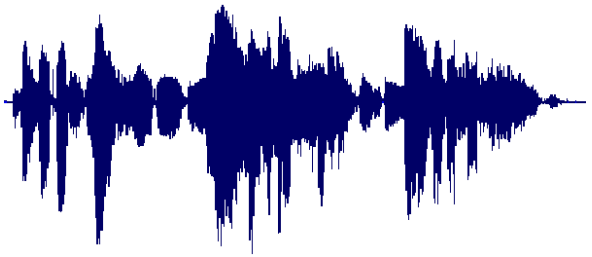
Deep Speech Factorization

Dong Wang

2017/06/05

Speech signal is a composition

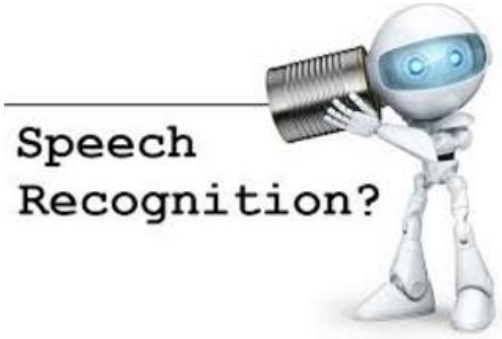
- You can tell multiple things from a waveform
  - Linguistic (what is pronounced)
  - Speaker (who pronounced)
  - Emotion (how pronounced)
  - Physical status



- Composition leads to uncertainty, the main challenge in speech processing
- This is not only for speech, but also for other signals

Deal with composition

- Discriminative training
  - Single-task oriented
  - Filtering interference driven by discriminative modeling, through deep learning
  - No information correlation utilized



Deal with composition

- Joint training
  - Multi-task feature learning
  - Conditional learning
  - Collaborative learning

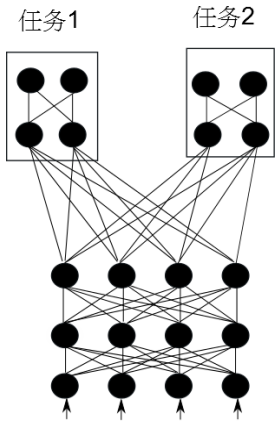


图2. 特征共享学习示意图。

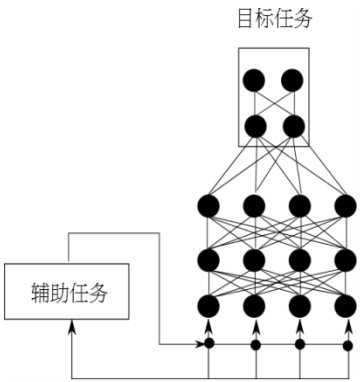


图3. 条件学习示意图。

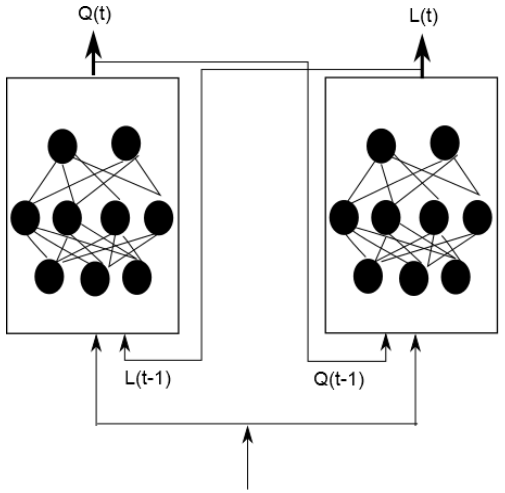
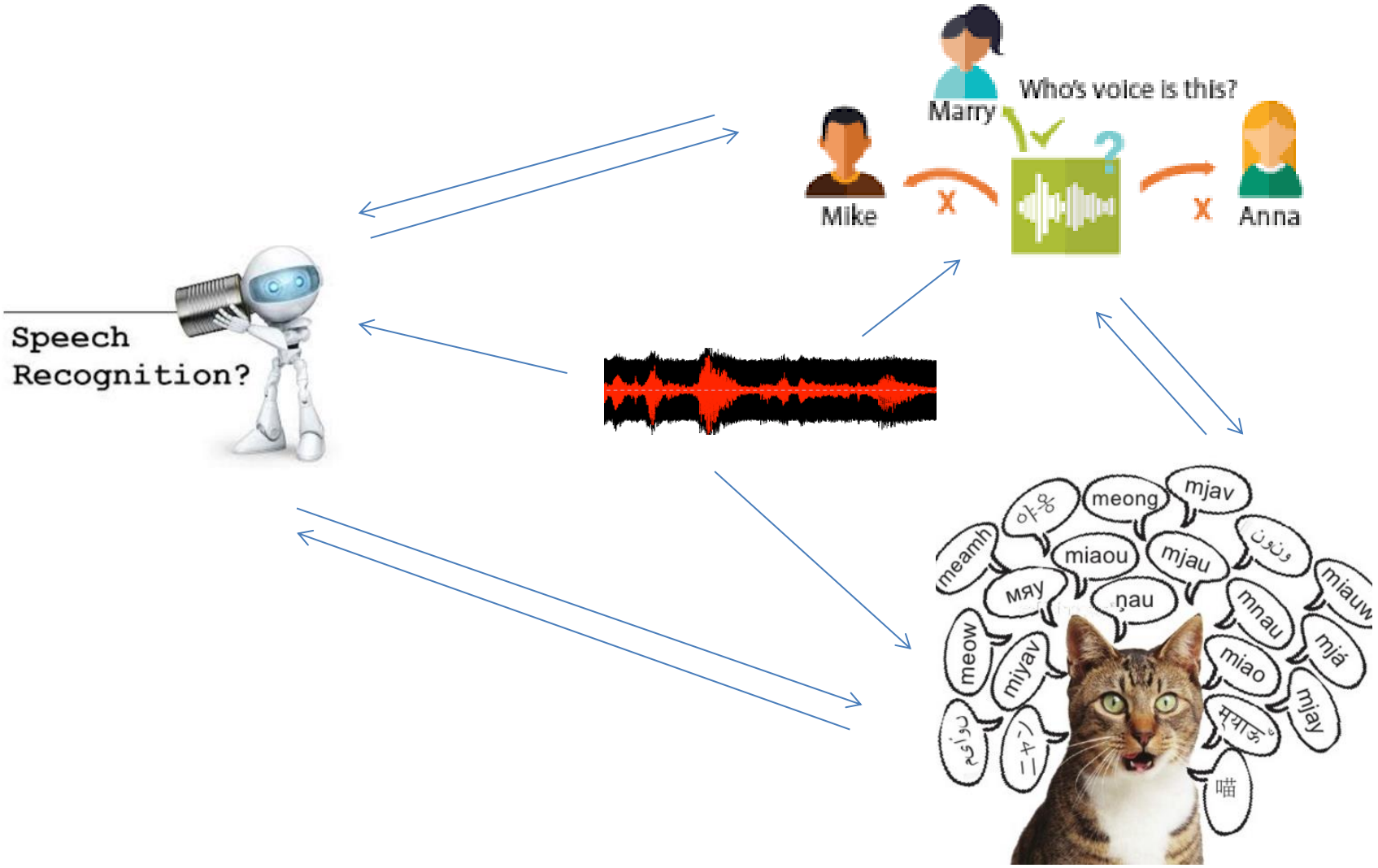


