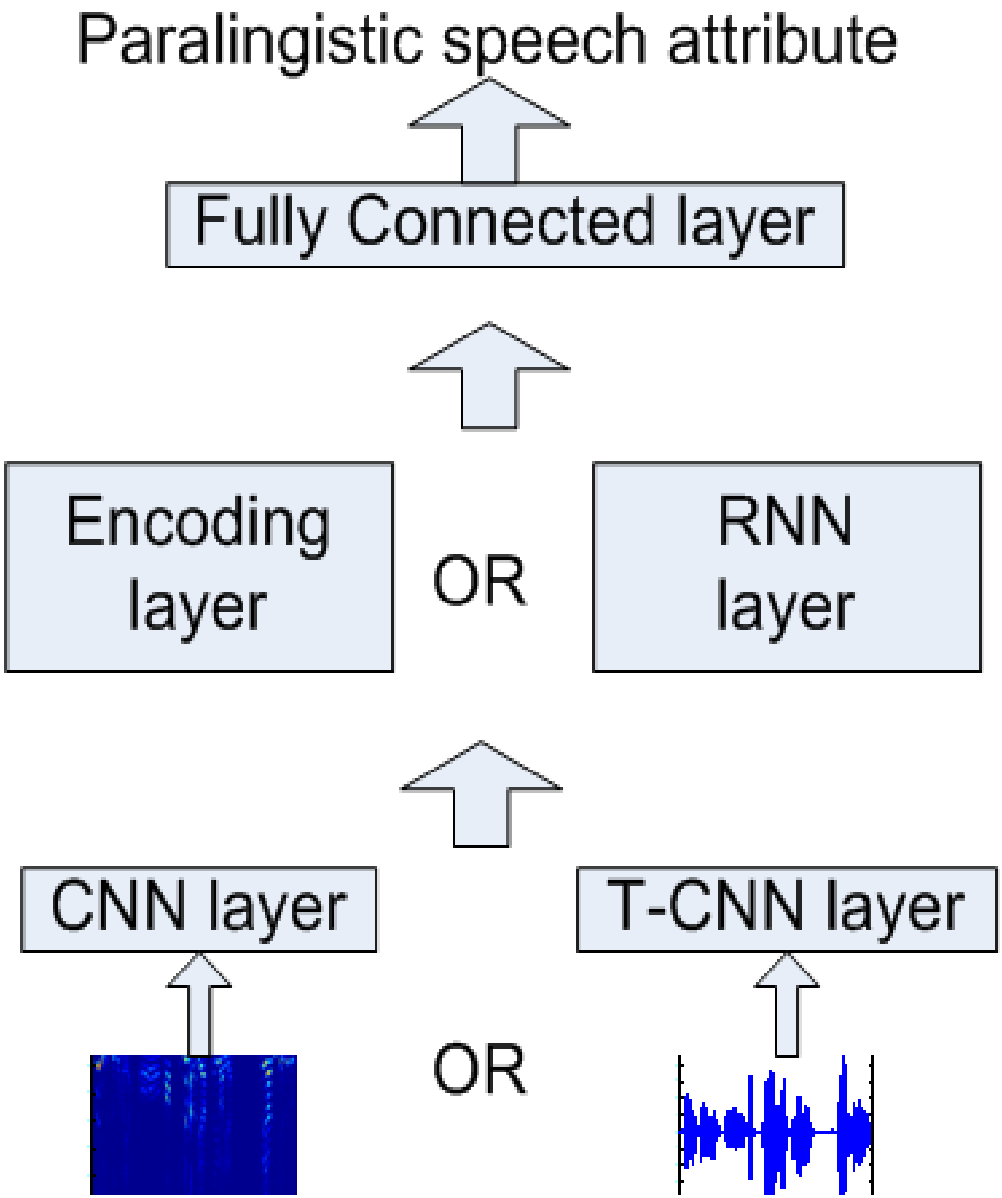
$$Score = \log N\left(\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{bmatrix}; \begin{bmatrix}\boldsymbol{m}\\ \boldsymbol{m}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}+\mathbf{\Phi\Phi}^{\boldsymbol{t}} & \mathbf{\Phi\Phi}^{\boldsymbol{t}}\\ \mathbf{\Phi\Phi}^{\boldsymbol{t}} & \boldsymbol{\Sigma}+\mathbf{\Phi\Phi}^{\boldsymbol{t}}\end{bmatrix}\right) - \log N\left(\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{bmatrix}; \begin{bmatrix}\boldsymbol{m}\\ \boldsymbol{m}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}+\mathbf{\Phi\Phi}^{\boldsymbol{t}} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}+\mathbf{\Phi\Phi}^{\boldsymbol{t}}\end{bmatrix}\right)$$

General framework



	Commonly Used Methods
Feature extraction	MFCC, PNCC, GFCC, CQCC, SDC, LLD, Tandem, Bottleneck, Acoustic-to-articulatory inversion, subglottal, etc.
Representation	GMM-MAP,GMM-supervector,GMM-Ivector,HMM-Ivector,Auto-encoder,DBN, Statistic Measurement, etc.
Variability Compensation	WCCN, JFA, LDA, NAP, NDA, LSDA, LFDA, etc.
Backend classification	SVM, PLDA, NN, ELM, Random Forest， Cosine Similarity, Joint Bayesian, Sparse Representation, etc.

