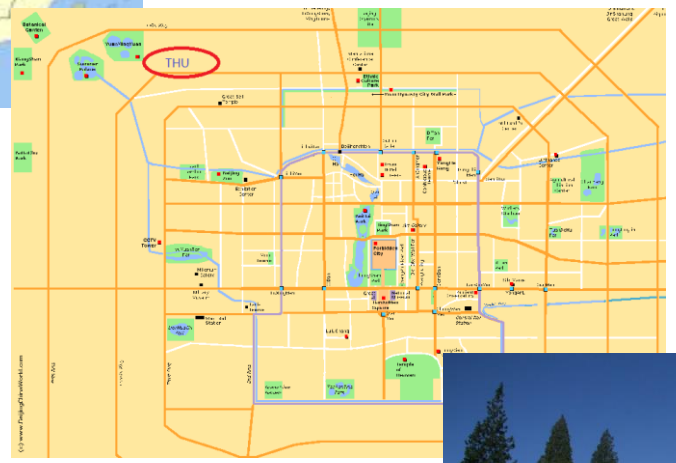
Deep Feature Learning and Normalization for Speaker Recognition

Dong Wang

CSLT, Tsinghua University

2019.07

Tsinghua University



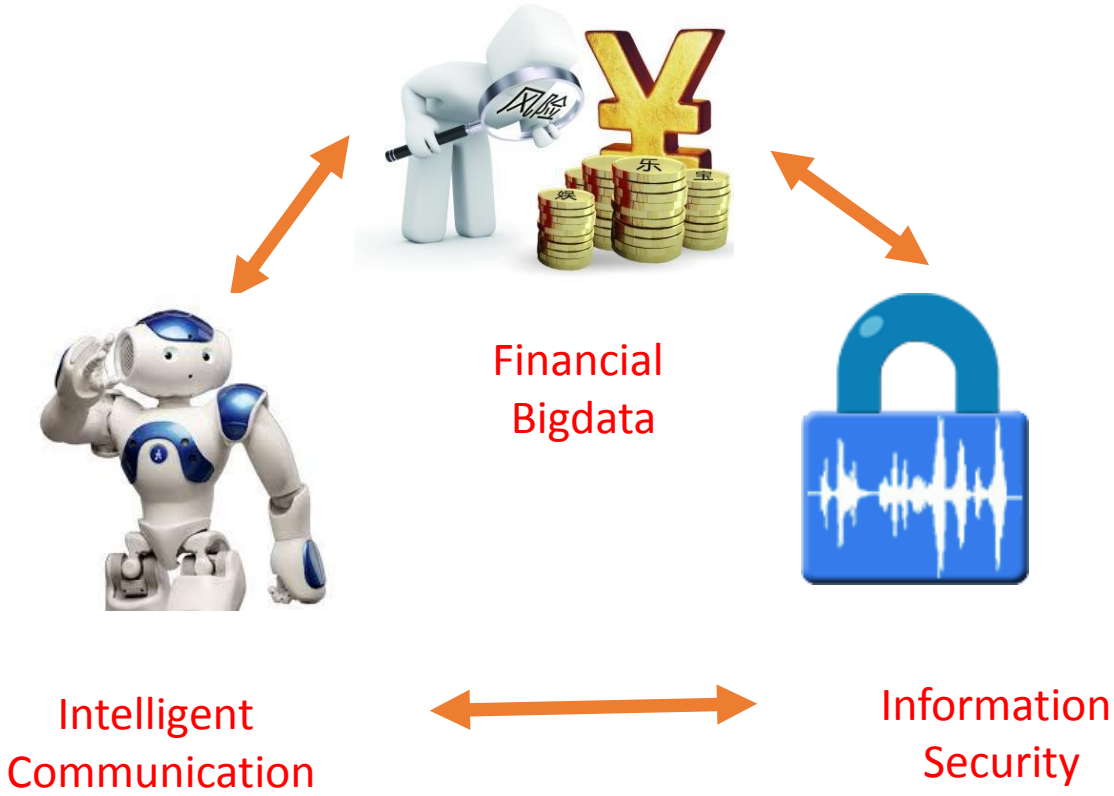
清华大学

Motto	自强不息、厚德载物[1]
Motto in English	Self-Discipline and Social Commitment[2]
Type	Public
Established	1911
President	Qiu Yong
Party Secretary	Chen Xu
Academic staff	3,133
Administrative staff	4,101
Undergraduates	15,184
Postgraduates	16,524
Location	Beijing, China
Campus	Urban, 395 hectares (980 acres)
Flower	Redbud and Lilac
Colors	Purple and White
Affiliations	AEARU, APRU, C9, BRICS Universities League

Center for Speech and Lanugage Technologies (CSLT)

- Established in 1979.
- Director Prof. Fang Zheng
- Focus on speech processing, language processing and finance processing

CSLT research goals



About me

- Dong Wang
  - Associate professor at Tsinghua University
  - Deputy director of CSLT@Tsinghua University
  - Chair of APSIPA SLA
- Brief resume
  - 1995-2002: Bachelor and Master at Tsinghua University
  - 2002-2004: Oracle China
  - 2004-2006: IBM China
  - 2006-2010: PhD candidate and Marie Curie Fellow at University of Edinburgh, UK
  - 2010-2011: Post-doc Fellow at EURECOM, France
  - 2011-2012: Nuance, US
  - 2012- present: Tsinghua University

APSIPA

- 7 TCs
- ASC
- Transactions
- News letter
- Friend lab
- Distinguished lecture

APSIPA DL program

- Promote education
- International collaboration

An objective of APSIPA is to provide educational activities in signal and information processing. A number of tutorials are organized at every APSIPA ASC. Other educational activities we will organize include distinguished lecturer program, summer school, etc.

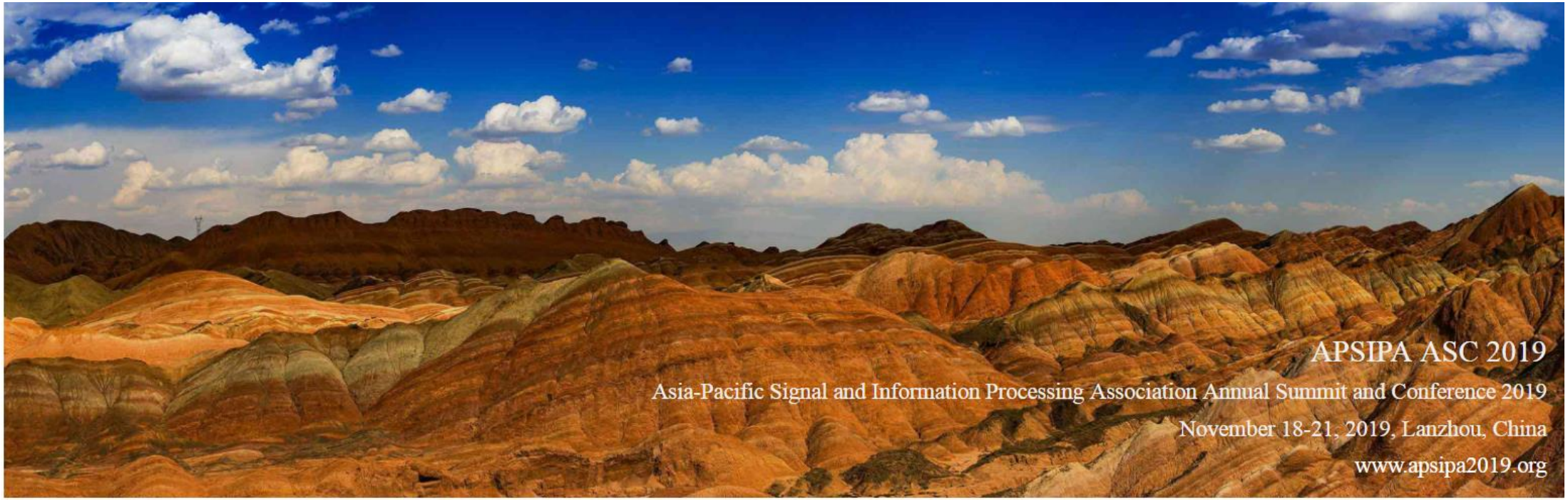
1. **Distinguished Lecturer Program**
2. **Institutional Relations and Education (IRE) Board**
3. **Previous Distinguished Lectures by DLs For 2017-2018**
4. **Previous Distinguished Lectures by DLs For 2016-2017**
5. **Previous Distinguished Lectures by DLs For 2015-2016**
6. **Previous Distinguished Lectures by DLs For 2014-2015**
7. **Previous Distinguished Lectures by DLs For 2013-2014**
8. **Previous Distinguished Lectures by DLs For 2012-2013**
9. **Winter/Summer School**

Distinguished Lecturer Program >> More Details Information

Distinguished Lecturers for 2018-2019:

Hemant A Patil, *Dhirubhai Ambani Institute of Information and Communication Technology, India*
Hiroshi Saruwatari, *The University of Tokyo, Japan*
Junosng Yuan, *Nanyang Technological University, Singapore*
Dong Wang, *Tsinghua University, China*
Jie Chen, *Northwestern Polytechical University, China*
Chuang Shi, *University of Electronic Science and Technology of China*
Xiangui Kang, *Sun Yat-sen University, China*
Jiaching Wang, *National Central University, Taiwan*
Xie Lei, *Northwestern Polytechnical University, China*

APSIPA 2019 in Lanzhou



The talk is about…

- Can we discover fundamental speaker features?

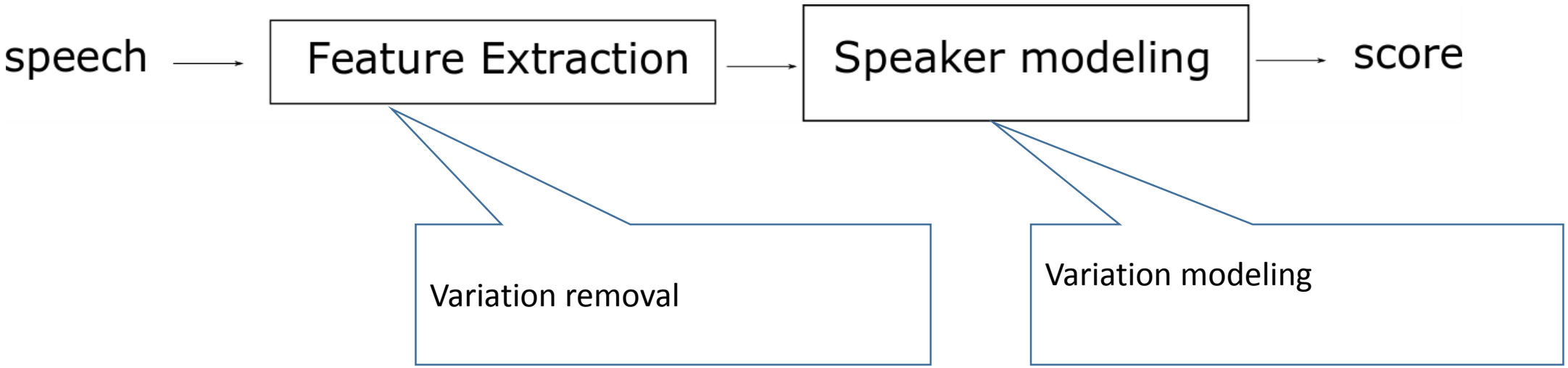
Two things we will talk

- How to extract features
- How to use those features

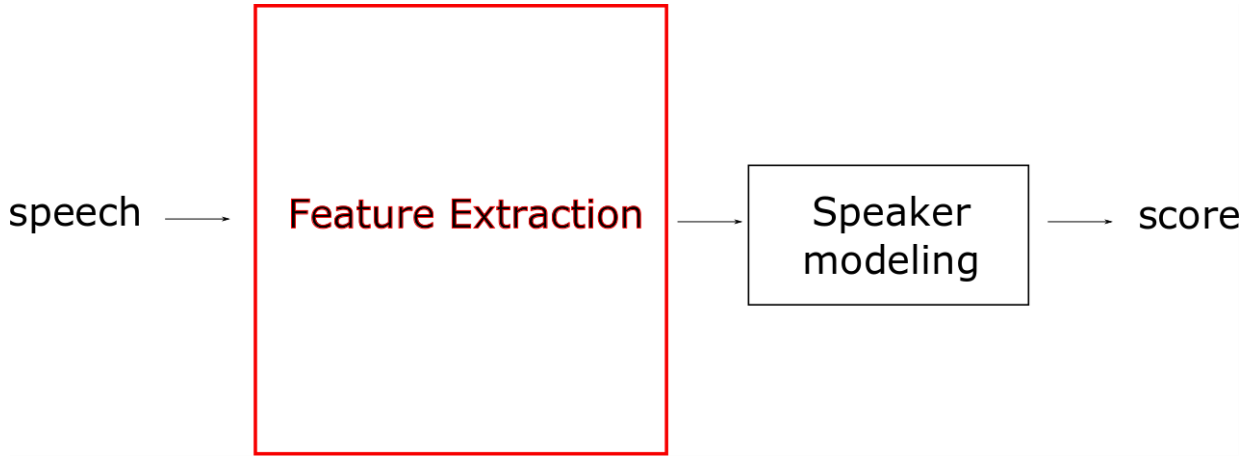
Deep Feature Learning

A classical view: variation compression

- Variation: phonetics, acoustics, physical, pysiological, emotional
- Duration is a very special variation