图4. 协同学习示意图。

joint training for ASR, SID, LID



Speech-speaker joint training

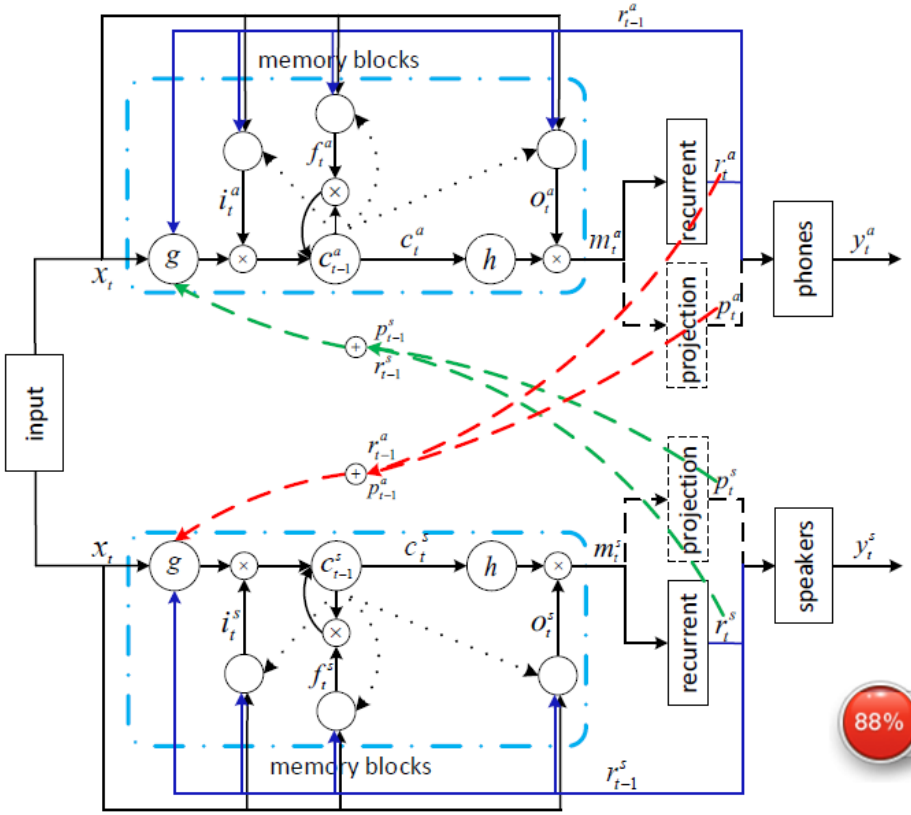


Fig. 4: *Multi-task recurrent model for ASR and SRE, an example.*

Feedback Input	SWB & Fisher Ratio	ASR WER%	SRE EER%
i f o g			
-	-	24.0 (SWB)	15.95 (Fisher)
√	1 : 0.5	23.2	11.44
√	1 : 0.5	**22.6**	10.91
√	1 : 1	24.2	11.09
√	1 : 1	23.2	10.85
√	1 : 2	26.0	**10.54**
√	1 : 2	24.5	11.07
√	1 : 3	28.8	11.13
√	1 : 3	26.5	12.05

Zhiyuan Tang, Lantian Li, Dong Wang, Collaborative Joint Training with Multi-task Recurrent Model for Speech and Speaker Recognition, IEEE TASLP 2017