End-to-end framework

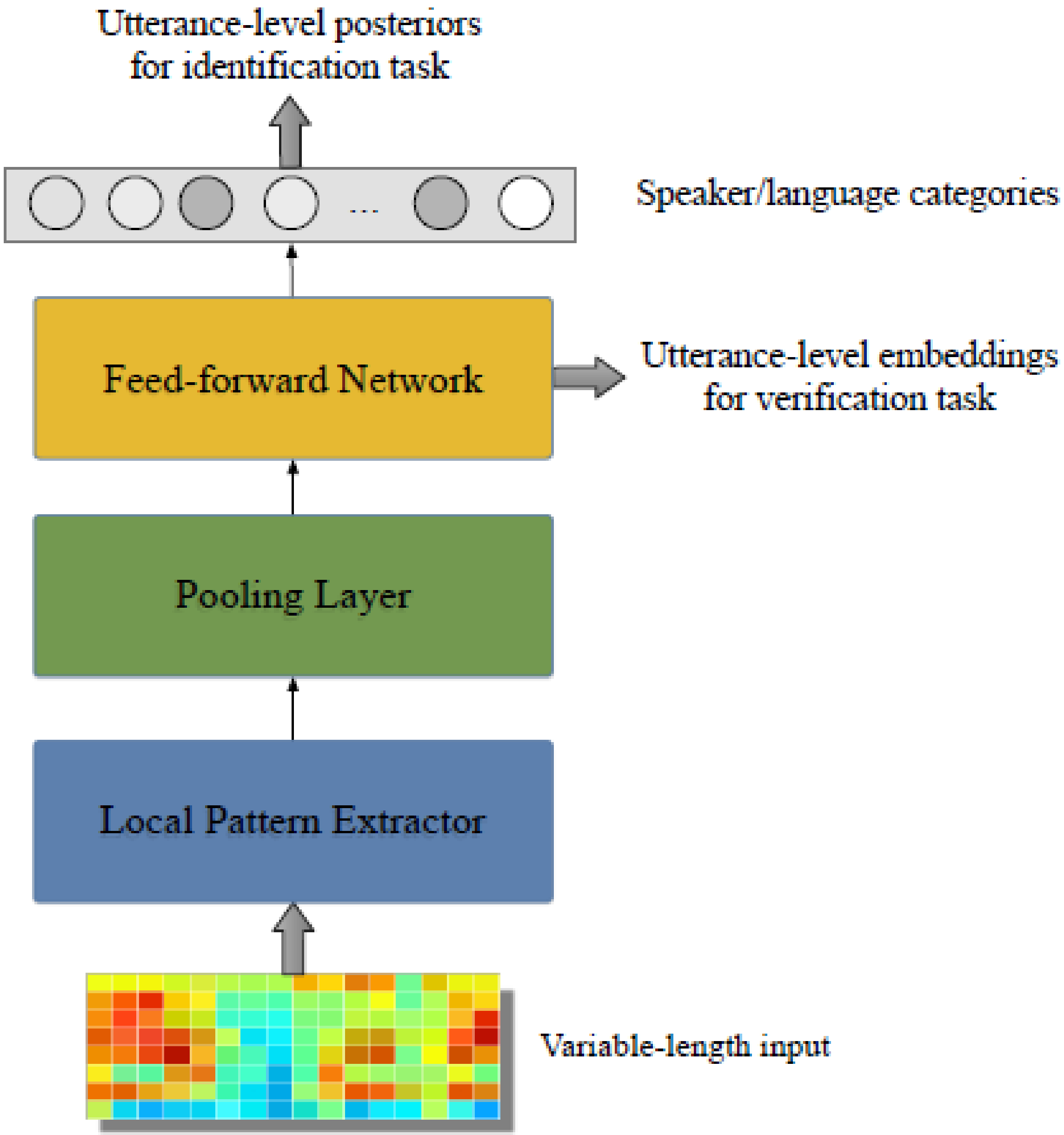


backend classifier

representation

feature extraction

End-to-end framework

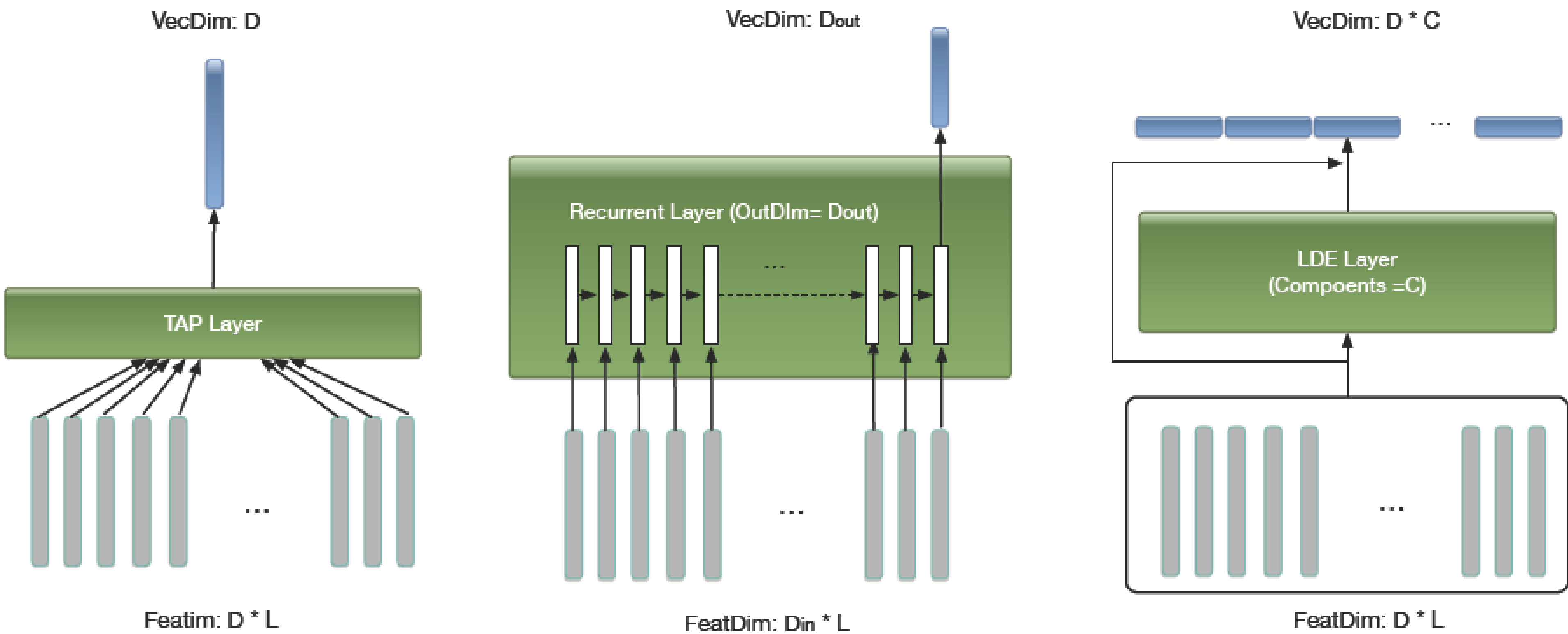


backend classifier