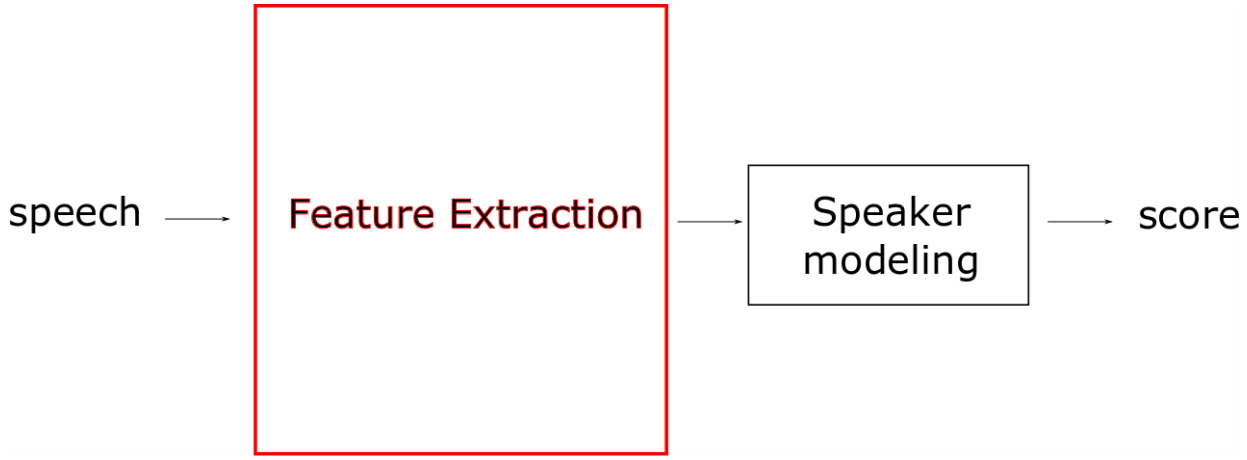


Feature-based approach



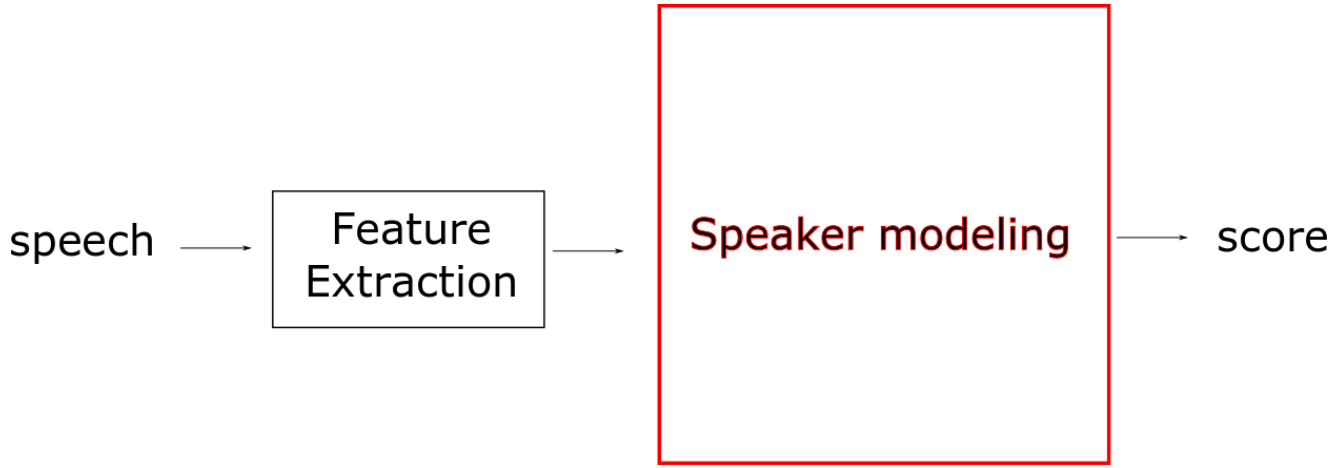
- Powerful feature plus simple model
  - Short-term features (MFCC, PLP)
  - Voice source features (LP)
  - Spectral-temporal features (delta, or long-term feature)
  - Prosodic features: F0, speaking rate, phone duration
  - High-level features: usaga of words and phones, pdf of articulary or acoustic units

Feature-based approach



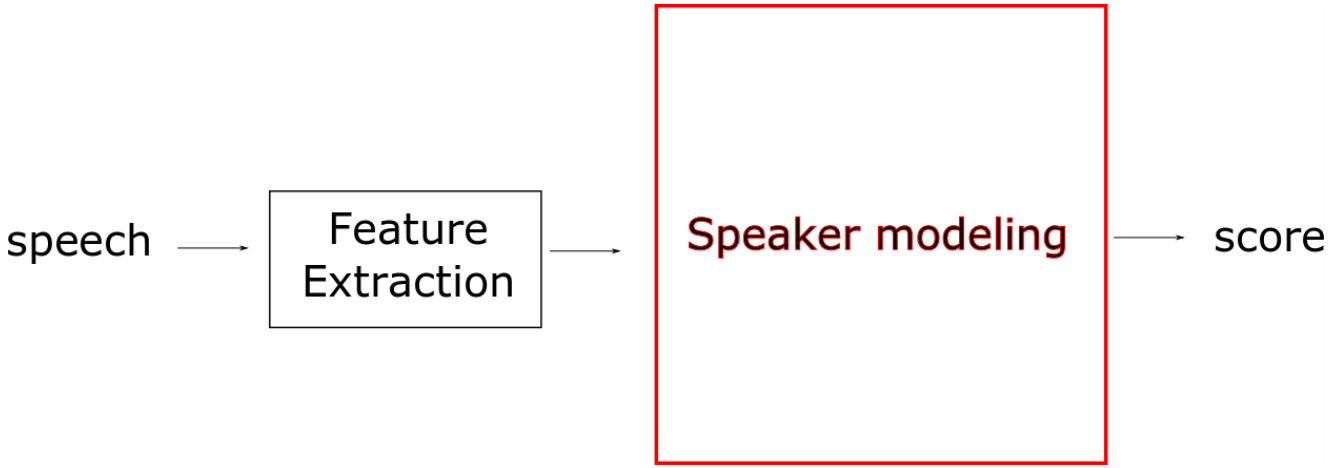
- Long term features tend to be changed by speaking style
- Short term features are noisy, so require probabilistic models

Model-based approach



- Primary feature plus comprehensive models
  - GMM-UBM
  - JFA/i-vector

Model-based approach



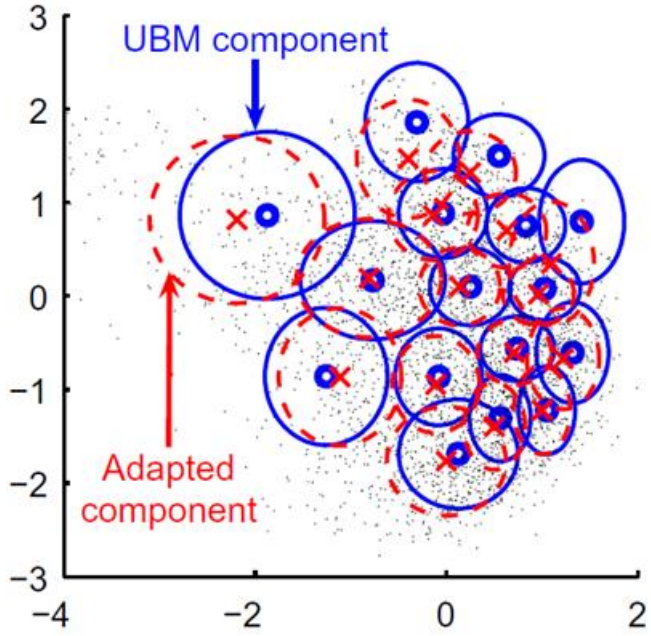
- Principles
  - Using probabilistic model to address variation
  - Length, residual noise…

Who won? A historical perspective

- In short, model-based approach largely wins
  - Long-term and complex features often vary much
  - Carefully designed features are fragile
  - Most importantly, they are hard to model (we come back later).
- Simple features plus a probabilistic model worked the best

What is means?

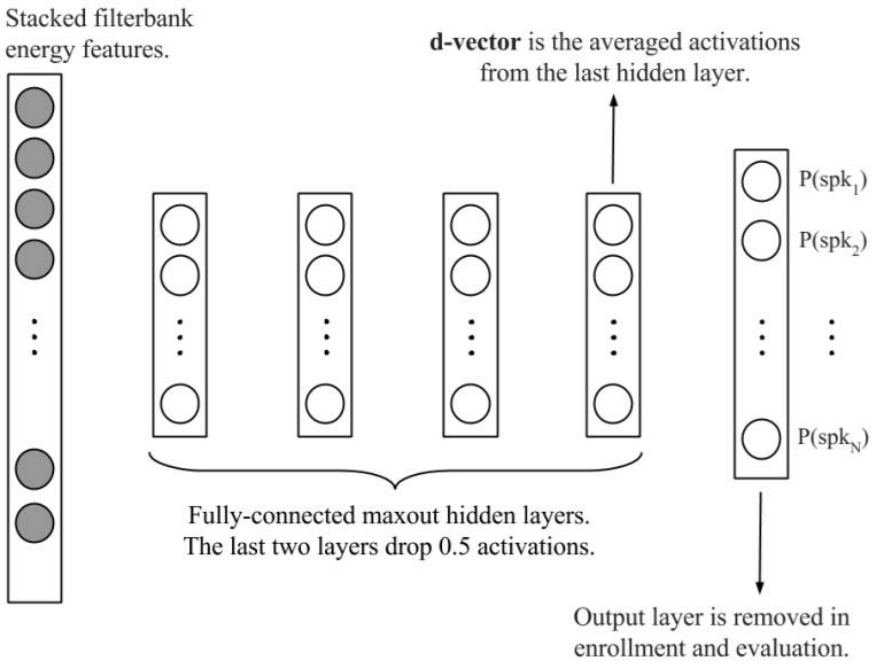
Speaker characteristics are probabilistic patterns!



But it is true?

- This ‘inference’ is based on experimental results
- Perceptual intuition seems an ‘a’ is discriminative
- We still believe some fundamental features exist, but:
  - Need a new approach to extract them
  - Need a new approach to use them

Deep Feature learning



- Learn speaker-dependent features driven by speaker discrimination
- Frame-based representation, average-based back-end

E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez- Dominguez, “Deep neural networks for small footprint text-dependent speaker verification,” ICASSP 2014.

More dedicated structure

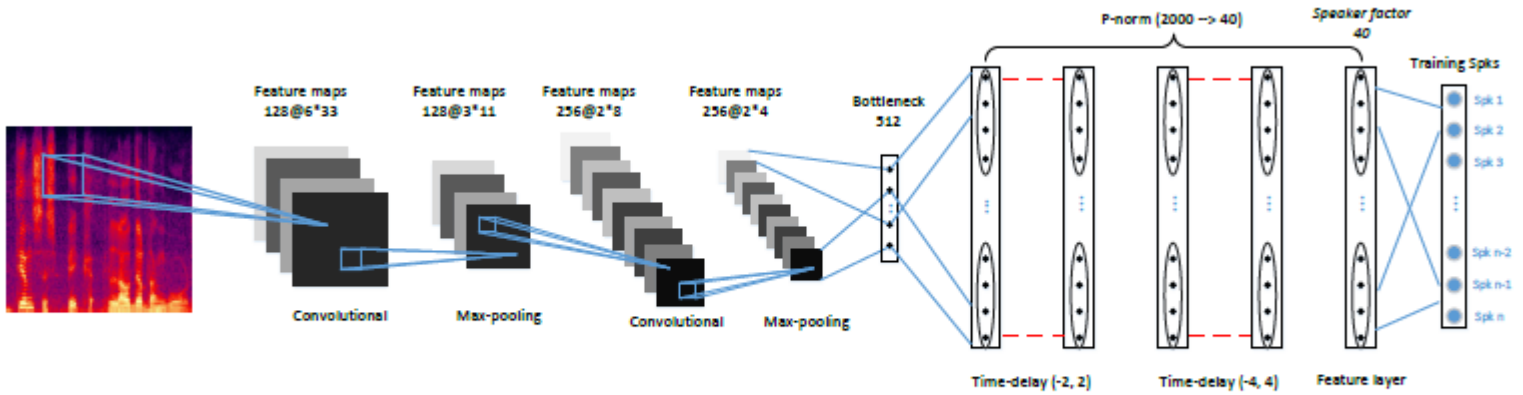
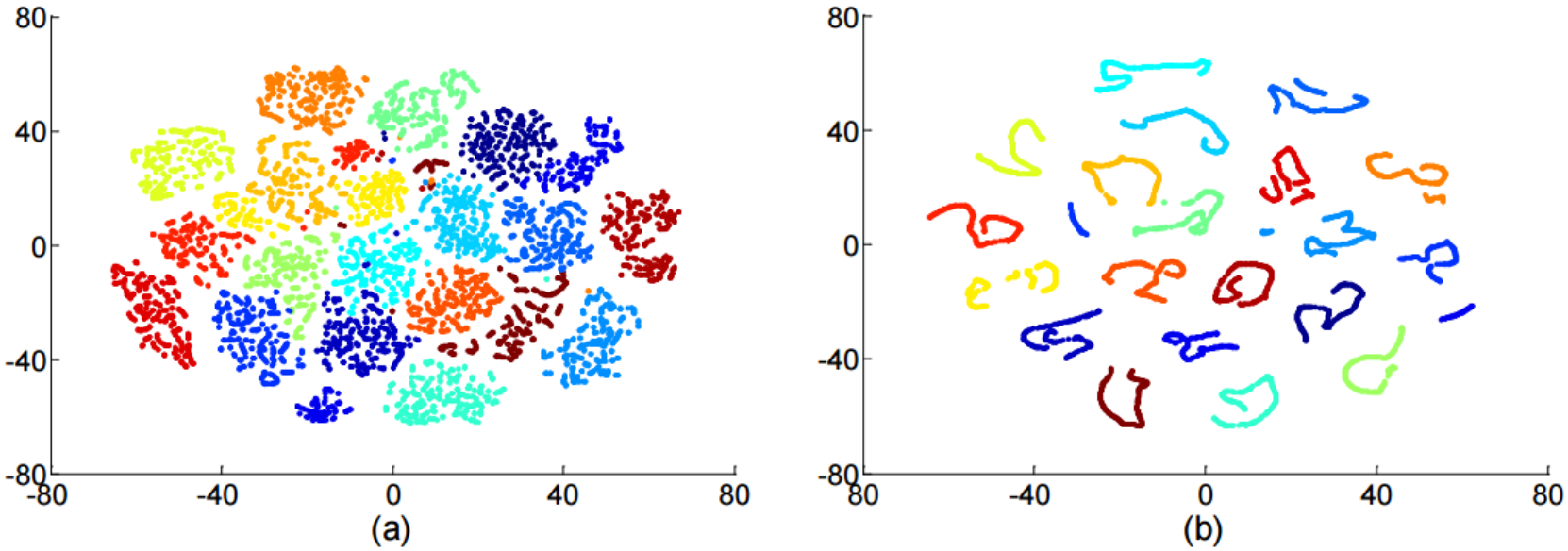


Figure 1: The DNN structure used for deep speaker factor inference.

- L.L et al, Deep Speaker Feature Learning for Text-independent Speaker Verification, Interspeech 2017.

Very discriminative short-term features



- L.L et al, Deep Speaker Feature Learning for Text-independent Speaker Verification, Interspeech 2017.

Systems	Metric	20 frames	50 frames	100 frames
i-vector	Cosine	30.01	18.23	11.14
	LDA	29.47	15.96	8.64
	PLDA	29.29	15.71	8.34
d-vector	Cosine	**7.68**	6.67	4.61
	LDA	7.88	**4.72**	**3.02**
	PLDA	20.81	15.02	8.98

- L.L et al, Deep Speaker Feature Learning for Text-independent Speaker Verification, Interspeech 2017.

What is means?

Speaker characteristics are largely short-term patterns!