Speaker-Language joint training

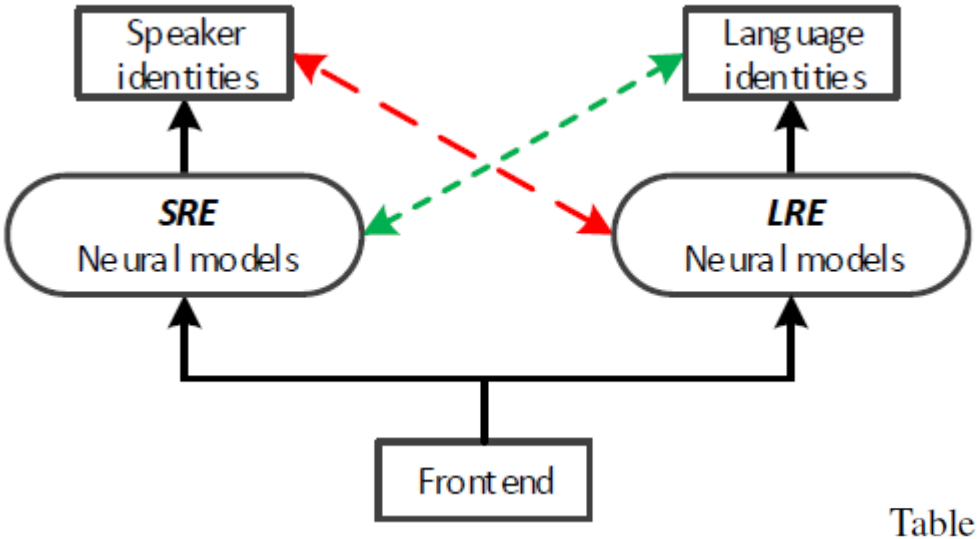


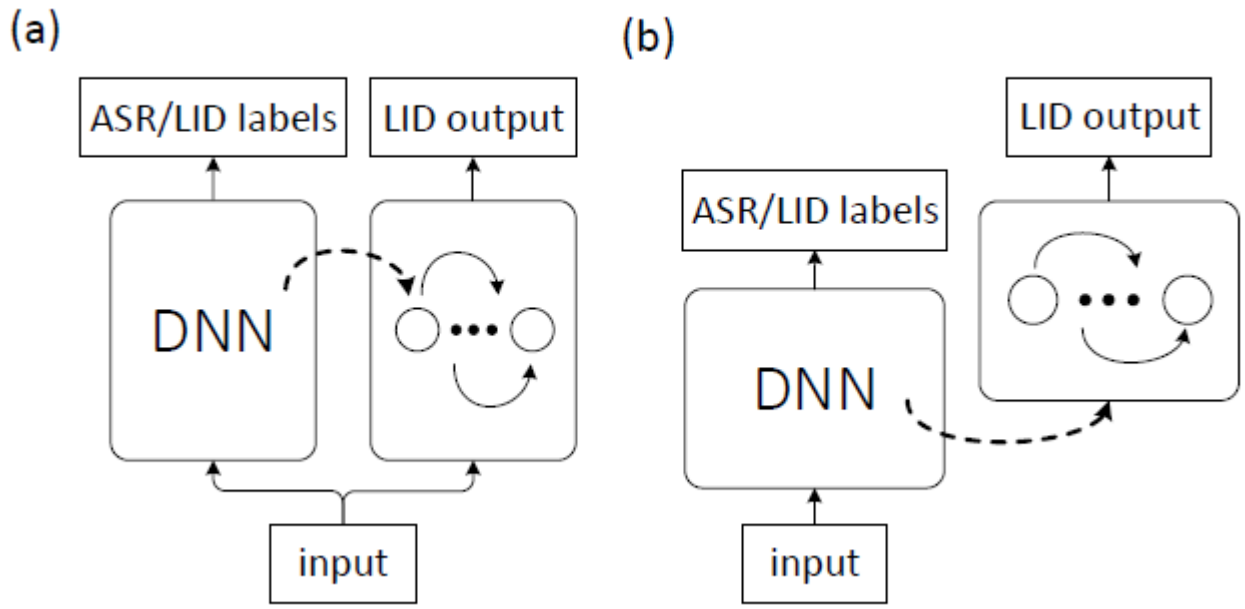
Table 3: *SRE results with collaborative learning.*

Feedback Input				EER(%)			
				Full		Short	
i	f	o	g	Eng.	Chs.	Eng.	Chs.
r-vector Baseline				1.38	1.61	2.70	3.99
√				1.27	1.43	2.50	3.61
	√			1.38	1.38	2.55	3.52
		√		**1.19**	**1.31**	**2.48**	3.66
			√	1.37	1.48	2.67	**3.52**
√	√	√	√	1.32	1.31	2.52	3.69

Table 4: *LRE results with collaborative learning.*

Feedback Input				IDE					
				Full			Short		
i	f	o	g	Cosine	SVM	Softmax	Cosine	SVM	Softmax
r-vector Baseline				25	47	29	218	139	129
√				5	2	0	11	6	2
	√			1	0	0	3	1	1
		√		11	2	0	21	8	3
			√	0	0	1	2	2	1
√	√	√	√	6	2	0	17	10	2

Speech-language joint training



Speech-language joint training

TABLE VI: Results of various LID systems on the 7 languages in Babel.

					C_{avg}		EER%	
System	Phonetic DNN	LID model	Info. Rec.	LID Feature	Fr.	Utt.	Fr.	Utt.
i-vector	-	i-vector	-	-	-	0.1696	-	16.52
Acoustic RNN	-	LSTM-RNN	-	Fbank	0.1987	0.1249	19.11	12.63
Ph. Aware	AG-TDNN-MLT	LSTM-RNN	g	Fbank	0.1280	0.0620	12.27	6.66
Ph. Aware	SWB-TDNN-ASR	LSTM-RNN	g	Fbank	0.1610	0.0786	15.49	8.24
PTN	AG-TDNN-MLT	LSTM-RNN	Input	Phonetic	**0.1165**	**0.0518**	**11.12**	**5.70**
PTN	SWB-TDNN-ASR	LSTM-RNN	Input	Phonetic	0.1726	0.0823	16.65	8.56

Need understand how factors are convolved

- Some factors are additional (white noise)
- Some factors are convolutive (linguistic and speaker)
- Some factors are hierarchical (linguistic and language)
- Some factors are explained away (speaker and language)

One speaker is trait
A group of speakers is dialect
A large group of speakers is language