representation

feature extraction

Encoding layer



(a) TAP Layer

(b) Recurrent encoding layer

(c) LDE Layer

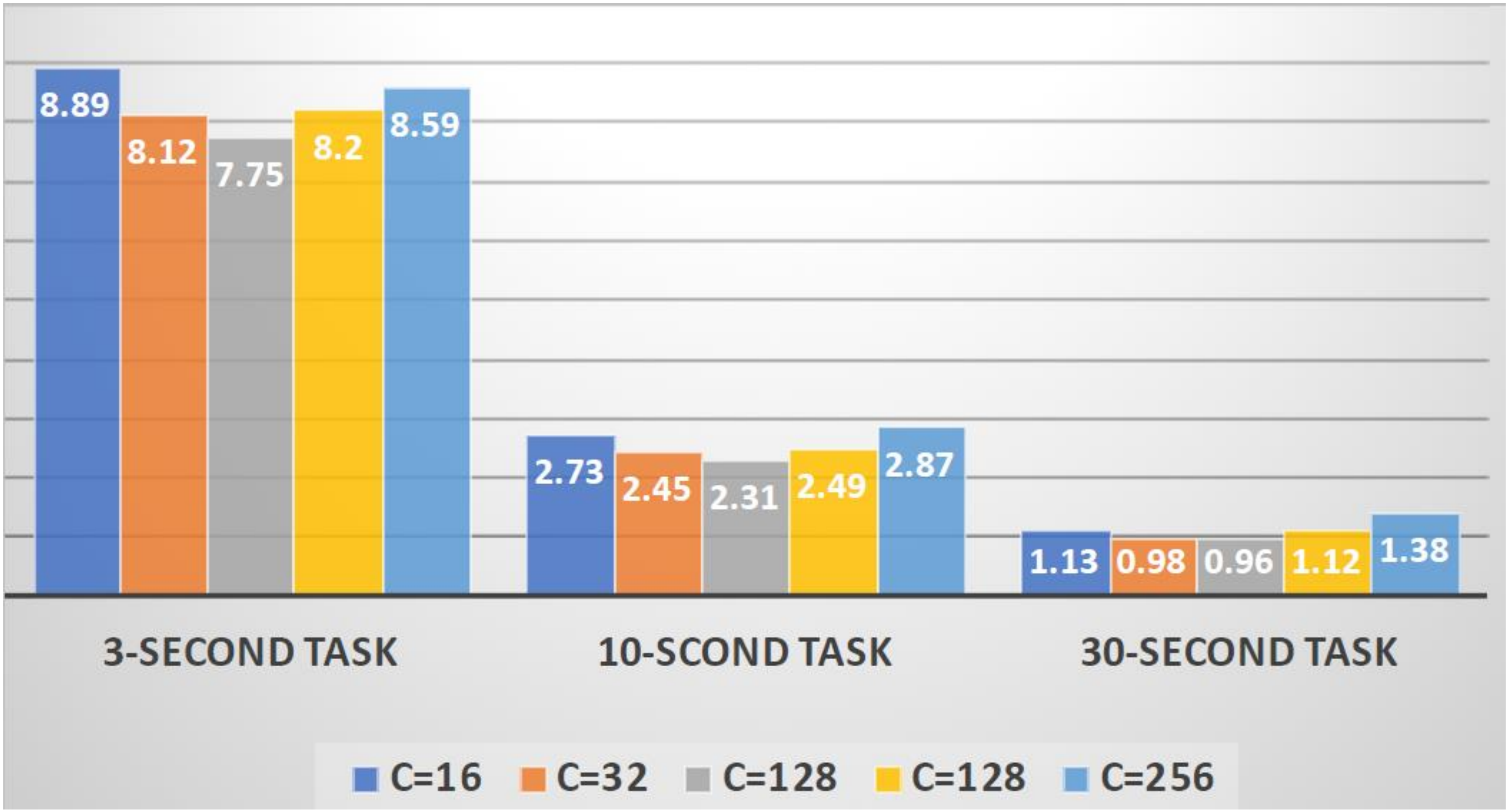
Results on NIST LRE 07 language identification