That is really interesting

- Our personalities can be determined in 0.3 second
- We can largely factorize/manipulate speech signals based on short spectrum
- …

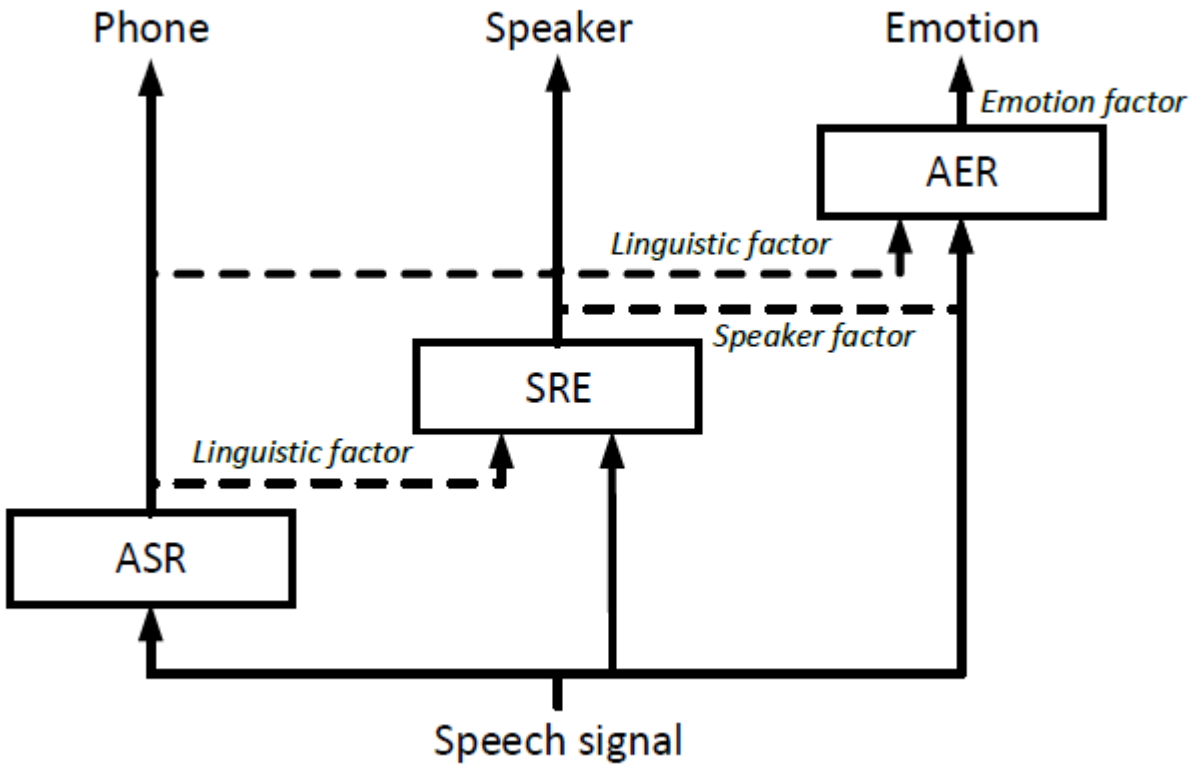
Let’s discriminate cough and laugh

DER%					
Cough	Laugh	‘Hmm’	‘Tsk-tsk’	‘Ahem’	Sniff
20.20	20.71	19.70	42.42	26.26	35.86

		EER%					
Systems	Metric	Cough	Laugh	‘Hmm’	‘Tsk-tsk’	‘Ahem’	Sniff
i-vector	Cosine	23.42	27.69	15.71	29.70	18.12	37.78
	LDA	26.14	27.99	15.54	31.79	20.83	37.74
	PLDA	27.82	25.79	14.28	33.57	21.85	34.76
d-vector	Cosine	8.89	12.43	**5.88**	16.75	10.44	**11.91**
	LDA	**8.33**	**11.20**	6.76	**15.95**	**9.71**	12.44
	PLDA	10.26	15.48	7.28	17.85	13.16	12.93

- Miao Zhang, Yixiang Chen, Lantian Li and Dong Wang, "Speaker Recognition with Cough, Laugh and `Wei'", APSIPA 2017
- Miao Zhang, Xiaofei Kang, Yanqing Wang, Lantian Li, Zhiyuan Tang, Haisheng Dai, Dong Wang, "HUMAN AND MACHINE SPEAKER RECOGNITION BASED ON SHORT TRIVIAL EVENTS", ICASSP 2018

Let’s do speech factorization



Lantian Li, Dong Wang, Yixiang Chen, Ying Shi, Zhiyuan Tang, "DEEP FACTORIZATION FOR SPEECH SIGNAL", ICASSP 2018

Completness of the factorization

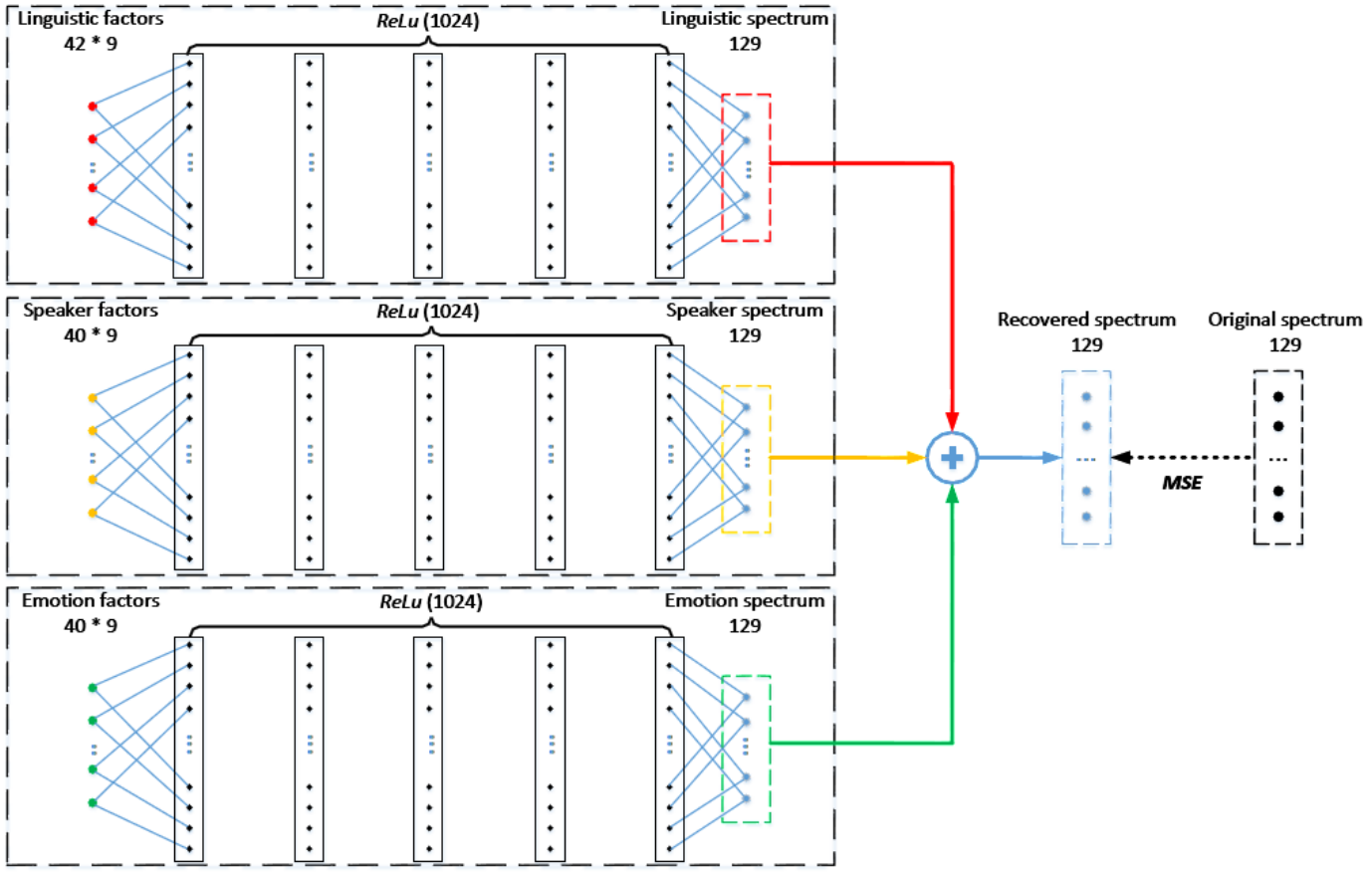
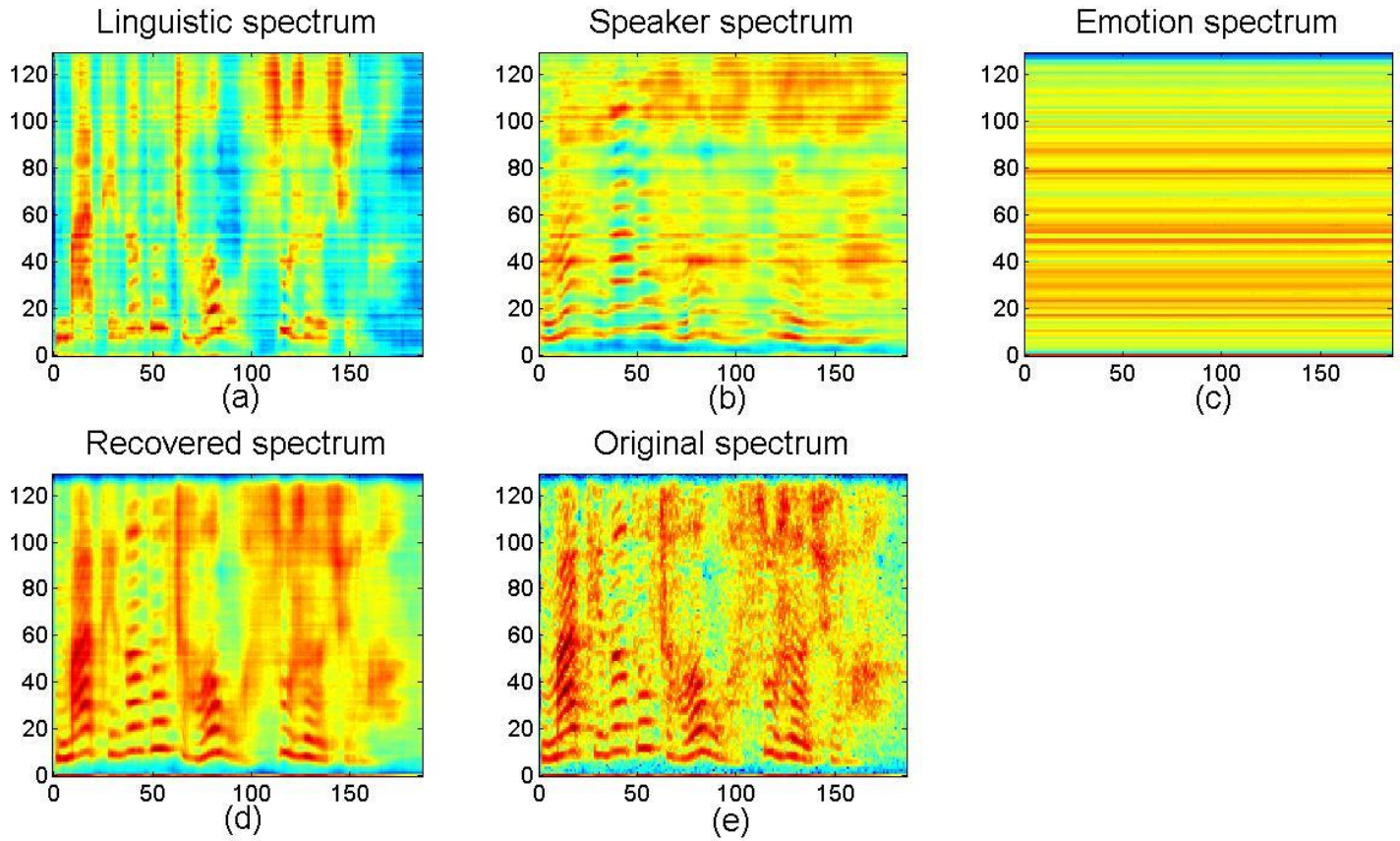


Figure 3: The architecture for spectrum reconstruction.

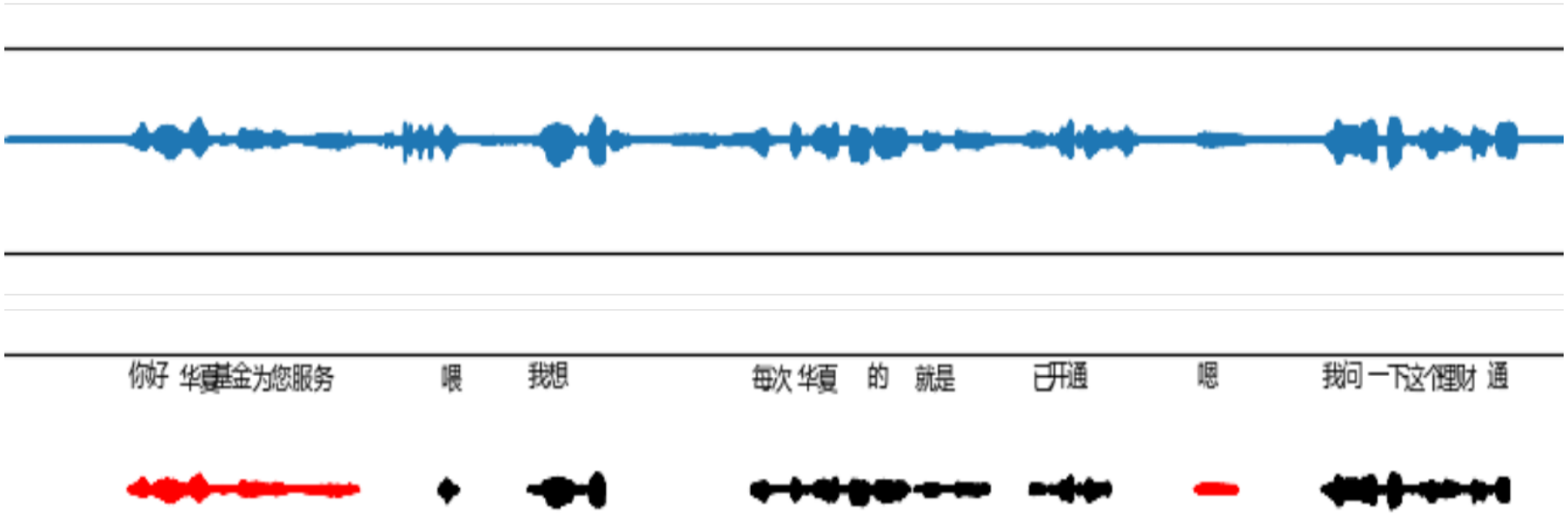
Lantian Li, Dong Wang, Yixiang Chen, Ying Shi, Zhiyuan Tang, "DEEP FACTORIZATION FOR SPEECH SIGNAL", ICASSP 2018

Truly factorized

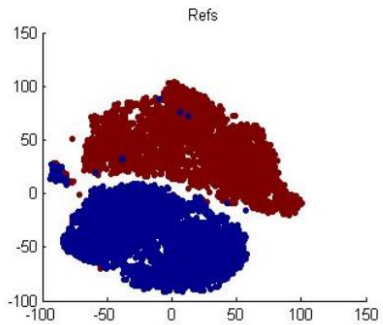


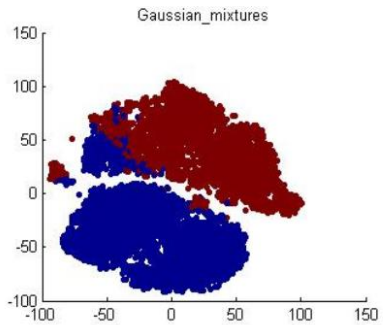
Lantian Li, Dong Wang, Yixiang Chen, Ying Shi, Zhiyuan Tang, "DEEP FACTORIZATION FOR SPEECH SIGNAL", ICASSP 2018