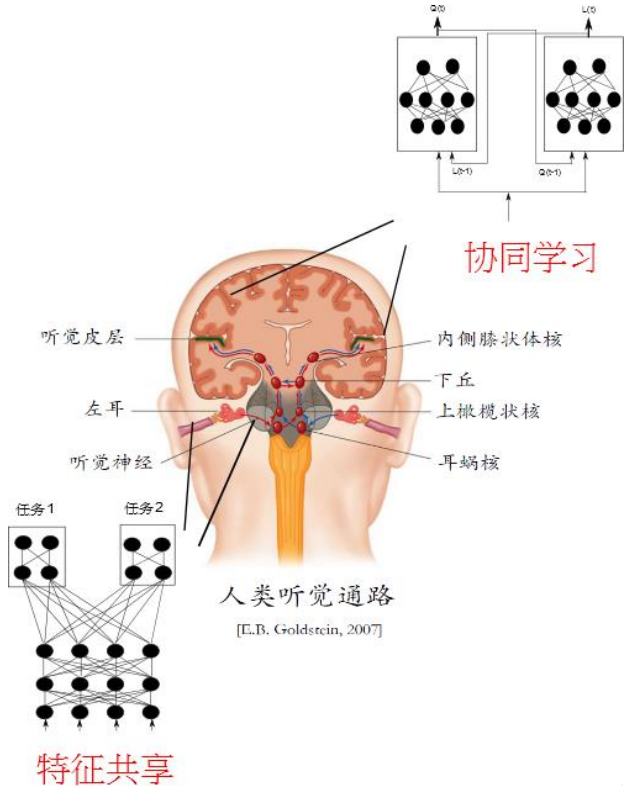


图5. 人类听觉系统中的特征共享和协同学习

A better way is to factorize

- Probabilistic factorization
  - Define a compositional rule, and find the optimal separation by following the rule with a criterion, usually maximal likelihood
  - PCA: $y=u_0+Tw+\varepsilon$
  - Gaussian MAP: $y=u0+Dw+\varepsilon$
  - PLDA: $M_{ij}=u_0 + Tw_i + \varepsilon_{ij}$

Difficulty of the shallow factorization

- Most of the existing factorization methods are shallow, linear and Gaussian (except ICA)
- They require explicit definition of the relationship
- Very difficult to apply to real signals like speech

Several possibilities

- Deep unsupervised factorization, RBM, AE. However it is blind to task-oriented factors.
- Linear, shallow, Gaussian factorization for deep features:

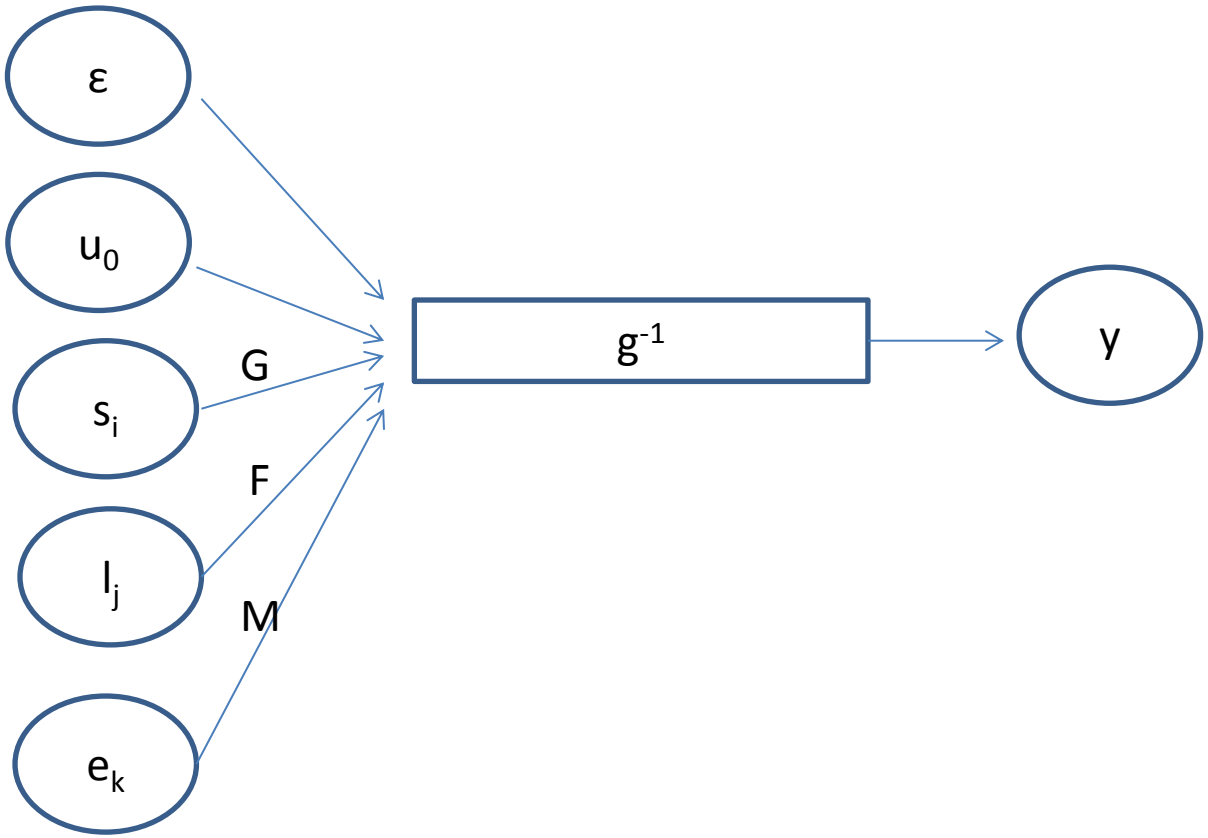
$$g(y)=u0+Gs_i+Fl_j+Me_k+\varepsilon$$

- Linear, Shallow, Gaussian factorization of deep recovery:

$$g(s)+f(l)+m(e)=y$$

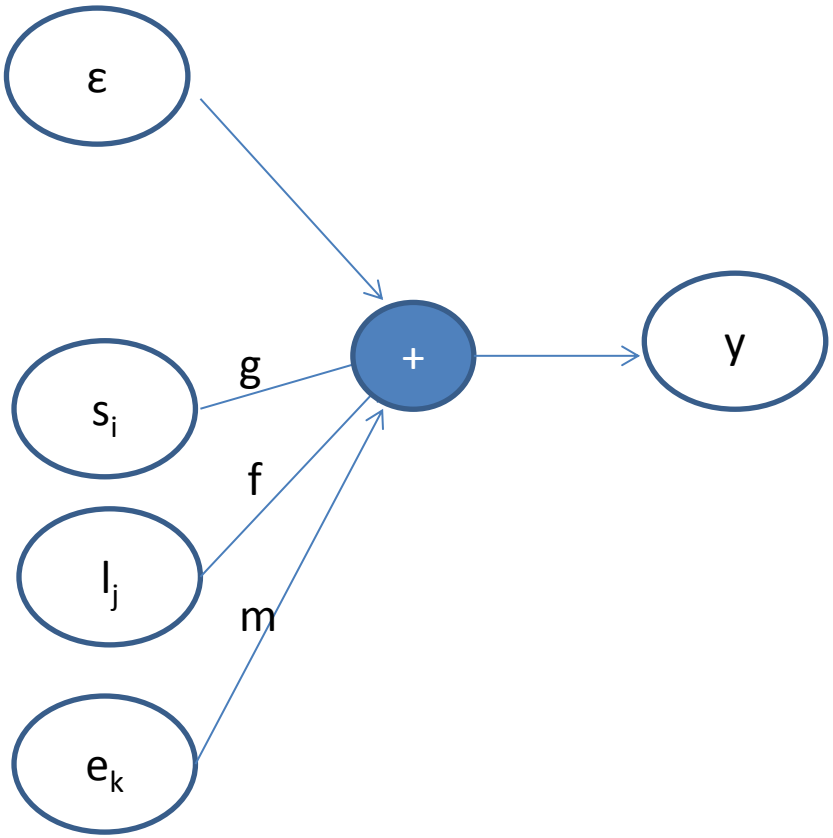
Factorization for deep feature

- $g(y)=u0+Gs_i+Fl_j+Me_k+\varepsilon$

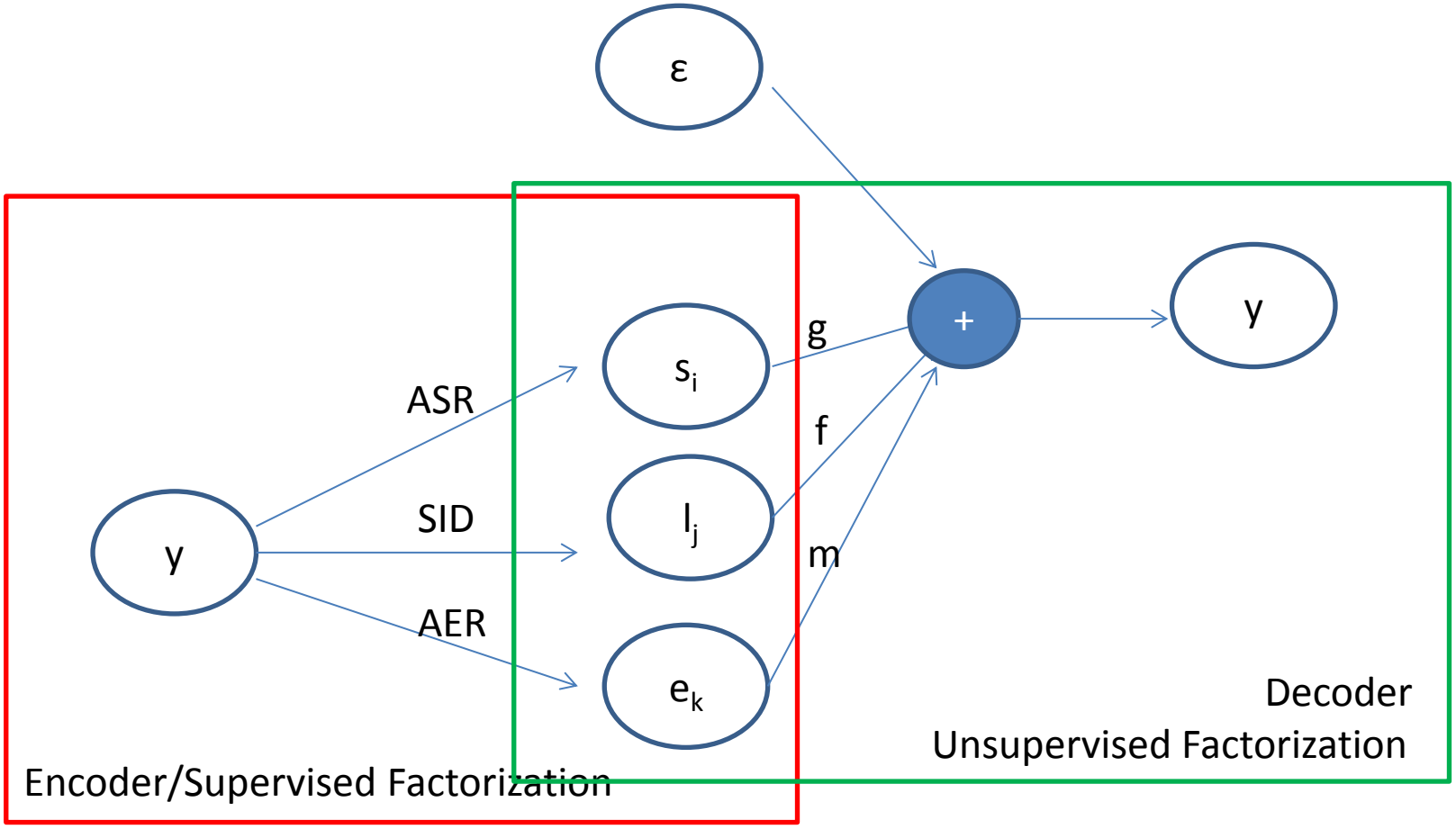


Factorization for recovery

- g(s)+f(l)+m(e)+ ε =y



Supervised deep factorization



Think a minute

- We can factorize the speech signal already, if we have all the single-task systems
- The factors can be used to recover the speech It is like an AE, and something like VADE, but **task-supervised**
- The decoder simulates convolutional, if it is in the log space (later to see). So prior knowledge applied.