PERFORMANCE ON THE 2007 NIST LRE CLOSED-SET TASK (LOWER IS BETTER). NR: NOT REPORTED

ID	System	$C_{avg}(\%)/EER(\%)$		
		3s Task	10s Task	30s Task
1	ResNet-TAP	9.98/11.28	3.24/5.76	1.73/3.96
2	ResNet-SAP	8.59/9.89	**2.49**/4.27	**1.09**/2.38
3	ResNet-LDE	**8.25/7.75**	2.61/**2.31**	1.13/**0.96**
4	GMM i-vector [12]	20.46/8.29	3.02/17.71	3.02/2.27
5	DNN i-vector [12]	14.64/12.04	6.20/3.74	2.60/1.29
6	DNN PPP feature [12]	**8.00/6.90**	**2.20/1.43**	**0.61/0.32**
7	DNN Tandem Feature [12]	9.85/7.96	3.16/1.95	0.97/0.51
8	DNN Phonotactic [43]	18.59/12.79	6.28/4.21	1.34/0.79
9	RNN D&C [43]	22.67/15.57	9.45/6.81	3.28/3.25
10	LSTM-Attention [44]	NR/14.72	NR	NR
11	ResNet-GRU [12]	11.31/10.74	5.49/6.40	NR
12	ResNet-LSTM [12]	10.17/9.80	4.66/4.26	NR

Weicheng Cai,Wenbo Liu, Zexin Cai, **Ming Li(*)**. "Insights into end-to-end learning scheme for language identification", ICASSP, 2018.

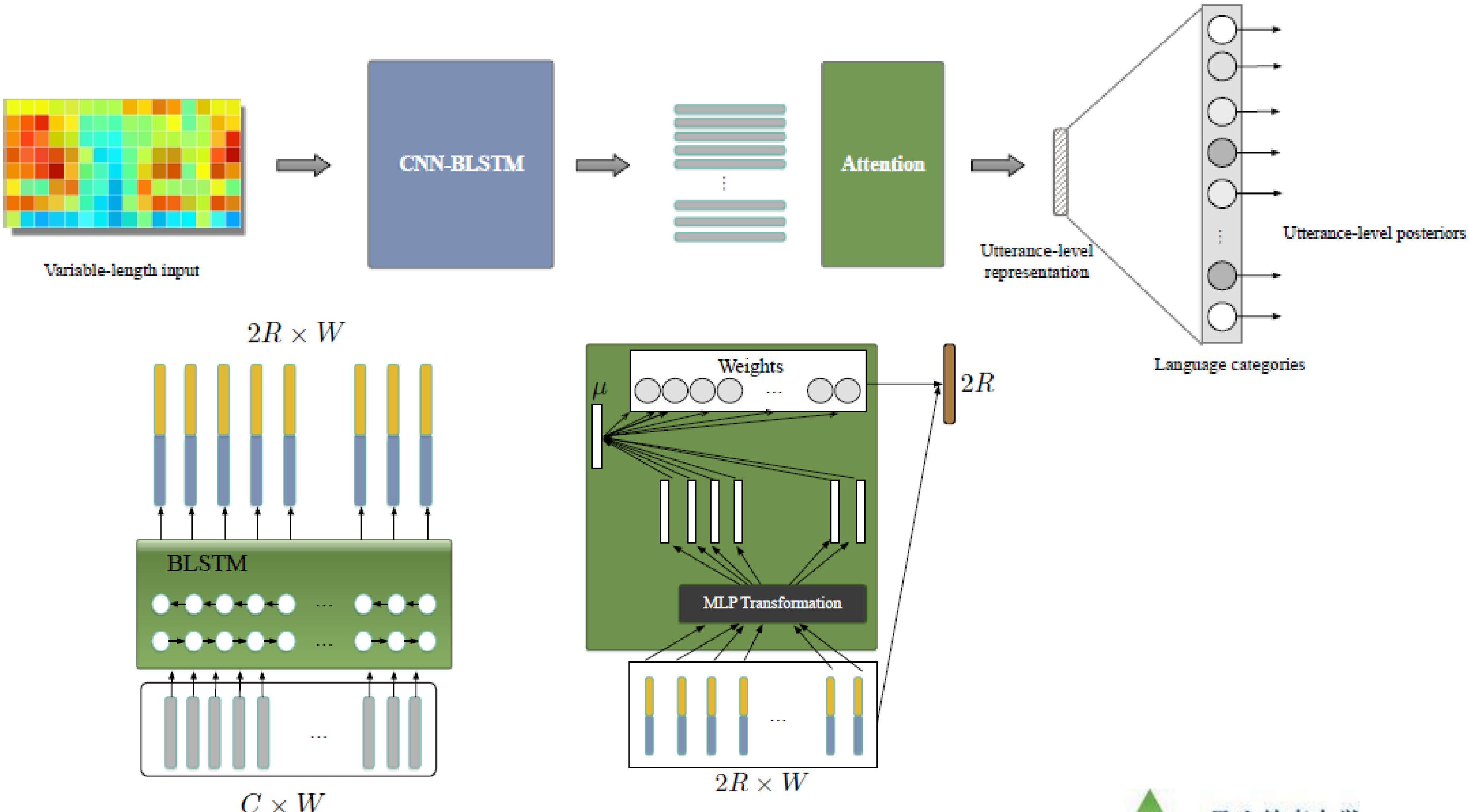
Results on NIST LRE 07 language identification



Weicheng Cai,Wenbo Liu, Zexin Cai, **Ming Li(*)**. "Insights into end-to-end learning scheme for language identification", ICASSP, 2018.