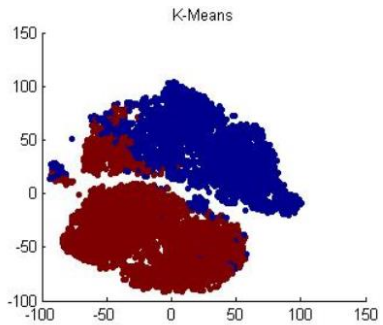
Segmentation

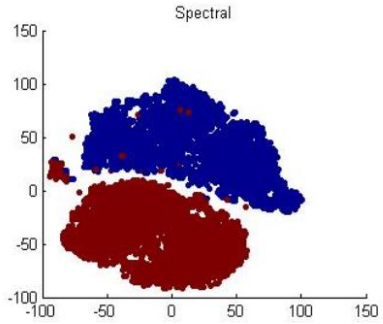


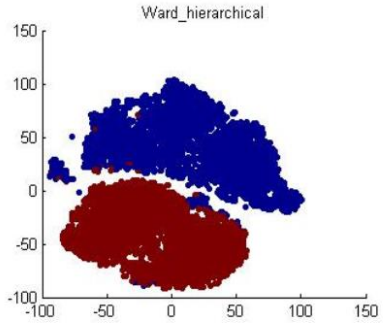
Human-music classification

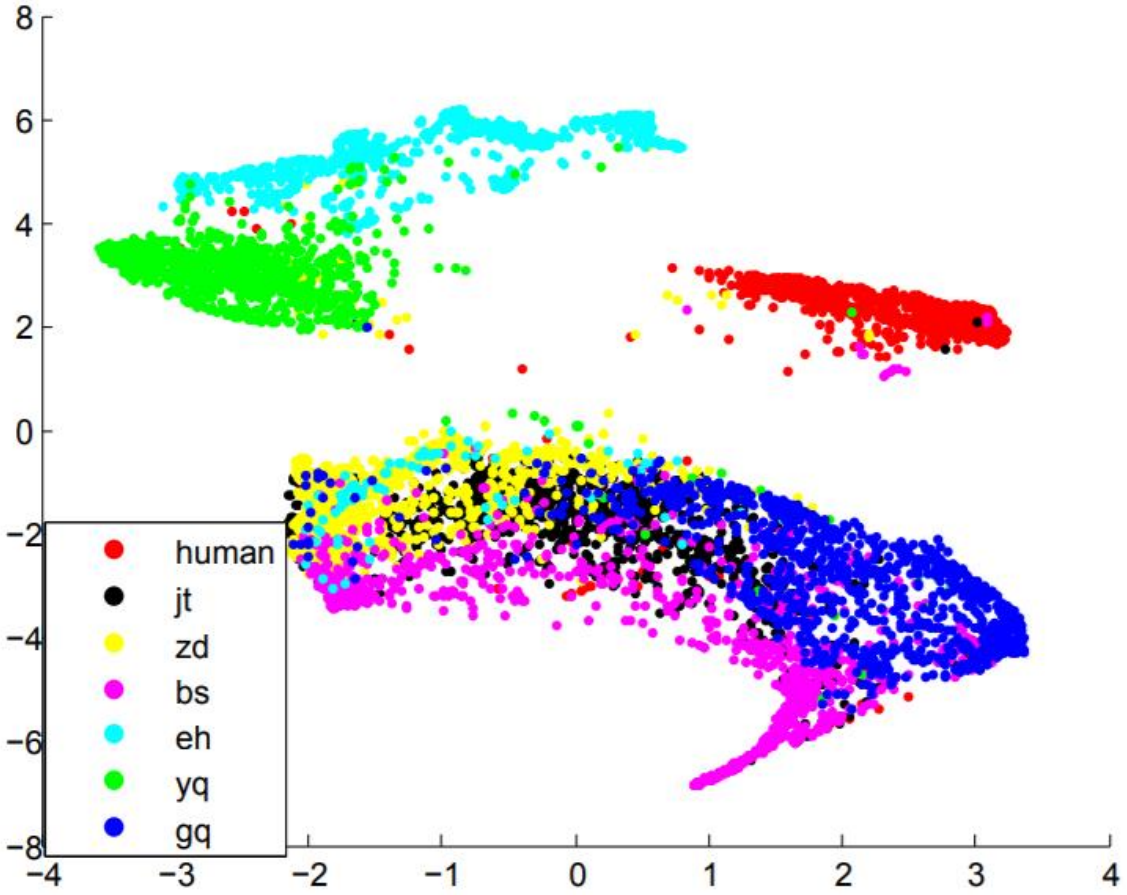




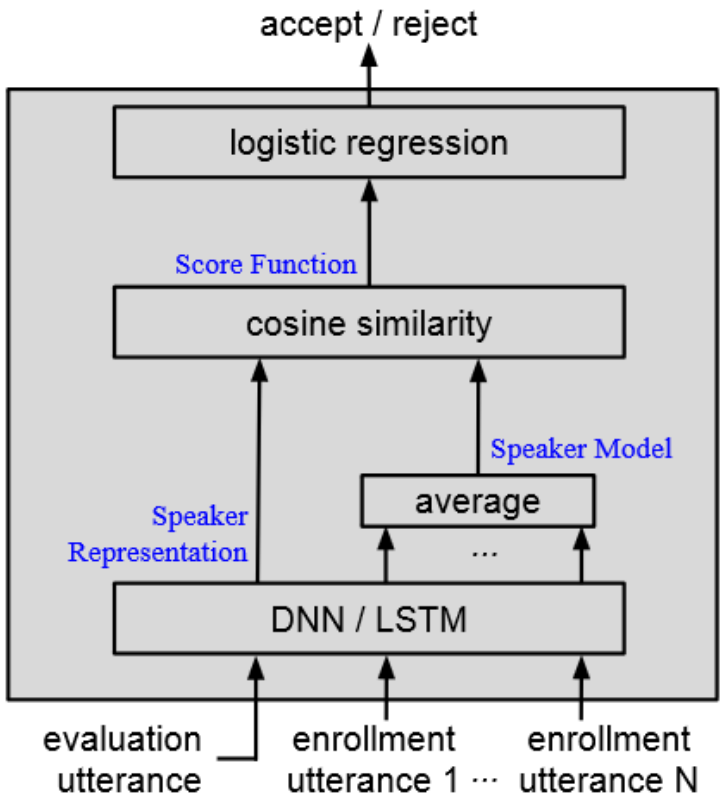




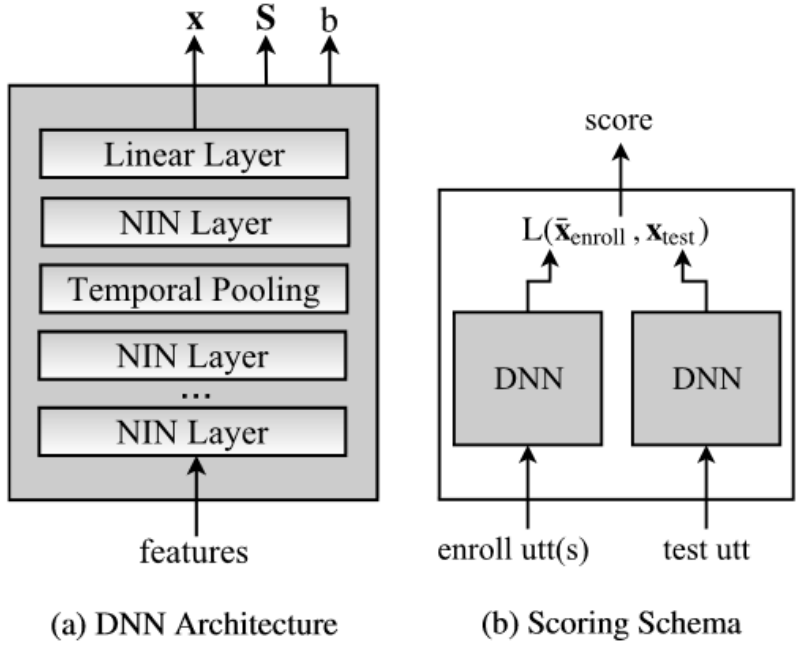




Compare to End-to-end learning



G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in Acoustics, Speech and Signal Pro-cessing (ICASSP), 2016

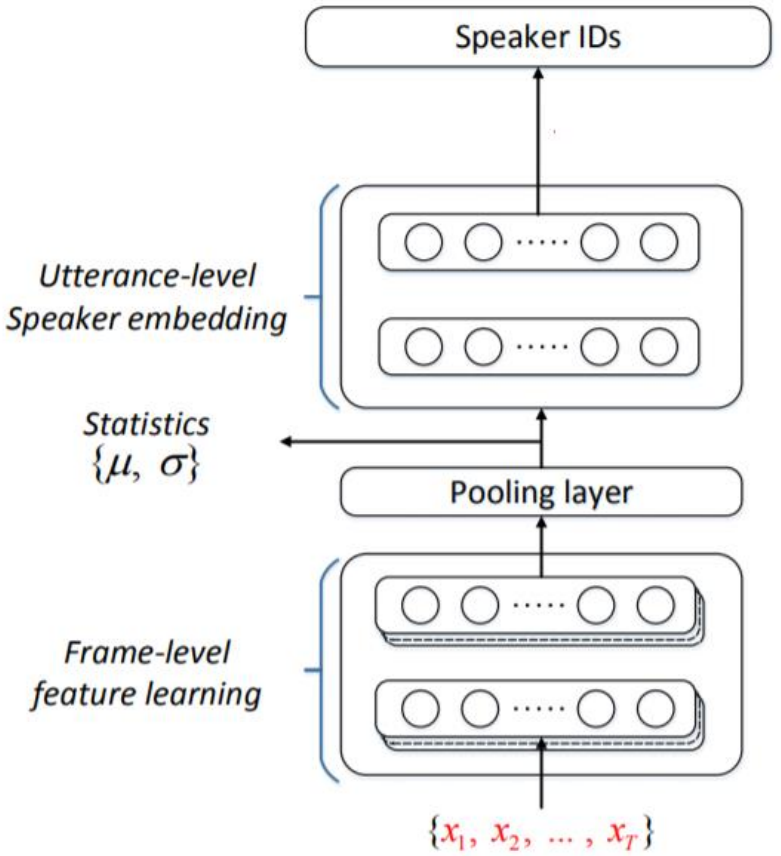


(a) DNN Architecture (b) Scoring Schema

D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in SLT'2016

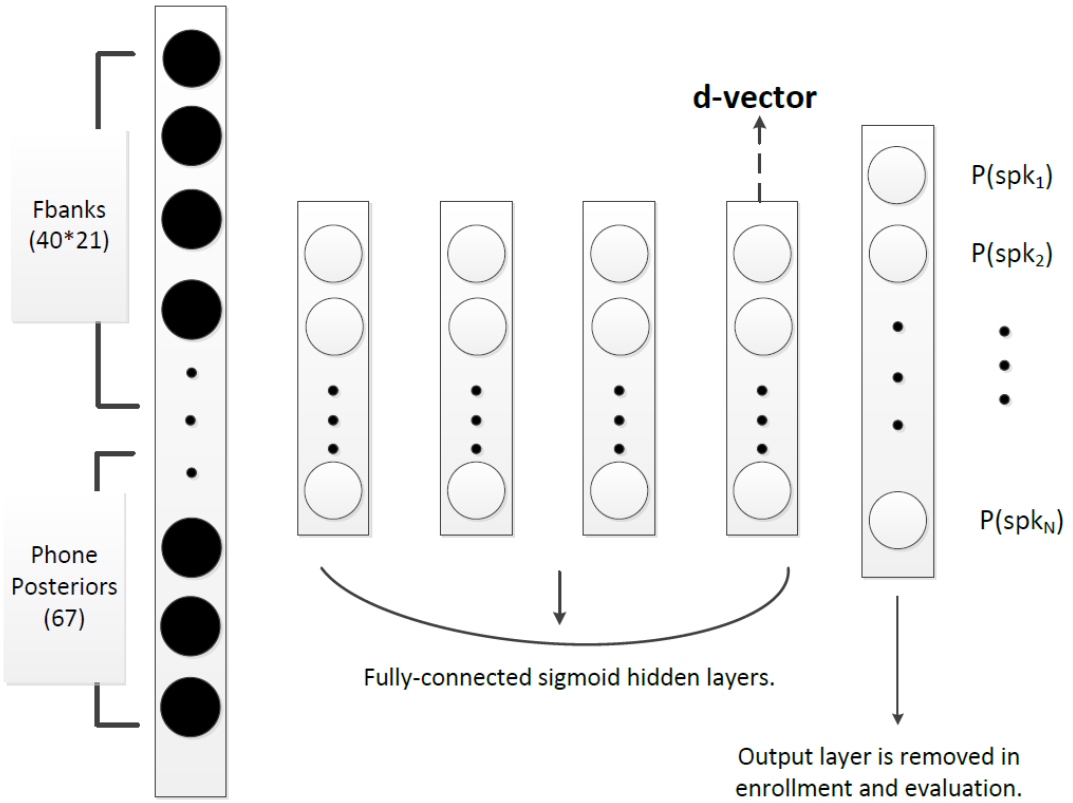
Compared to x-vector

- A trade-off between feature learning and end-to-end
- Specifically good for speaker recognition
- New model/architecture for speaker embedding



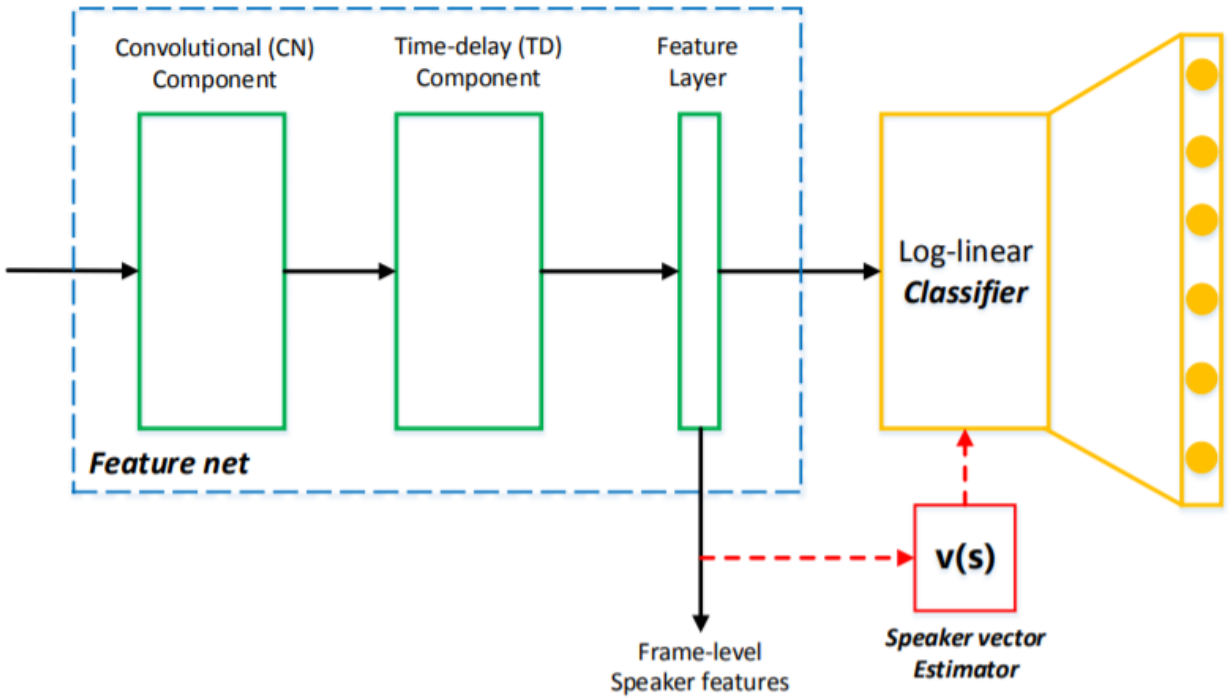
D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Recent advance: phone-aware training