Think more

- But, if we have individual task systems, why we factorize? !$@^&@#$%
- At least two reasons:
  - Only significant factors can be learned by individual task system, due to the limited data amount and the significance of the factor
  - Factorization is not only for tasks for these factors, but an analysis tool for a broad range of potentials

Two contributions we made

- Conditional deep factorization
- Deep recovery

Conditional factorization

- Only significant factors that with a large amount of data can be factorized individually by deep learning
- These factors will benefit other factors that are not so significant and not too much data
- Conditioning on the significant factors will make things much more easy
- Form an iterative form that decipher speech signal

Speaker factor is the impediment

- Factors, if can be factorized out, has to be short-term
- Speech factor is short-term, but speaker factor is not necessarily the case, considering the great success of model-based approach
- Fortunately, we found they are short term

Speaker factor learning

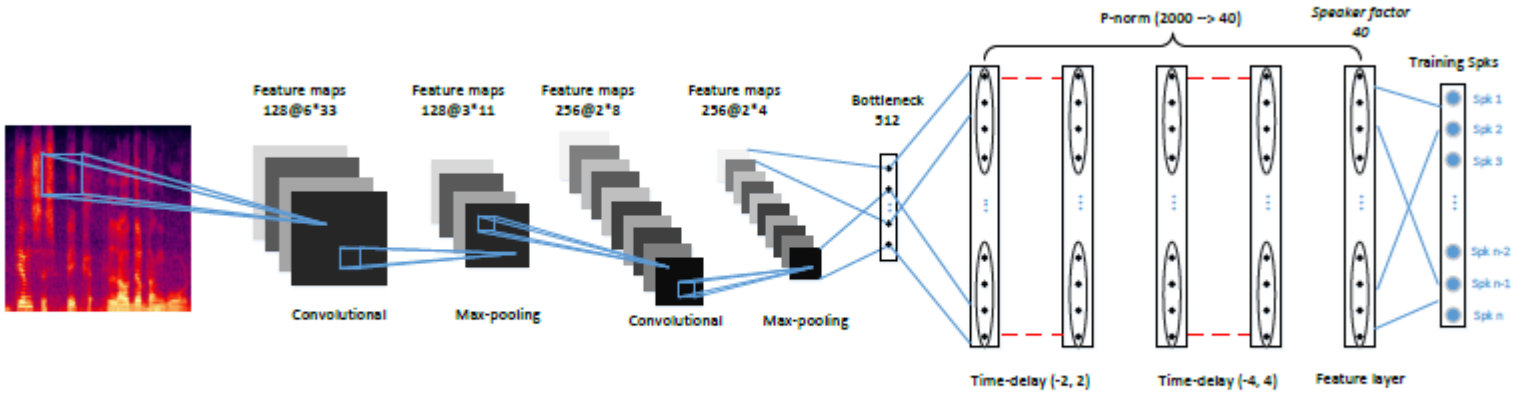
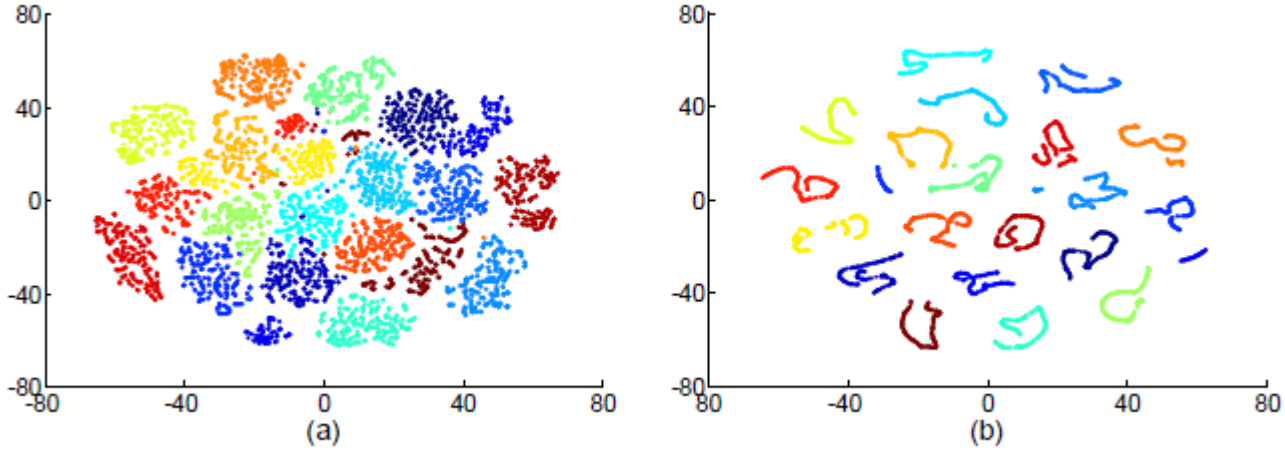


Figure 1: The DNN structure used for deep speaker factor inference.