Encoding layer

Attention based CNN-BLSTM



Weicheng Cai, Shen Huang and Ming Li (*), "utterance-level end-to-end language identification using attention-based cnn-blstm", ICASSP 2019

Results on NIST LRE 07 language identification

Table 1. Performance on the 2007 NIST LRE closed-set task

System ID	System Description	Front-end module	Encoding layer	$C_{avg}(\%)$			$EER(\%)$		
				3s	10s	30s	3s	10s	30s
1	CNN-TAP [10]	CNN	GAP	9.98	3.24	1.73	11.28	5.76	3.96
2	CNN-SAP [12]	CNN	SAP	**8.59**	**2.49**	1.09	9,89	4.27	2.38
3	CNN-LSTM [10]	CNN	LSTM	10.17	4.66	N/R	9.80	4.26	N/R
4	CNN-GRU [10]	CNN	GRU	11.31	5.49	N/R	10.74	6.40	N/R
5	LSTM-Attention [24]	LSTM	Attention	14.72	N/R	N/R	N/R	N/R	N/R
6	tandem CNN-BLSTM TAP	CNN-BLSTM	TAP	9.83	3.31	2.03	11.22	5.26	3.67
7	**tandem CNN-BLSTM SAP**	CNN-BLSTM	SAP	9.22	2.54	**0.97**	**9.50**	**3.48**	**1.77**
8	**Fusion ID2 + ID7**			**7.98**	**2.30**	**0.89**	**8.03**	**3.05**	**1.56**

Weicheng Cai, Shen Huang and Ming Li (*), "utterance-level end-to-end language identification using attention-based cnn-blstm", ICASSP 2019
Jinkun Chen, Weicheng Cai and Ming Li(*), "End-to-end Language Identification using NetFV and NetVLAD", ISCSLP 2018.

Loss design

Table 2: Results for verification on VoxCeleb (lower is better)

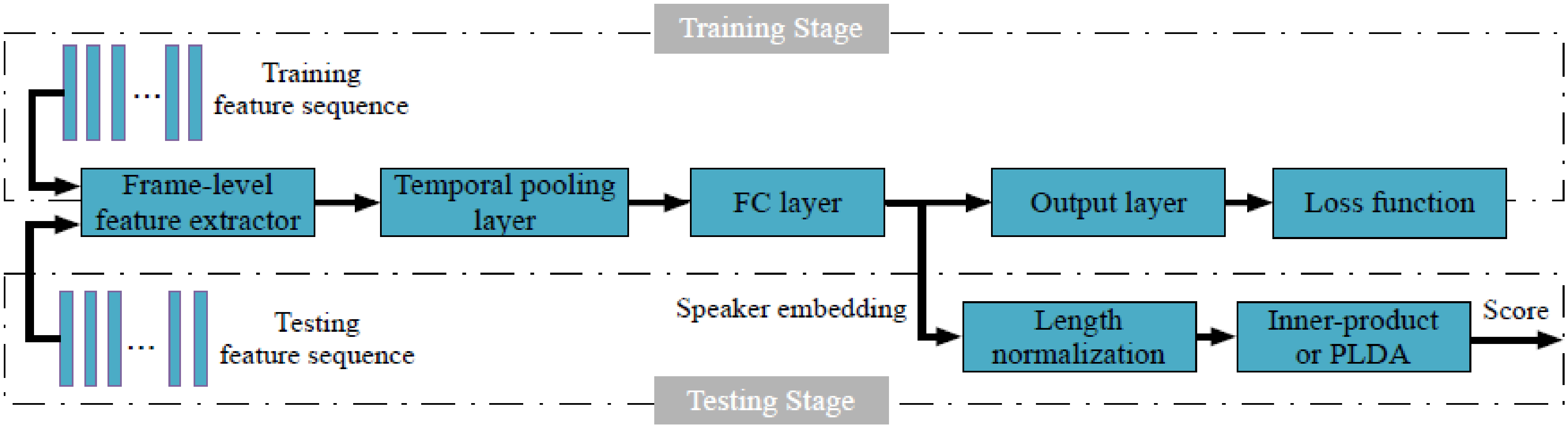
System ID	System Description	Encoding Procedure	Loss Function	Similarity Metric	$C_{det}(\%)$	$EER(\%)$
1	i-vector + cosine	Supervector	GNLL	cosine	0.829	20.63
2	i-vector + PLDA	Supervector	GNLL + GNLL	PLDA	0.639	7.95
3	TAP-Softmax	TAP	softmax	cosine	0.553	5.48
4	TAP-Softmax	TAP	softmax + GNLL	PLDA	0.545	5.21
5	TAP-CenterLoss	TAP	center loss	cosine	0.519	4.99
6	TAP-CenterLoss	TAP	center loss+ GNLL	PLDA	0.608	4.82
7	TAP-ASoftmax	TAP	A-Softmax	cosine	**0.439**	5.27
8	TAP-ASoftmax	TAP	A-Softmax + GNLL	PLDA	0.577	**4.46**
9	SAP-Softmax	SAP	softmax	cosine	0.522	5.51
10	SAP-Softmax	SAP	softmax + GNLL	PLDA	0.545	5.08
11	SAP-CenterLoss	SAP	center loss	cosine	0.509	5.15
12	SAP-CenterLoss	SAP	center loss+ GNLL	PLDA	0.581	4.58
13	SAP-ASoftmax	SAP	A-Softmax	cosine	0.509	4.90
14	SAP-ASoftmax	SAP	A-Softmax + GNLL	PLDA	0.622	4.40
15	LDE-Softmax	LDE	softmax	cosine	0.516	5.21
16	LDE-Softmax	LDE	softmax + GNLL	PLDA	0.519	5.07
17	LDE-CenterLoss	LDE	center loss	cosine	0.496	4.98
18	LDE-CenterLoss	LDE	center loss + GNLL	PLDA	0.632	4.87
19	LDE-ASoftmax	LDE	A-Softmax	cosine	**0.441**	**4.57**
20	LDE-ASoftmax	LDE	A-Softmax + GNLL	PLDA	0.576	4.48