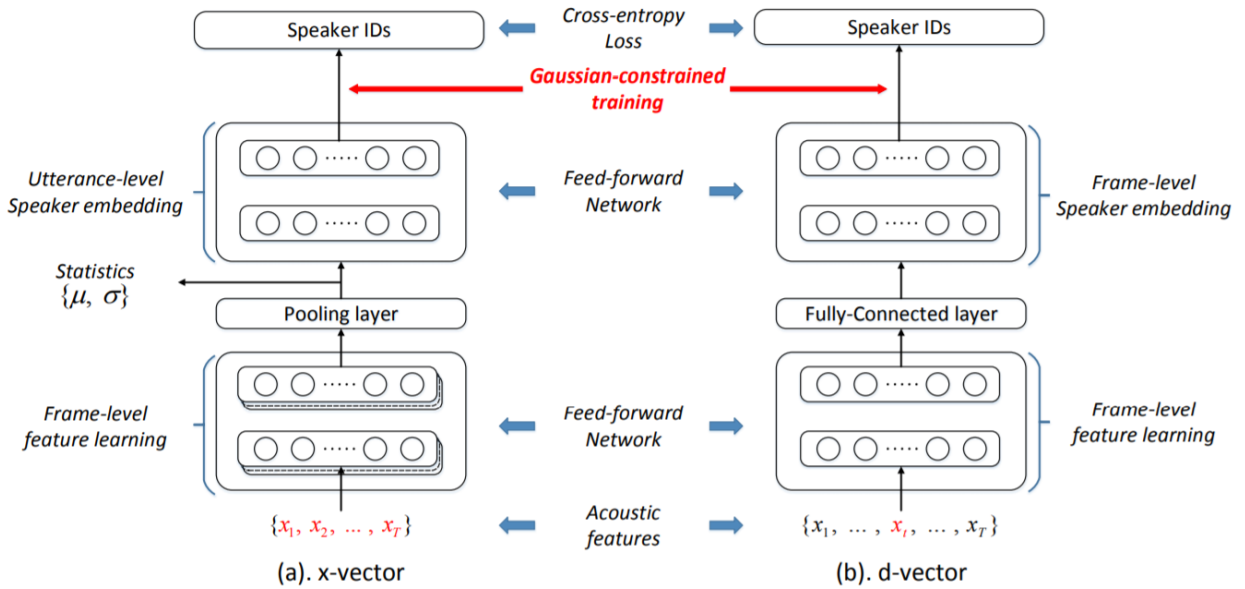
Lantian Li, Yiye Lin, Zhiyong Zhang, Dong Wang, "Improved Deep Speaker Feature Learning for Text-Dependent Speaker Recognition", APSIPA 2015

Recent advance :Full-info training



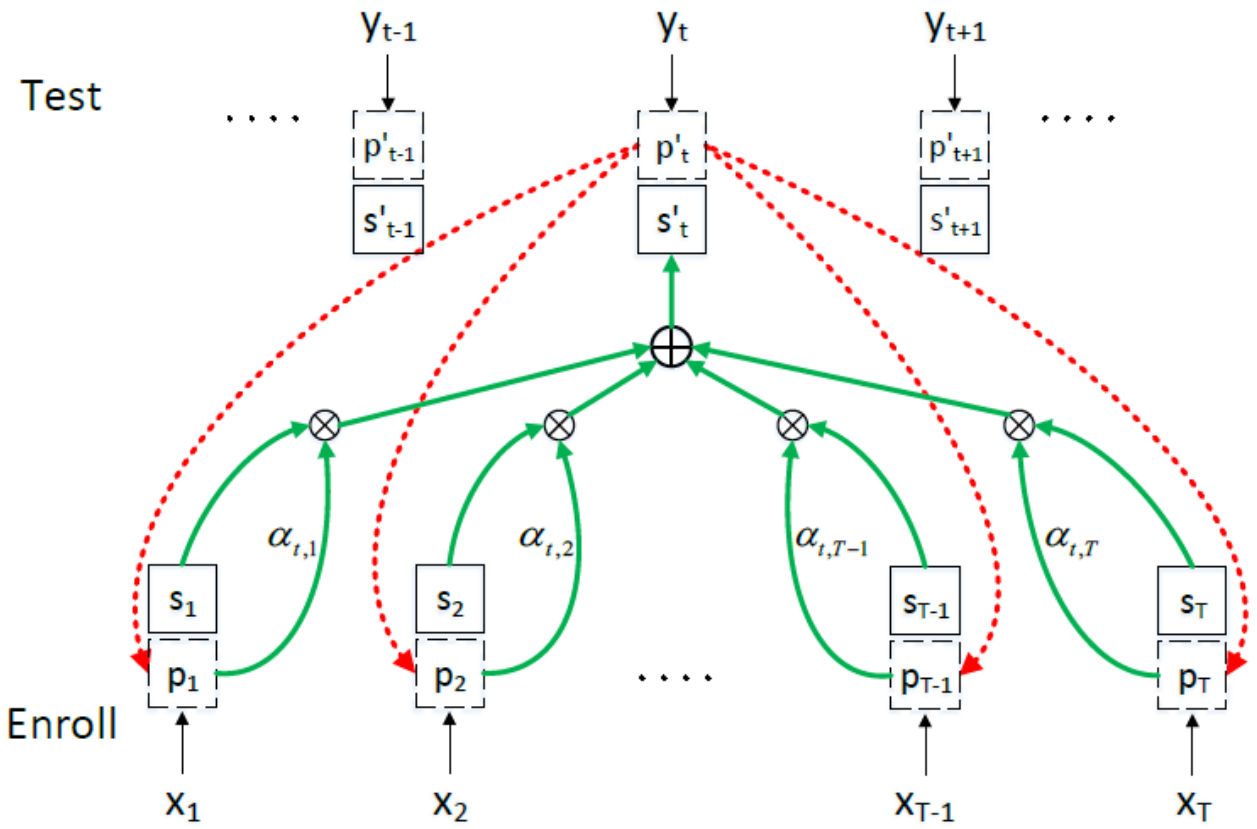
Lantian Li, Zhiyuan Tang, Dong Wang, "FULL-INFO TRAINING FOR DEEP SPEAKER FEATURE LEARNING", ICASSP 2018.

Recent advance :Gaussian constrained learning

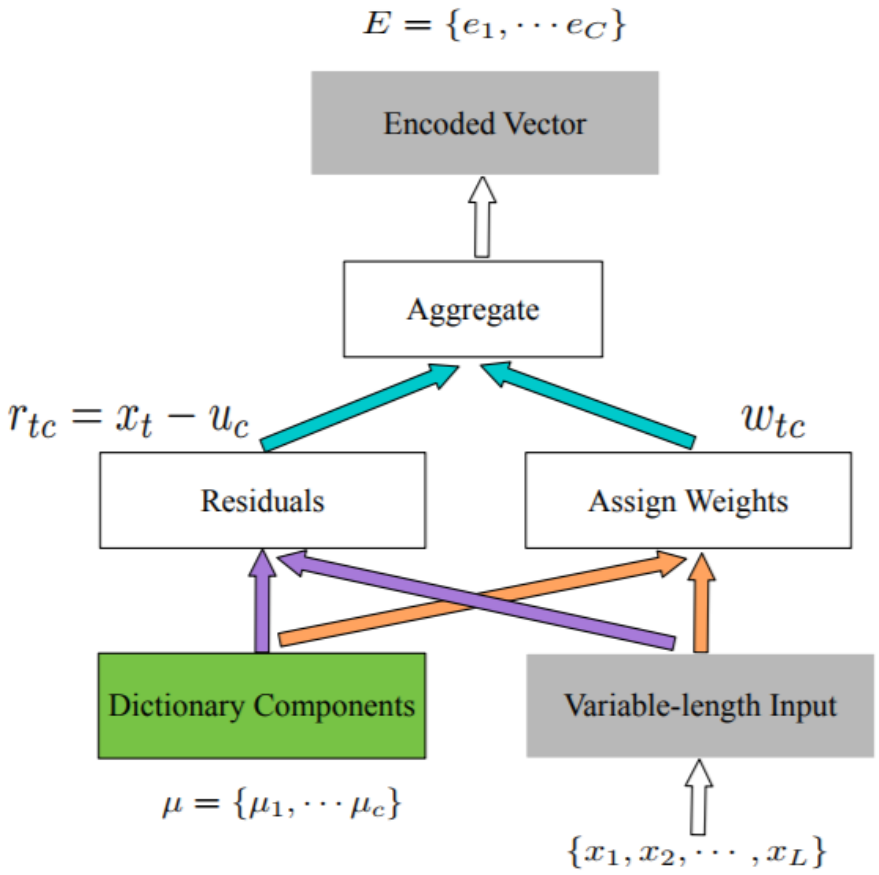


Lantian Li,Zhiyuan Tang,Ying Shi,Dong Wang, "Gaussian-Constrained Training for Speaker Verification", ICASSP 2019

Recent advance: Phonetic attention

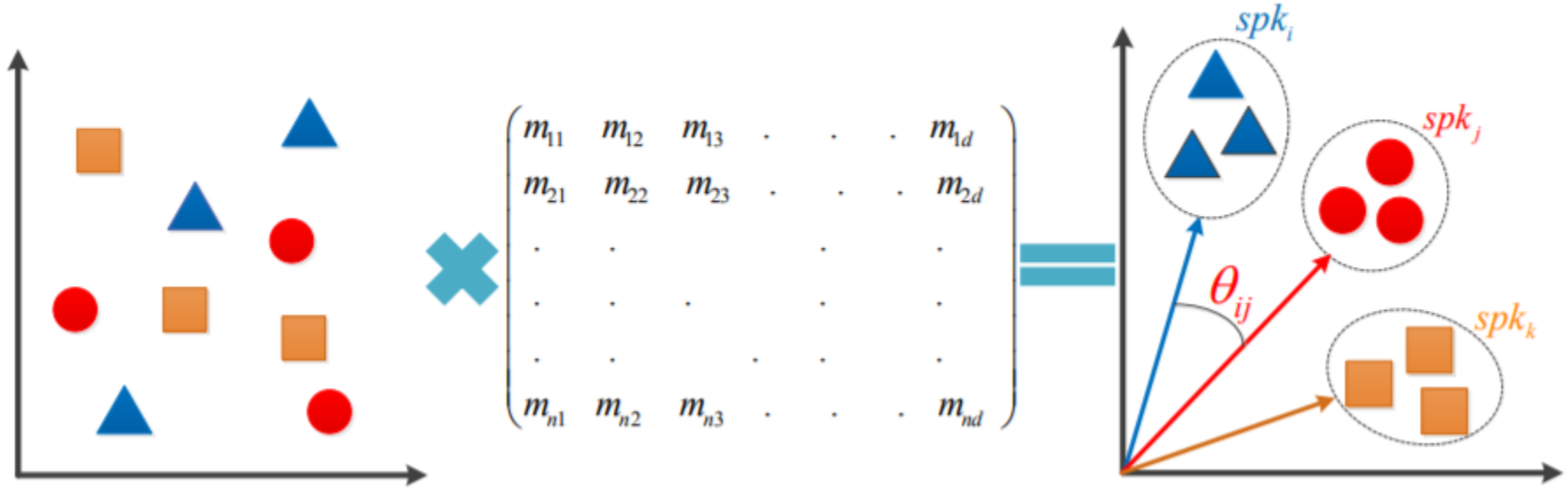


Recent advance: dictionary learning



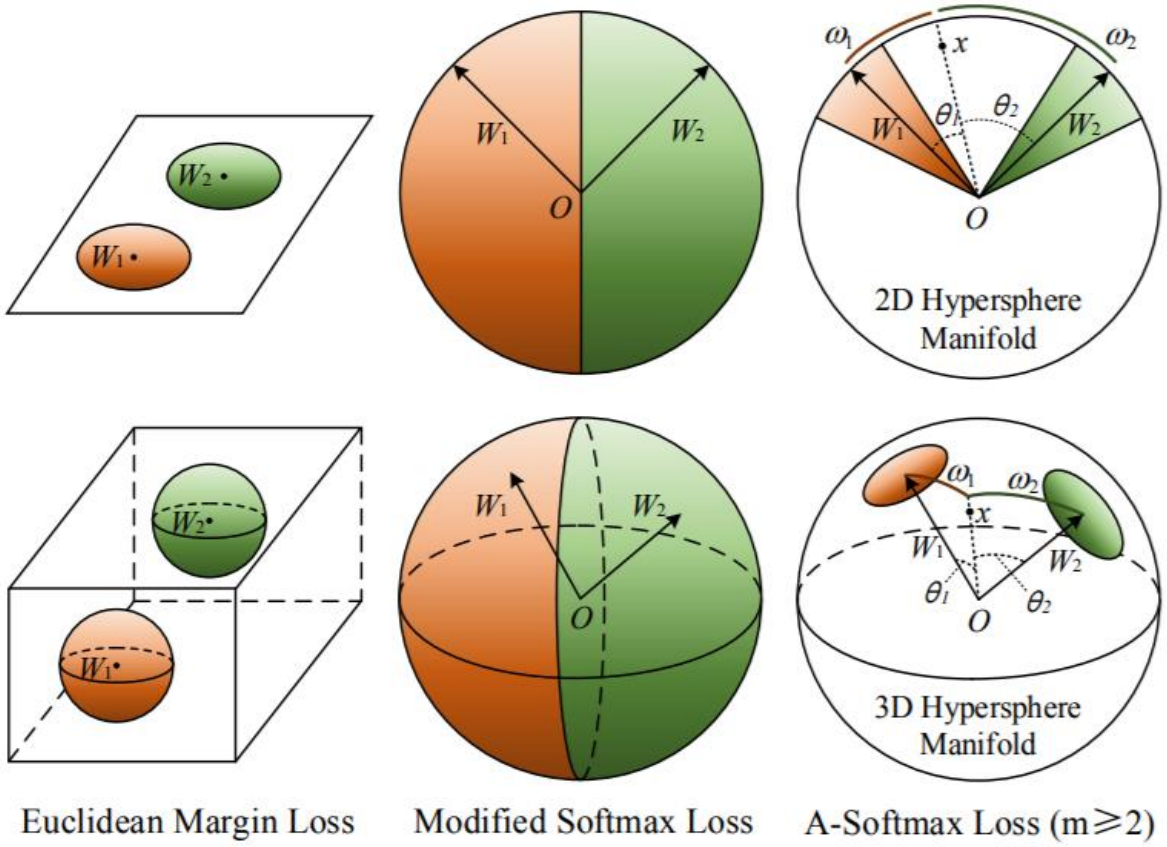
Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System, Weicheng Cai, Jinkun Chen, Ming Li, Odyssey, 2018.

Recent advance : Max margin



Lantian Li, Dong Wang, Thomoas Fang Zheng, "Max-Margin Metric Learning for Speaker Recognition", ISCSLP 2016

Recent advance: Angle loss



Loss Function	Decision Boundary
Softmax Loss	$(\boldsymbol{W}_1-\boldsymbol{W}_2)\boldsymbol{x}+b_1-b_2=0$
Modified Softmax Loss	$\|\boldsymbol{x}\|(\cos\theta_1-\cos\theta_2)=0$
A-Softmax Loss	$\|\boldsymbol{x}\|(\cos m\theta_1-\cos\theta_2)=0$ for class 1 $\|\boldsymbol{x}\|(\cos\theta_1-\cos m\theta_2)=0$ for class 2

- W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017
- Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System, Weicheng Cai, Jinkun Chen, Ming Li, Odyssey, 2018.