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

Speaker factor learning



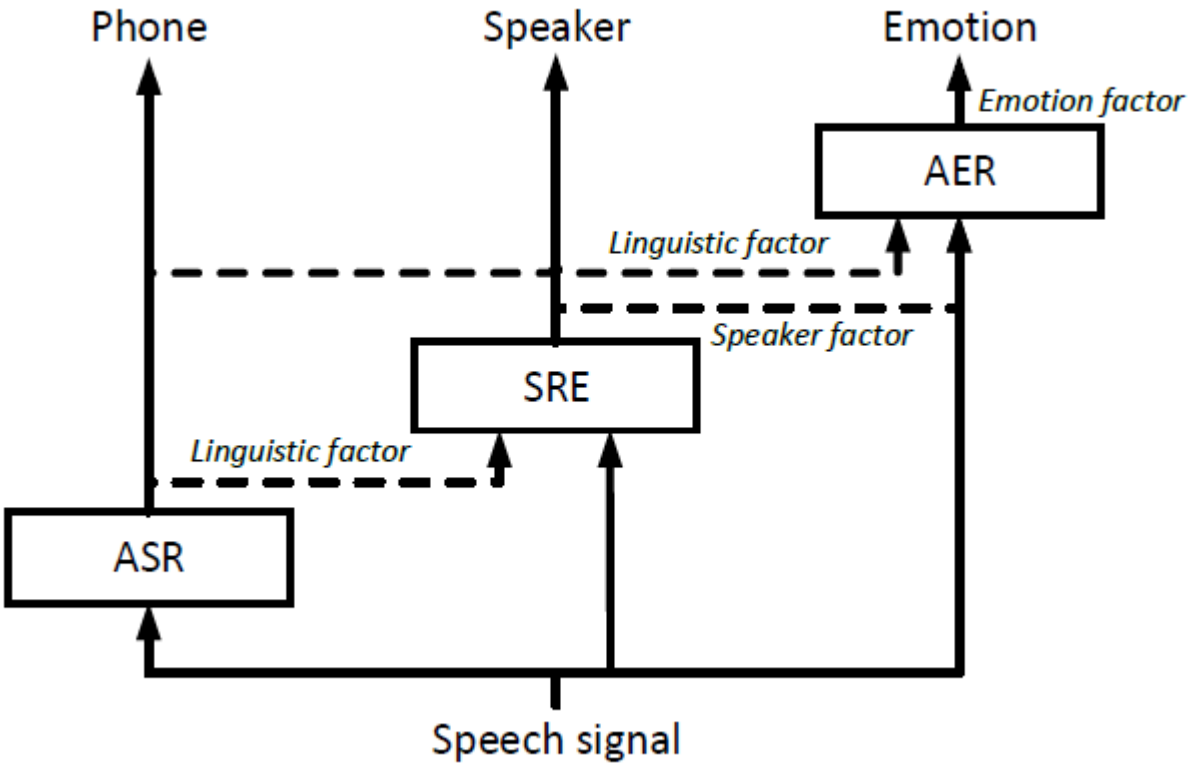
Systems	Metric	20 frames	50 frames	100 frames
i-vector	Cosine	30.01	18.23	11.14
	LDA	29.47	15.96	8.64
	PLDA	29.29	15.71	8.34
d-vector	Cosine	**7.68**	6.67	4.61
	LDA	7.88	**4.72**	**3.02**
	PLDA	20.81	15.02	8.98

Speaker factor

This will make great impact

- Speaker traits are spectrum properties rather than distributional pattern
- Speaker traits can be identified with a short segment (0.3 seconds, 93% accuracy)
- Very impressive in research and industry
- Linguistic content and speaker traits are two of the major factors in speech signal, therefore speech signals are largely factorizable!

Cascaded deep factorization



CDF for speaker factorization

Table 1: The ***Top-1*** IDR(%) results on the short-time speaker identification with the i-vector and two d-vector systems.

		IDR%		
Systems	Metric	S(30-20f)	S(30-50f)	S(30-100f)
i-vector	PLDA	5.72	27.77	55.06
d-vector (IDF)	Cosine	37.18	51.24	**65.31**
d-vector (CDF)	Cosine	**47.63**	**57.72**	64.45

CDF for emotional detection

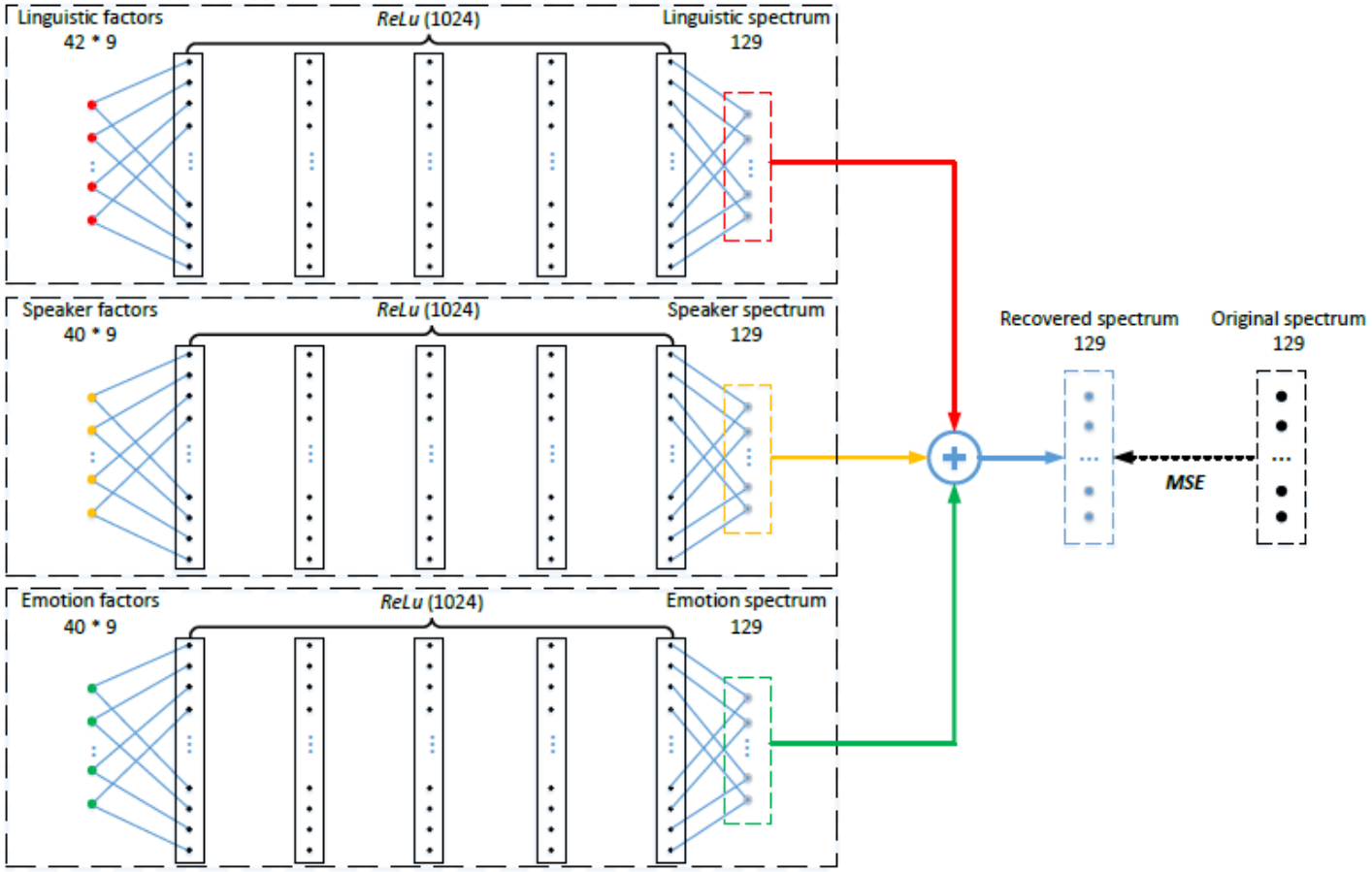
Table 2: Accuracy (ACC) and macro average precision (MAP) on the training set.