Angular loss, center loss, softmax loss

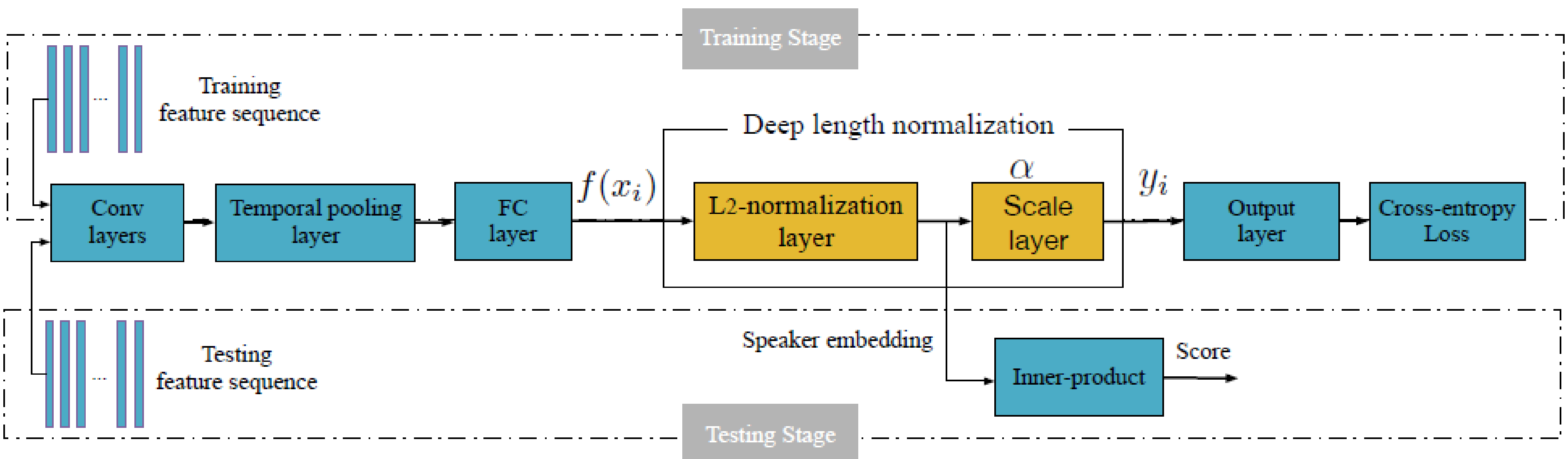
Liu, Weiyang, Yandong Wen, Zhiding Yu, **Ming Li**, Bhiksha Raj, and Le Song. "Sphereface: Deep hypersphere embedding for face recognition." CVPR, vol. 1. 2017.

Weicheng Cai, Jinkun Chen, **Ming Li(*)**. "Exploring the Encoding Layer and Loss function in End-to-End Speaker and Language Recognition System", Odyssey, 2018.

Length normalization layer



Conventional pipeline



The proposed framework with length normalization layer

How to tune Alpha

Table 3: *Verification performance on VoxCeleb1 for various scale parameter α (lower is better)*

System Description	DCF10^{-2}	DCF10^{-3}	$EER(\%)$
Deep embedding baseline	0.553	0.713	5.48
fixed $\alpha = 1$	0.922	0.967	10.18
fixed $\alpha = 4$	0.601	0.828	6.36
fixed $\alpha = 8$	0.515	0.687	5.49
fixed $\alpha = 12$	**0.475**	**0.586**	**5.01**
fixed $\alpha = 16$	0.499	0.696	5.32
fixed $\alpha = 20$	0.503	0.637	5.46
fixed $\alpha = 24$	0.502	0.638	5.54
fixed $\alpha = 28$	0.497	0.640	5.52
trained $\alpha = 26.1$	0.486	0.599	5.60

$$\alpha_{low} = \log \frac{p(C-2)}{1-p}$$

For voxceleb1, C=1211, p=0.9, then Alpha_low=9

昆山杜克大学 DUKE KUNSHAN UNIVERSITY

Weicheng Cai,Jinkun Chen, **Ming Li(*)**. "Analysis of Length Normalization in End-to-End Speaker Verification System", Interspeech, 2018.

Length normalization layer

Table 2: *Voxceleb1 open-set verification task performance, in comparing the effect of our introduced deep length normalization strategy and traditional extra length normalization step (lower is better)*

System Description	Deep L_2-norm	Traditional L_2-norm	Similarity Metric	DCF10^{-2}	DCF10^{-3}	$EER(\%)$
i-vector + inner-product	N/A	✗	inner-product	0.736	0.800	13.80
i-vector + cosine	N/A	✓	inner-product	0.681	0.771	13.80
i-vector + PLDA	N/A	✗	PLDA	0.488	0.639	5.48
i-vector + L_2-norm + PLDA	N/A	✓	PLDA	0.484	0.627	5.41
Deep embedding + inner-product	✗	✗	inner-product	0.758	0.888	7.42
Deep embedding+ cosine	✗	✓	inner-product	0.553	0.713	5.48
Deep embedding+ PLDA	✗	✗	PLDA	0.524	0.739	5.90
Deep embedding + L_2-norm + PLDA	✗	✓	PLDA	0.545	0.733	5.21
L_2-normalized deep embedding + inner-product	✓	✗	inner-product	**0.475**	**0.586**	5.01
L_2-normalized deep embedding + PLDA	✓	✗	PLDA	0.525	0.694	**4.74**

昆山杜克大学
DUKE KUNSHAN UNIVERSITY

Weicheng Cai,Jinkun Chen, **Ming Li(*)**. "Analysis of Length Normalization in End-to-End Speaker Verification System", Interspeech, 2018.