Conclusions for part I:

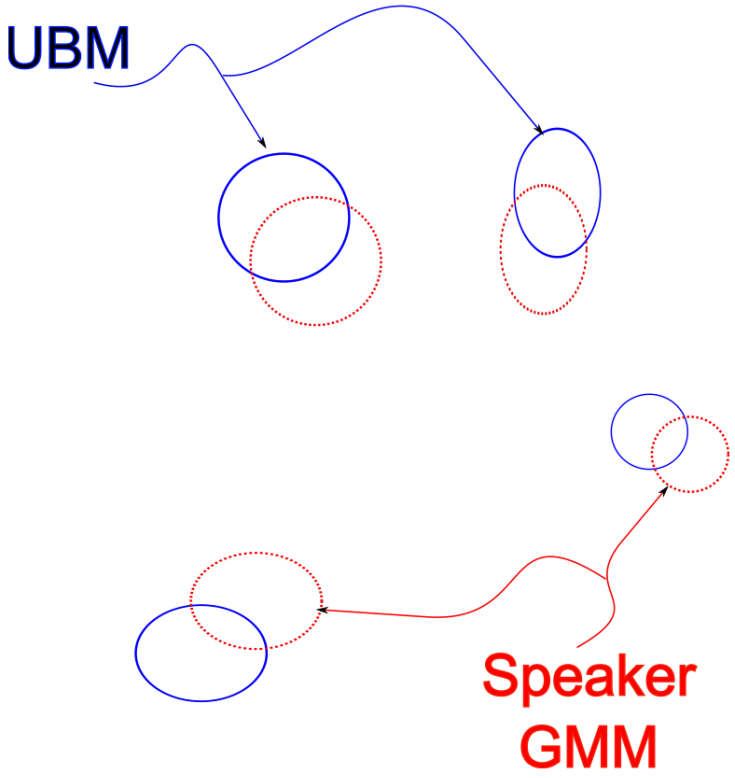
- Model-based won feature-based approach in history
- Deep learning learns short-term frame-based foundamental features
- The learned features can do many interesting things~

Feature/Embedding Normalization

Motivation

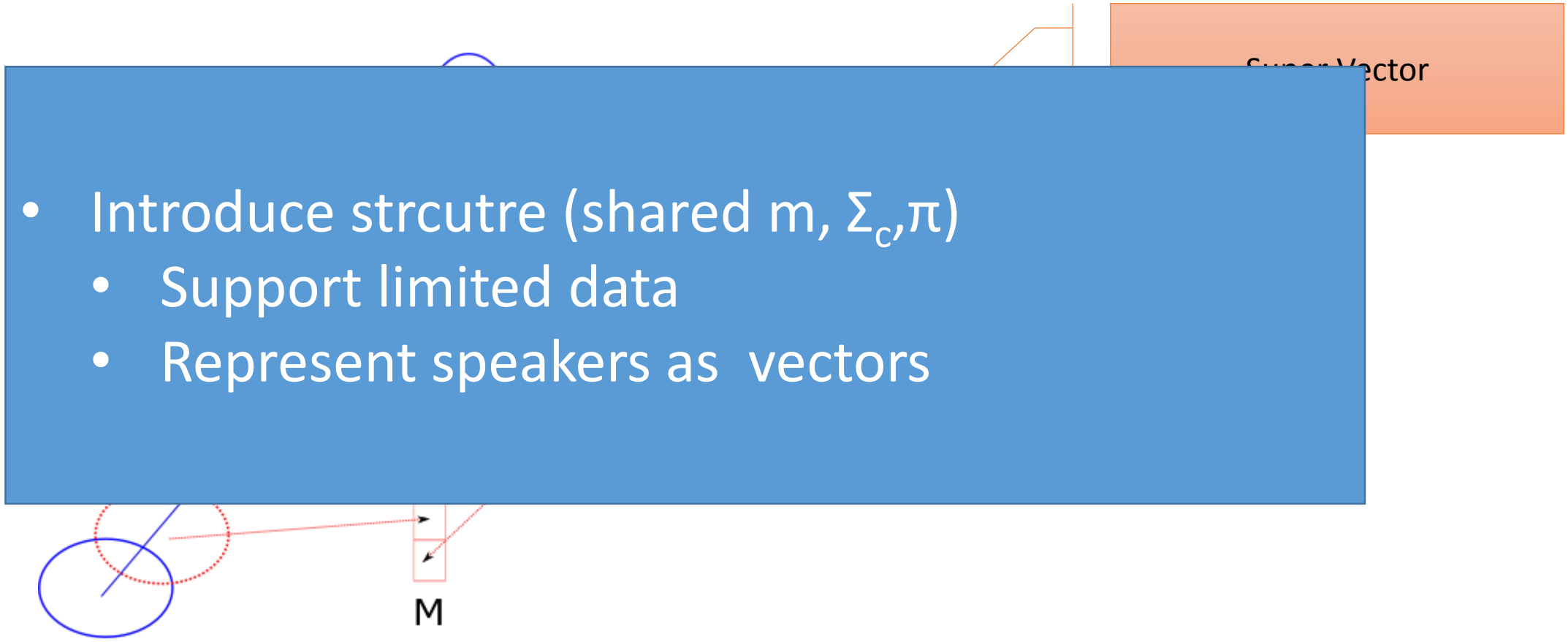
- We have (partly) solved the problem of learning speaker features
- Now we move to how to use them
  - Right now, they are mostly used as usually features
  - For frame-based, stack to utterance-based
  - For utterance-based, treated as i-vector and employ LDA/PLDA.
  - But are these correct and optimal?

Starting from GMM-UBM



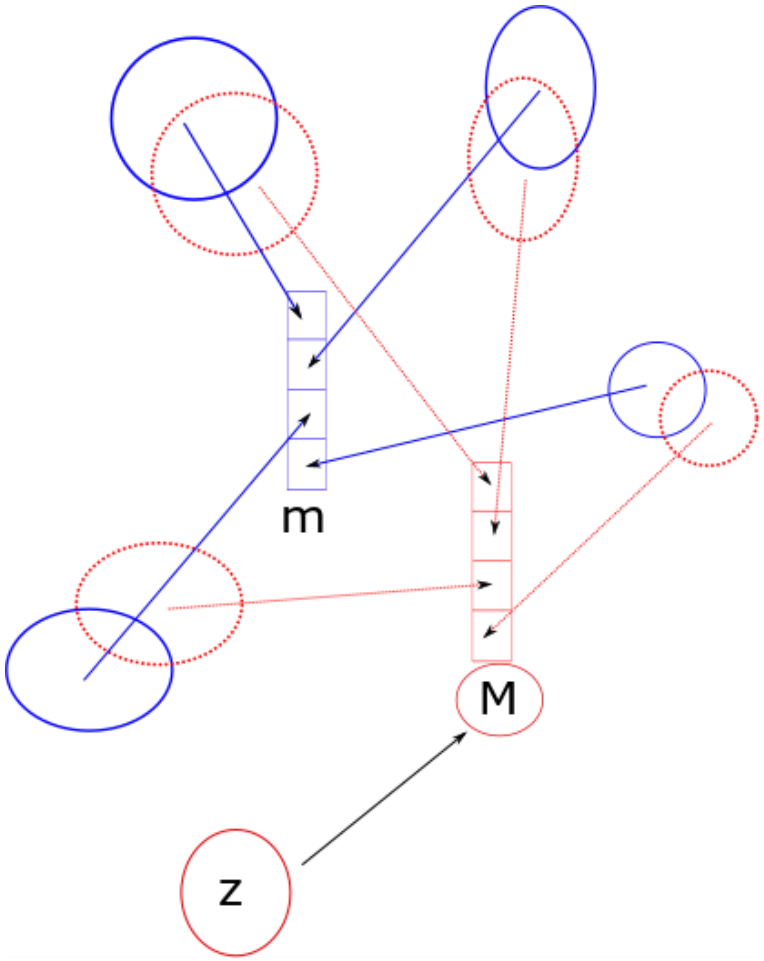
D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” Digital signal processing, vol. 10, no. 1-3, pp. 19–41, 2000.

It is a generative model



- Introduce strcutre (shared m, Σ_c,π)
  - Support limited data
  - Represent speakers as vectors

Factorization view

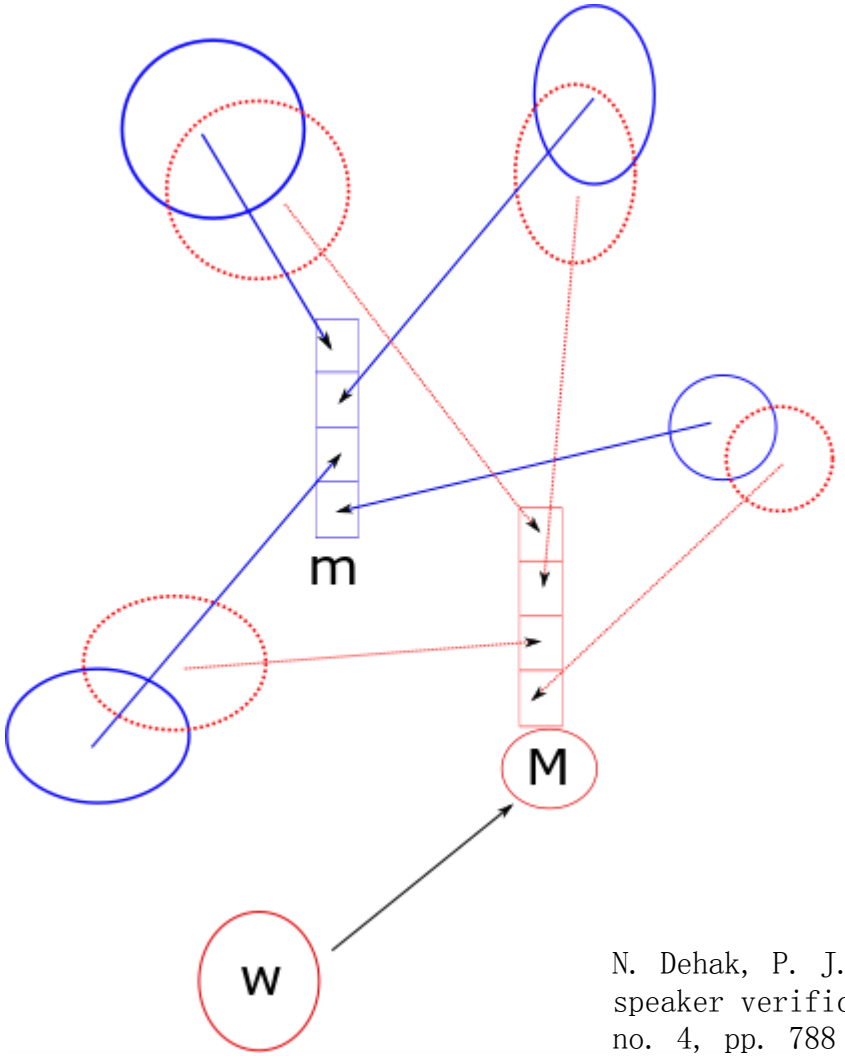


Embedding!!

$$x_i = m_c + Dzc + \varepsilon_c$$

$z = N(0,1)$; $\varepsilon_c = N(0, \Sigma_c)$; $c = \text{Multi}(\pi)$

i-vector: More structured factorization



$$x_i = m_c + [Tw]_c + \varepsilon_c$$

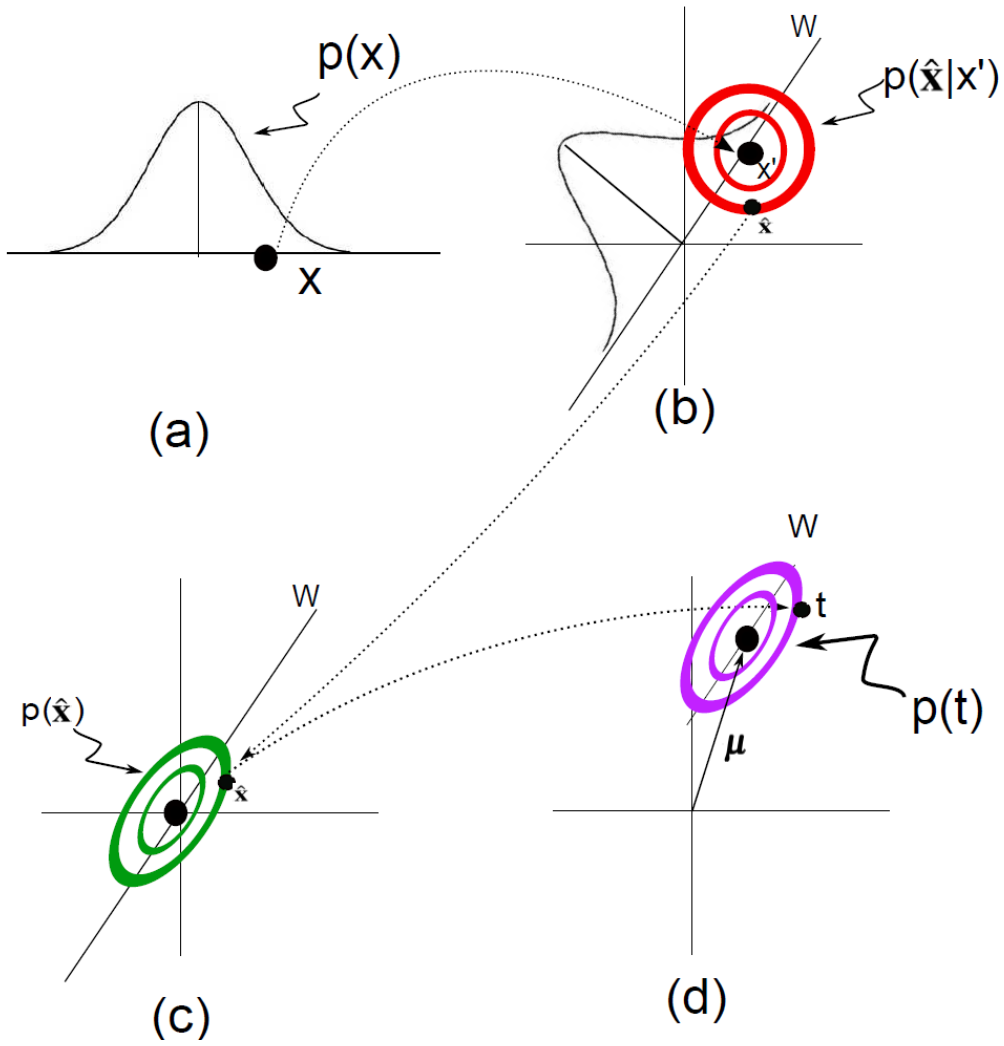
$w = N(0,1)$; $\varepsilon_c = N(0,\Sigma_c)$; $c = \text{Multi}(\pi)$

- Embedding!
  - Low dimensional
  - Component dependent

N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788–798, 2011.

Key properties

- Generative model, relying on Bayesian inference
- Pseduo-Linear Gaussian
- Two layers (shallow)
- Extended PPCA
- Weakly discriminative



Improving discrimination

- WCCN
- LDA
  - Partly generative
  - Shared-variance Gaussian
  - Mean as parameters
- PLDA
  - Fully generative
  - Shared-variance Gaussian
  - Mean as Gaussian variables

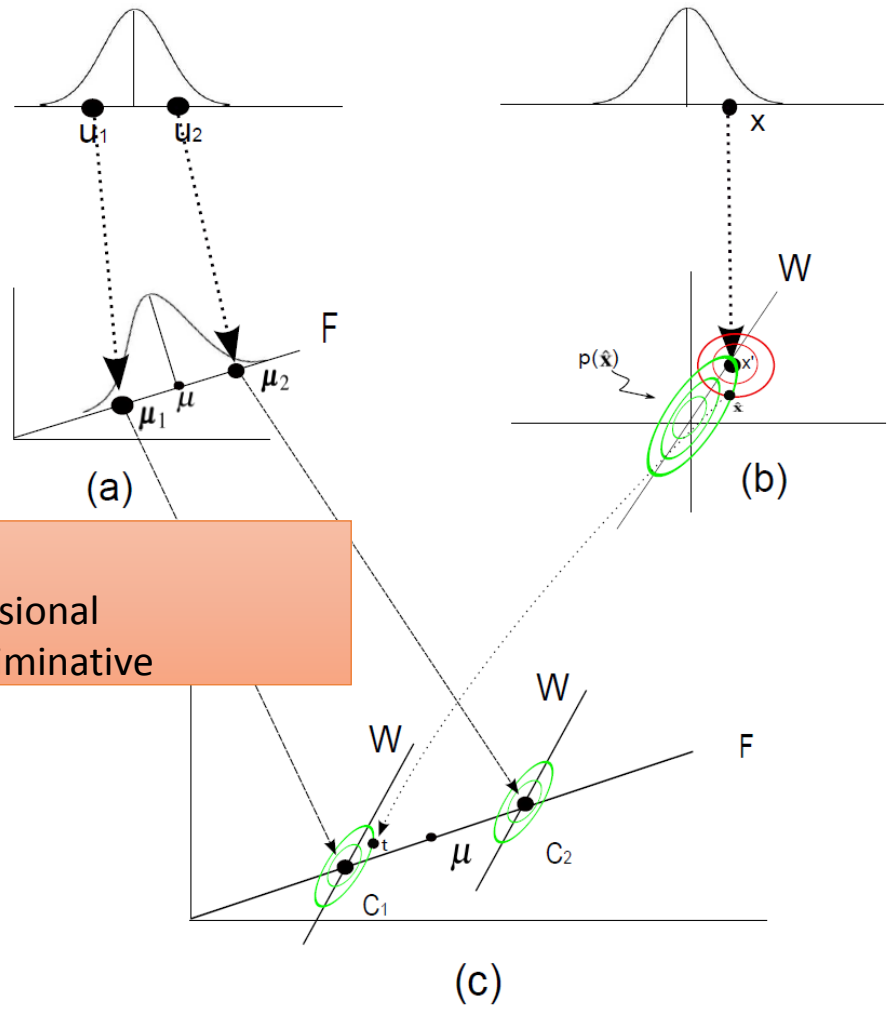
PLDA

- Linear Gaussian
- Generative model, but discriminatively trained
- Discriminative decision by Bayesian rule

- Embedding!
  - Low dimensional
  - More discriminative

$$\mathbf{t} = \boldsymbol{\mu} + \mathbf{F}\mathbf{u} + \mathbf{W}\mathbf{x} + \boldsymbol{\varepsilon}$$

$$\mathbf{u} \sim N(\mathbf{0}, \mathbf{I}), \quad \mathbf{x} \sim N(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}).$$

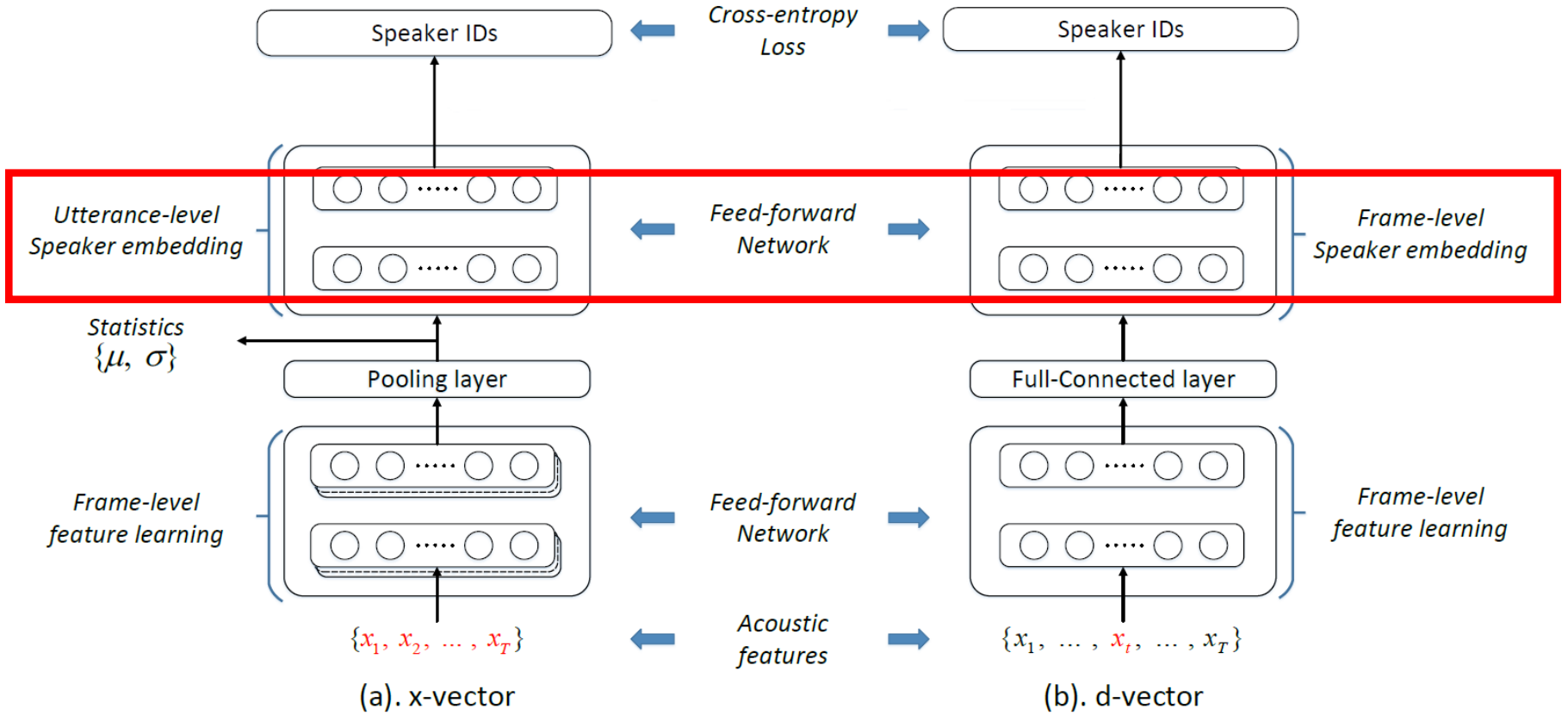


S. Ioffe, "Probabilistic linear discriminant analysis," Computer Vision–ECCV 2006, pp. 531–542, 2006.

i-vector and PLDA is consistent

- PLDA assumption
  - Gaussian prior
  - Gaussain conditional
  - Hence Gaussian marginal
- i-vectors are mostly Gaussian

Neural-based embedding



E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint textdependent speaker verification," in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014, pp. 4052－4056.

D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

Properties of neural embeddings

- Inferred from discriminative models (differnet from i-vectors)
- Less probabilistic meaning (different from i-vectors)
- Highly discriminative (different from i-vectors)

An interesting observation

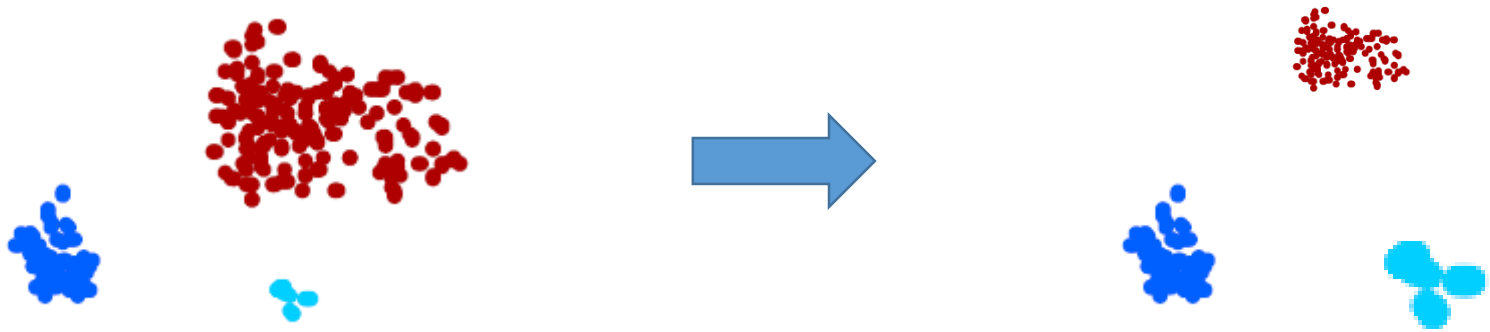
	Cosine EER	PLDA EER	LDA PLDA
d-vector	38.39%	17.71%	9.511%
x-vector	15.67%	9.087%	3.157%
d-vector-LDA	12.94%	9.665%	9.511%
x-vector-LDA	5.198%	3.735%	3.157%

- Why LDA works?
- Why PLDA works?
- Why LDA+PLDA works?

Why discriminative embeddings need discriminative back-end?

Because of normalization…

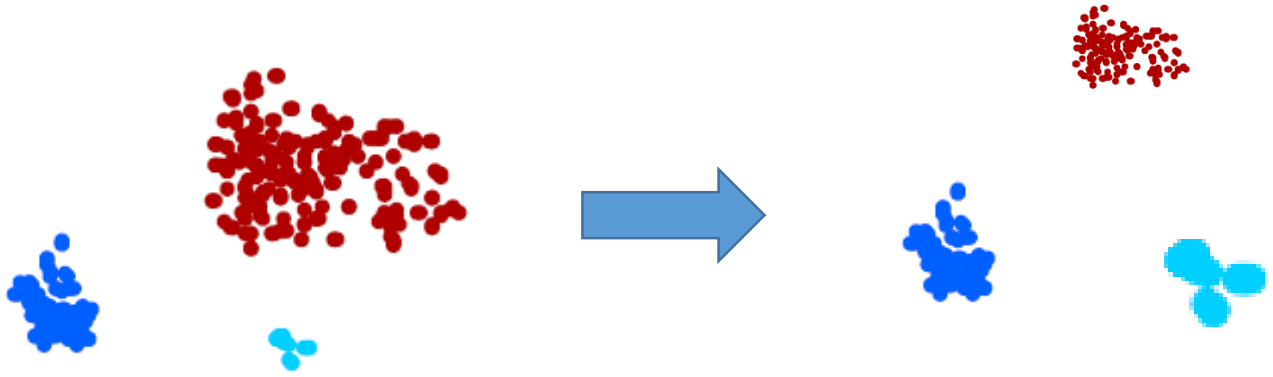
- Normalization
  - Different speaker embeddings should have identical covariance(WCCN)
  - Different speaker scores (own, imposter) should have identical variance (ZT norm)
- Normalization is important for generalization
- Normalization is important for thresholding



Review LDA/PLDA

- LDA
  - Partly generative
  - Shared-variance Gaussian
  - Mean as parameters
- PLDA
  - Fully generative
  - Shared-variance Gaussian
  - Mean as Gaussian variables

- The assumptions of these models regularize the embeddings, hence scores….



Therefore…

- LDA works
- PLDA works

	Cosine EER	PLDA EER	LDA PLDA
d-vector	38.39%	17.71%	9.511%
x-vector	15.67%	9.087%	3.157%
d-vector-LDA	12.94%	9.665%	9.511%
x-vector-LDA	5.198%	3.735%	3.157%

But why LDA+PLDA works?

- PLDA does not only normalize per se, but also requires normalization.
- Prior is Gaussian, conditional is Gaussian, and marginal is Gaussian.
- LDA thus helps PLDA

$$\mathbf{t} = \boldsymbol{\mu} + \mathbf{F}\mathbf{u} + \mathbf{W}\mathbf{x} + \boldsymbol{\varepsilon}$$

$$\mathbf{u} \sim N(\mathbf{0}, \mathbf{I}), \quad \mathbf{x} \sim N(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}).$$

	Cosine EER	PLDA EER	LDA PLDA
d-vector	38.39%	17.71%	9.511%
x-vector	15.67%	9.087%	3.157%
d-vector-LDA	12.94%	9.665%	9.511%
x-vector-LDA	5.198%	3.735%	3.157%

Normalization test

- Skew and Kurt

	Skew	Kurt
d-vector	0.1092	-0.11609
x-vector	-0.04228	-0.3603
d-vector-LDA	-0.005515	0.028395
x-vector-LDA	-0.01074	-0.0089

$$Skew(x) = \frac{E[(x-\mu_x)^3]}{\sigma^3} \quad Kurt(x) = \frac{E[x-\mu_x]^4}{\sigma_x^4} - 3$$

Why LDA+PLDA works?

- LDA makes the conditional embeddings more Gaussian, hence suitable for PLDA.

PCA also works

- LDA regurlize conditional distribution
- PCA regularize marginal distribution

	Cosine	PCA	PLDA	L-PLDA	P-PLDA
x-vector	15.67	16.17	9.09	**3.12**	4.16

LDA/PCA does not work for ivector+PLDA

- i-vector is Gaussian constrained (marginally)

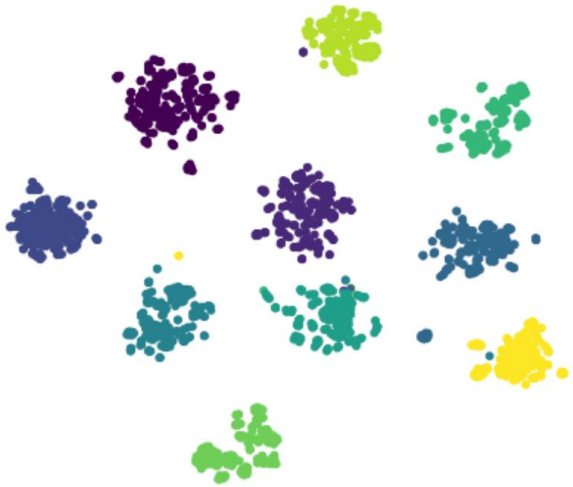
	Cosine	LDA	PLDA	LDA+PLDA	PCA+PLDA
i-vector	3.744	4.032	3.6	3.672	4.536
x-vector	5.256	4.104	4.032	4.004	3.888

Quick summary

- i-vector is probabilistic embedding, and d/x vector is neural embedding.
- i-vector is regularized but not discriminative, and d/x vector is the opposite.
- PLDA works in both i-vector and d/x vectors, but perform differently: former is discrimination, latter is normalization.
- PCA and LDA help PLDA, by providing normalized vectors: former is via marginalized Gaussian, latter is via conditional Gaussian.

Problem of PCA/LDA normalization

- PLDA requires prior and conditional to be Guassian; neither PCA nor LDA matches all.
- Linear shallow models cannot derive Gaussian prior/conditional with complex observed marginal and observed conditional of d/x vectors.