Results on Voxceleb1 data

PERFORMANCE RESULTS ON VOXCELEB1 (LOWER IS BETTER). DA: DATA AUGMENTATION

ID	System	DA	Training Set	Loss + Scoring	C_{det}	$EER(\%)$
1	ResNet-TAP	✗	Voxceleb1	Softmax + Cosine	0.553	5.48
2	ResNet-SAP	✗	Voxceleb1	Softmax + Cosine	0.522	5.51
3	ResNet-LDE	✗	Voxceleb1	Softmax + Cosine	0.516	5.21
4	ResNet-TAP	✗	Voxceleb1+Voxceleb2	Softmax + Cosine	0.331	3.28
5	ResNet-SAP	✗	Voxceleb1+Voxceleb2	Softmax + Cosine	0.307	3.11
6	ResNet-LDE	✗	Voxceleb1+Voxceleb2	Softmax + Cosine	**0.291**	**2.89**
7	i-vector	✗	Voxceleb1	PLDA	0.484	5.41
8	i-vector	✗	Voxceleb1+Voxceleb2	LDA+PLDA	0.493	5.32
9	i-vector [16]	✓	Voxceleb1+PRISM	PLDA	0.479	5.39
10	x-vector	✗	Voxceleb1	Softmax + Cosine	0.726	11.42
11	x-vector	✓	Voxceleb1 + MUSAN	Softmax + Cosine	0.727	10.11
12	x-vector	✗	Voxceleb1	Softmax + PLDA	0.570	7.74
13	x-vector	✓	Voxceleb1 + MUSAN	Softmax + PLDA	0.485	6.20
14	x-vector	✓	Voxceleb1 + MUSAN	Softmax + LDA+PLDA	0.480	5.64
15	x-vector [16]	✓	Voxceleb1 + PRISM	Softmax + PLDA	0.413	4.19
16	x-vector	✓	Voxceleb1+Voxceleb2+MUSAN	Softmax + LDA+PLDA	**0.325**	**3.12**
17	Chung *et al.* [11]	✗	Voxceleb1	Softmax + Cosine	0.75	10.2
18	Chung *et al.* [11]	✗	Voxceleb1	Contrastive + Cosine	0.71	7.8
19	Cai *et al.* [17]	✗	Voxceleb1	A-Softmax + Cosine	0.441	4.56
20	Hajibabaei *et al.* [45]	✗	Voxceleb1	AM-Softmax + Cosine	0.413	4.30
21	Chung *et al.* [40]	✗	Voxceleb2	Contrastive + Cosine	0.429	3.95

Results on SITW data

PERFORMANCE RESULTS ON SITW (LOWER IS BETTER). DA: DATA AUGMENTATION. N/A: NOT APPLICABLE

ID	System	DA	Training Set	Loss + Scoring	SITW Dev		SITW Eval	
					C_{det}	$EER(\%)$	C_{det}	$EER(\%)$
1	ResNet-TAP	✗	Voxcele1+Voxceleb2	Softmax + Cosine	0.376	4.96	0.454	5.66
2	ResNet-SAP	✗	Voxceleb1+Voxceleb2	Softmax + Cosine	0.334	4.34	0.405	5.17
3	ResNet-LDE	✗	Voxceleb1+Voxceleb2	Softmax + Cosine	**0.298**	**3.95**	**0.349**	**4.52**
4	i-vector	✓	Voxceleb1+Voxceleb2+MUSAN	LDA+PLDA	0.425	4.81	0.463	5.65
5	x-vector	✗	Voxceleb1+Voxceleb2	Softmax + Cosine	0.827	16.55	0.887	17.19
6	x-vector	✓	Voxceleb1+Voxceleb2+MUSAN	Softmax + Cosine	0.777	15.05	0.818	15.30
7	x-vector	✗	Voxceleb1+Voxceleb2	Softmax + LDA+PLDA	0.377	3.77	0.410	4.31
8	x-vector	✓	Voxceleb1+Voxxceleb2+MUSAN	Softmax + LDA+PLDA	**0.313**	**3.08**	**0.348**	**3.41**
9	x-vector [29]	✓	SITW Dev+NIST SREs+Voxxceleb1+MUSAN	Softmax + LDA+PLDA	N/A	N/A	0.393	4.16

昆山杜克大学
DUKE KUNSHAN UNIVERSITY

Results on NIST SRE 2018

Table 1: *NIST SRE 2018 CMN2 results for fixed condition (EER[%] / minC / actC)*

x-vector	LDA + inW + PLDA	-	07.77 / 0.587 / 0.605	08.89 / 0.587 / 0.596
	LDA + CORAL + inW + PLDA	-	07.09 / 0.469 / 0.559	07.43 / 0.518 / 0.584
	LDA + PLDA	AS-Norm2	07.17 / 0.479 / 0.779	**07.68 / 0.492 / 0.770**
	LDA + CORAL + PLDA	AS-Norm2	**07.32 / 0.419 / 0.715**	07.50 / 0.504 / 0.730