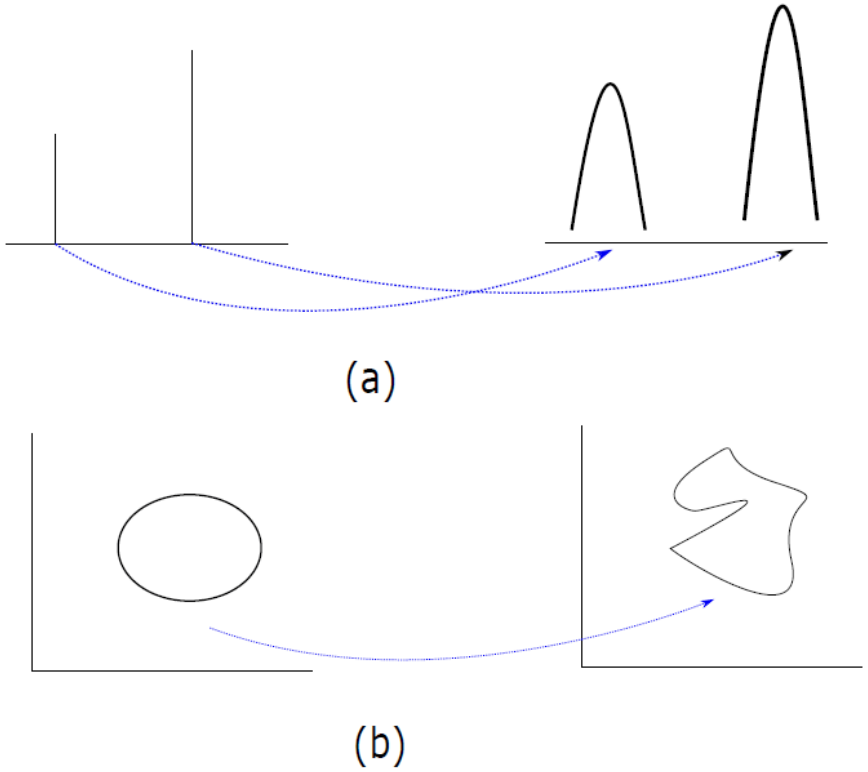
Marginal

Conditional

Moving to distribution mapping

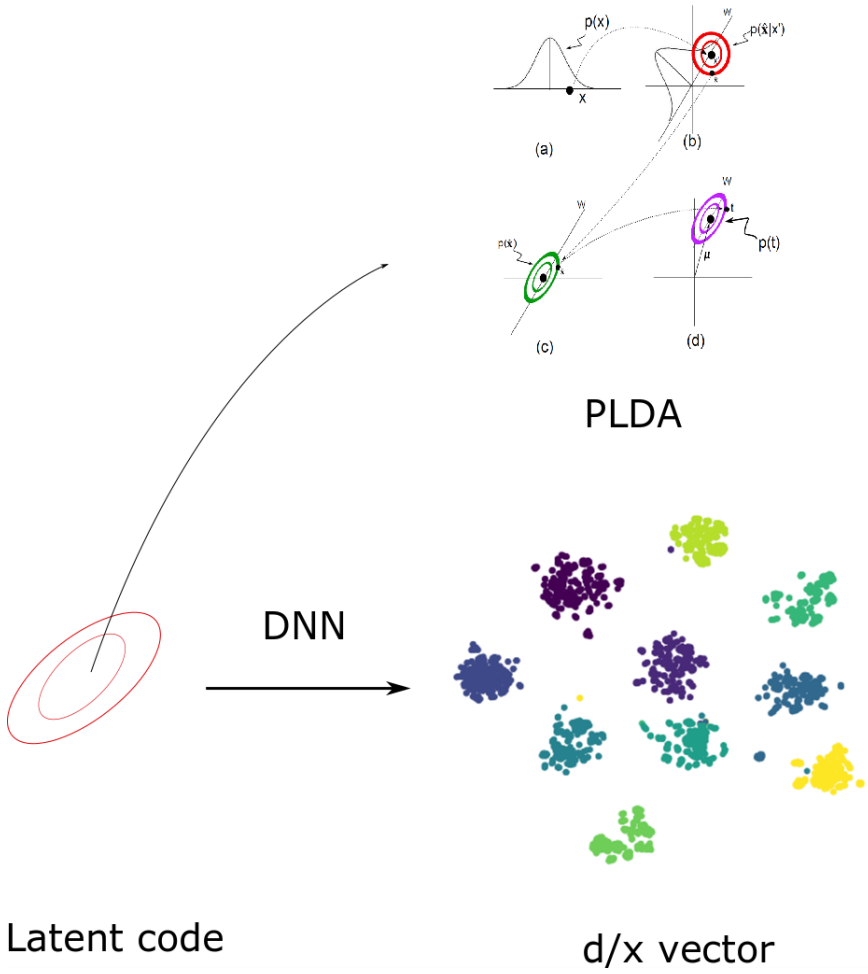
- A complex distribution can be generated from a simple distribution with a complex transforming.

$$\log p(x) = \log p(z) + \log \left| \det \frac{\mathrm{d} f^{-1}(x)}{\mathrm{d} x} \right|$$



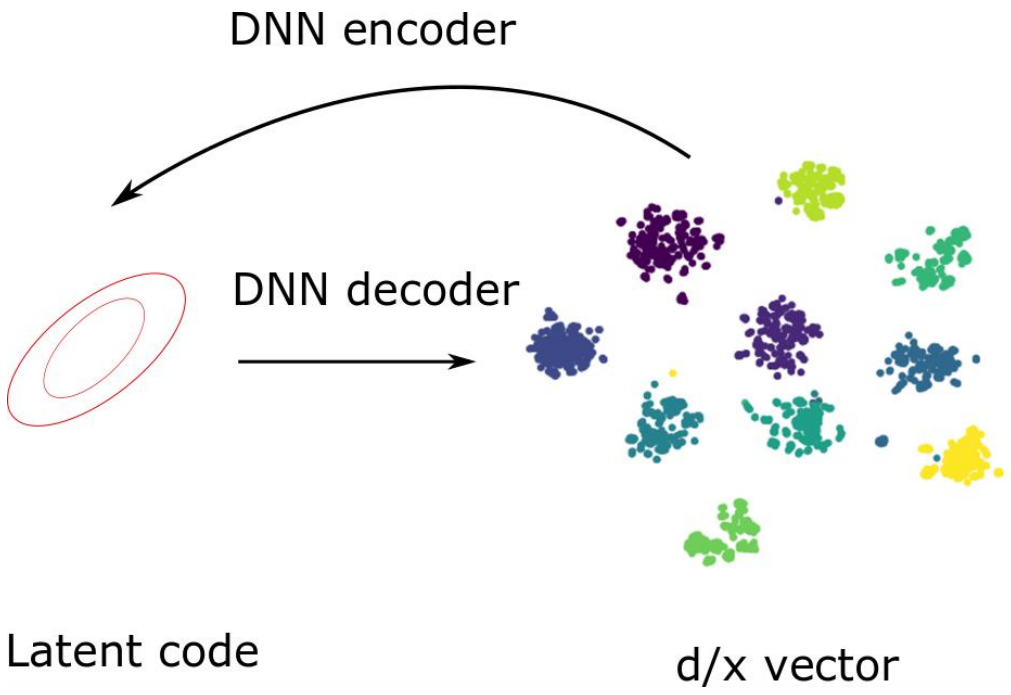
We therefore hope a deep generative model

- That can use Gaussian latent code to generate complex d/x vectors.
- The latent code will be used as normlized vectors.
- The noramlized vectors will be more PLDA ameable.



But how do we genreate the code?

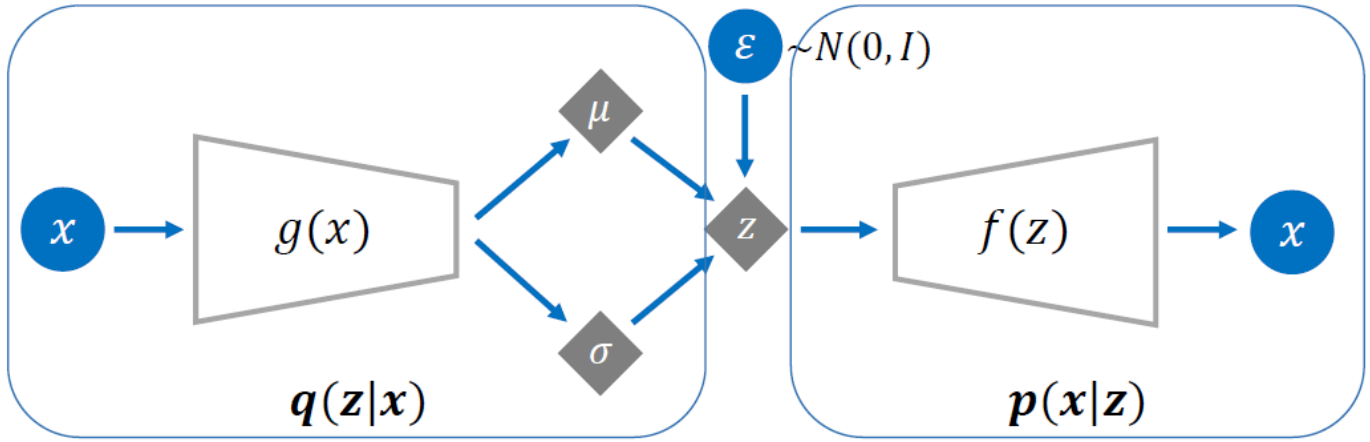
- A wake/sleep game.
- A stochastic VB approach for approximation.
- VAE architecture.



Hinton G E, Dayan P, Frey B J, et al. The" wake-sleep" algorithm for unsupervised neural networks[J]. Science, 1995, 268(5214): 1158-1161.
D. P. Kingma and M.Welling, “Auto-encoding variational bayes,” arXiv preprint arXiv:1312.6114, 2013.

VAE Architecture

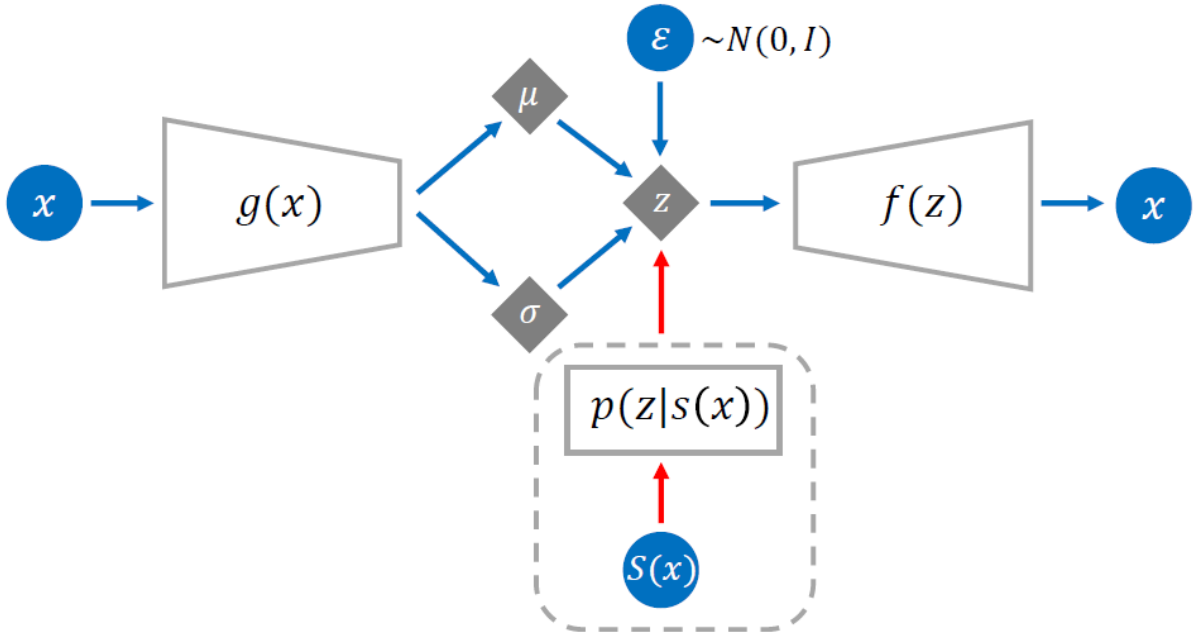
- Roughly regularize **marginal distribution** as Gaussian.
- Deal with complex observed marginal.
- Extended pesudo-VAE.



$$L(f,g) = \sum_i \{-D_{KL}[q(z|x_i)||p(z)] + \mathbb{E}_{q(z|x_i)}[\ln p(x_i|z)]\}$$

Further constraine conditional

- Conhensive loss, like central loss and Gaussian-constrained training.



$$L_C(f,g) = \sum_i \ln p(\mu(x)|s(x)) = \sum_i \ln N(\mu(x); s(x), I)$$

- W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," in Proc. Odyssey 2018 The Speaker and Language Recognition Workshop, 2018, pp. 74–81.
- L. Li, Z. Tang, Y. Shi, and D. Wang, "Gaussian-constrained training for speaker verification," in ICASSP, 2019.

SITW test

- X-vector: baseline
- V-vector:VAE-regularized
- C-vector: with cohensive constrained
- A-vector: AE-regularized (VAE without KL constrained to Gaussian, without hidden sampling)

SITW test

SITW Eval. Core

	Cosine	PCA	PLDA	L-PLDA	P-PLDA
x-vector	16.79	17.22	9.16	3.80	4.84
a-vector	16.05	16.81	12.14	4.27	5.09
v-vector	10.11	10.03	**3.64**	3.64	4.43
c-vector	**9.05**	**8.83**	3.77	**3.53**	**4.10**

- V/C vector works even with Cosine, though PCA/a-vector does not. Means VAE with random sampling really important.
- V/C vector with cosine get similar performance as PLDA. They all do normalization!

SITW test

SITW Eval. Core

	Cosine	PCA	PLDA	L-PLDA	P-PLDA
x-vector	16.79	17.22	9.16	3.80	4.84
a-vector	16.05	16.81	12.14	4.27	5.09
v-vector	10.11	10.03	**3.64**	3.64	4.43
c-vector	**9.05**	**8.83**	3.77	**3.53**	**4.10**

- V-vector works for PLDA, better than P-PLDA(unsuperivsed), comparable with L-PLDA(supervised).
- C-vector works mostly better than v-vector; worse when helping PLDA (PLDA also supervised).
- C-vector plus LDA provides the best performance. Something complementary.

Normalization test

- V/C AE normalize both marginal and prior
- Regulization on marginal can transfer to prior!
- CVAE gives better marginal, but worse prior (strange).
- AE reduces Skew but increases Kurt.

	Skew(utt)	Kurt(utt)	Skew(spk)	Kurt(spk)
x-vector	-0.0423	-0.3604	0.0018	-0.4499
a-vector	-0.0072	-0.7740	0.0014	-0.9765
v-vector	-0.0055	0.1324	-0.0042	-0.0285
c-vector	-0.0043	0.1154	-0.0076	-0.0298

Test on a more realistic data

- Similar trend on SITW.
- V/C normalization is highly effective.
- V/C+PCA+PLDA performs the best.

	Cosine	PCA	PLDA	L-PLDA	P-PLDA
x-vector	16.65	16.89	16.91	15.39	13.29
v-vector	13.55	13.71	**12.46**	12.06	12.02
c-vector	**12.98**	**13.13**	12.48	**12.01**	**11.98**

Conclusions for part II

- VAE can describe complex d/x embeddings.
- VAE-based code is Gaussian-constrained in marginal.
- Cohensive-constrained VAE further constraines conditional.
- The constrained marginal and conditional leads to better regularized prior.
- The normalized embeddings perform better by themselves or with PLDA.

Wrap up

- Deep learning can discover fundamental features, either frame-based or utterance-based.
- Deep features should be accompanies with a careful design to ensure consistence with the back-end.

- Thanks!