Encoding Layer	Loss	CMN2	
		Development	Evaluation
GAP	softmax	7.85 / 0.501 / 0.790	7.43 / 0.557 / 0.794
GAP	A-softmax	6.03 / 0.420 / 0.636	6.61 / 0.474 / 0.654
GSP	softmax	7.03 / 0.481 / 0.550	7.12 / 0.489 / 0.541
GSP	A-softmax	5.94 / 0.418 / 0.704	**6.14 / 0.463 / 0.700**
LDE	softmax	**7.50 / 0.408 / 0.716**	7.17 / 0.503 / 0.731
LDE	A-softmax	**6.03 / 0.354 / 0.425**	**6.20 / 0.430 / 0.448**

Results on Voices 2019 fixed condition

Table 1: *Development subset results for the speaker recognition task of the VOiCES from a distance challenge (SN represents Score Normalization, devW represents whitening using development sub-set)*

Front-end	Back-end	WPE	SN	Development sub-set			Evaluation		
				minC	actC	EER[%]	minC	actC	EER[%]
MFCC i-vector	PLDA	-	√	0.4935	0.6747	6.33	0.8037	0.8294	12.92
	CORAL + devW + PLDA	√	√	0.4527	0.4703	6.12	0.6870	0.6891	11.89
PNCC i-vector	PLDA	-	√	0.5073	0.6745	6.12	0.6791	0.7803	10.18
	CORAL + devW + PLDA	√	-	0.4594	0.4697	5.29	0.6498	0.7152	10.09
x-vector	CORAL + PLDA	-	√	0.4018	0.4151	4.96	0.6377	0.6492	09.13
	CORAL + PLDA	√	-	0.3617	0.3688	4.52	0.5417	0.5544	07.54
Mfbank-8k ResNet + Softmax	CORAL + devW + PLDA	-	-	0.4557	0.5246	5.41	0.6608	0.7128	10.92
	CORAL + devW + PLDA	√	-	0.3934	0.4611	4.59	0.5929	0.6424	09.75
Mfbank-16k ResNet + Softmax	cosine similarity	-	-	0.3608	1	3.81	0.6262	1	08.75
	cosine similarity	√	-	0.3245	1	3.02	0.5507	1	07.91
Mfbank-16k ResNet + A-Softmax	cosine similarity	-	-	0.2735	1	**2.73**	0.4156	1	**05.84**
	cosine similarity	√	-	**0.2485**	1	**2.41**	**0.3668**	1	**05.58**
Gfbank ResNet + A-Softmax	cosine similarity	-	-	0.3065	1	3.52	0.4411	1	06.78
	cosine similarity	√	-	**0.2680**	1	3.14	**0.4056**	1	06.49

Results on Voices 2019 fixed condition

Fusion strategy	Development sub-set				Evaluation			
	minC	actC	EER[%]	Cllr	minC	actC	EER[%]	Cllr
Best single system	0.2485	1	2.41	0.8060	0.3668	1	5.58	0.8284
Each embedding with top 1 back-end	0.1831	0.1857	1.93	0.0808	**0.3205**	**0.3214**	**4.60**	**0.2335**
Each embedding with top 2 back-end	0.1644	0.1659	1.48	0.0710	0.3555	0.3578	4.79	0.2684
Each embedding with top 3 back-end (submission)	**0.1473**	**0.1484**	**1.21**	**0.0577**	0.3532	0.3609	4.96	0.2683

Results on OLR 2018 dev dataset

Table 1. AP18-OLR development set performance

Feature	Modeling	$C_{avg} \times 100$	
		Full-length	1 second
MFCC	i-vector + LR	3.58	14.23
PPP	i-vector + LR	2.23	14.54
Tandem	i-vector + LR	2.77	13.21
BNF	i-vector + LR	3.17	20.74
MFCC	x-vector + LR	3.45	11.85
PPP	x-vector + LR	1.78	11.47
BNF	x-vector + LR	1.97	15.48
Fbank	CNN-GAP	4.63	8.98
PPP	CNN-GAP	1.49	11.02
Tandem	CNN-GAP	2.08	9.62
Fusion		0.85	5.76

Challenges & opportunities for the end-to-end speaker and language recognition task

Network structure

Data augmentation

Loss function design

Transfer learning

Joint learning & multitask learning

…

Thank you very much!

ming.li369@duke.edu

https://scholars.duke.edu/person/MingLi

2019年声纹识别研究与应用学术研讨会
4月20日　昆山杜克大学

主办方：中国计算机学会；昆山杜克大学
协办方：昆山杜克大学大数据研究中心
清华大学媒体大数据认知计算研 究中心
中国计算机学会语音对话与听觉专业工作组

参会专家（按姓氏首字母排列）

杜俊	中国科学技术大学
郭武	中国科学技术大学
洪青阳	厦门大学
何亮	清华大学
黄申	腾讯
胡伟湘	华为
李明	昆山杜克大学
欧智坚	清华大学
钱彦旻	上海交通大学
宋彦	中国科学技术大学
王东	清华大学
谢磊	西北工业大学
杨琳	联想
张鹏远	中国科学院声学研究所
张卫强	清华大学
张翔	腾讯
张晓雷	西北工业大学
周瑜	阿里巴巴

This research was funded in part by the National Natural Science Foundation of China (61773413), Natural Science Foundation of Guangzhou City (201707010363) and Six talent peaks project in Jiangsu Province (JY-074).