Dataset	Training set			
	ACC% (fr.)	MAP% (fr.)	ACC% (utt.)	MAP% (utt.)
Baseline	74.19	61.67	92.27	83.08
+ling.	86.34	81.47	96.94	96.63
+spk.	92.56	90.55	97.75	97.16
+ling. & spk.	**94.59**	**92.98**	**98.02**	**97.34**

Table 3: Accuracy (ACC) and macro average precision (MAP) on the evaluation set.

Dataset	Evaluation set			
	ACC% (fr.)	MAP% (fr.)	ACC% (utt.)	MAP% (utt.)
Baseline	23.39	21.08	28.98	24.95
+ling.	27.25	27.68	**33.12**	**33.28**
+spk.	27.18	28.99	32.01	32.62
+ling. & spk.	**27.32**	**29.42**	32.17	32.29

Deep speech recovery



Deep speech recovery

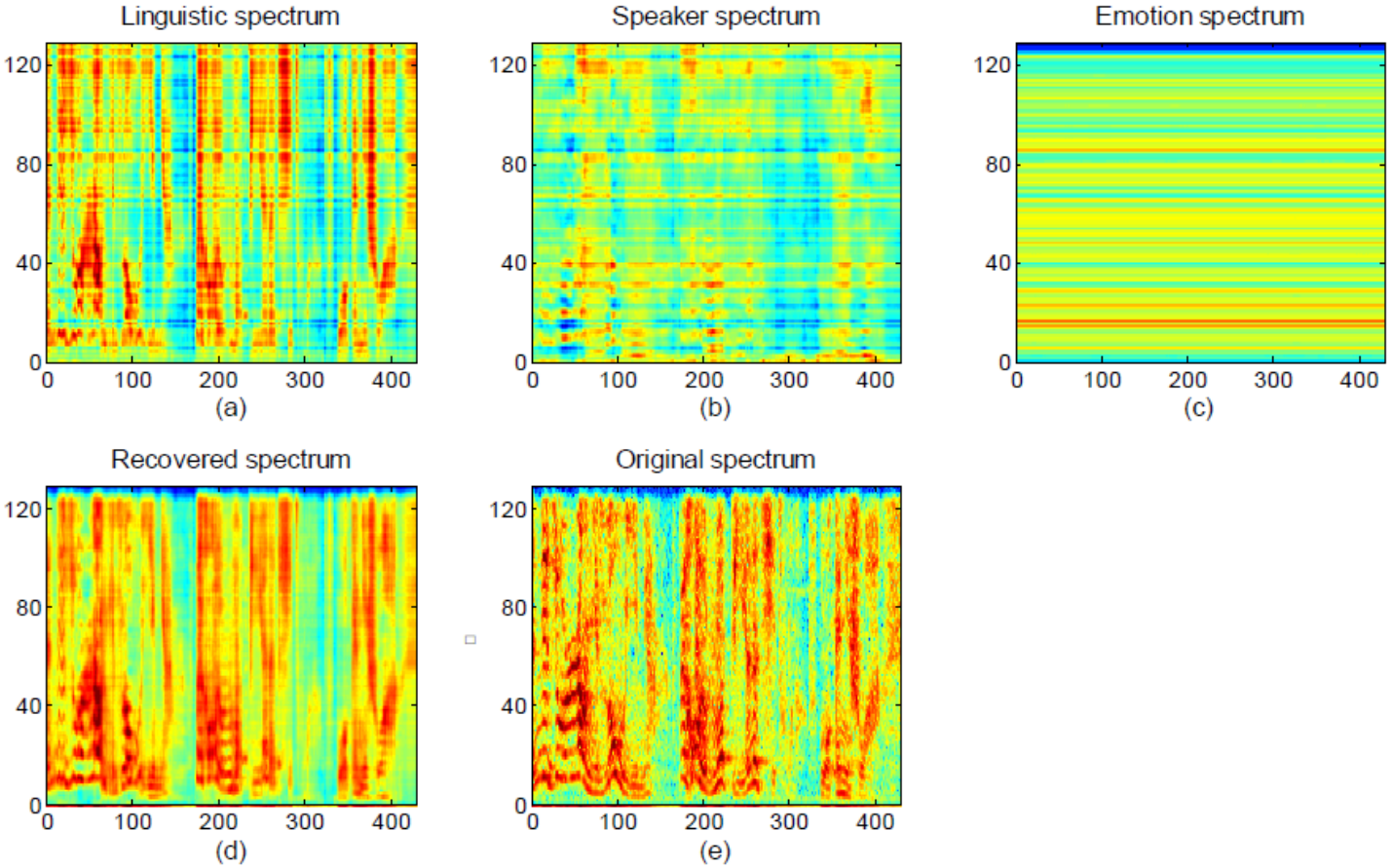
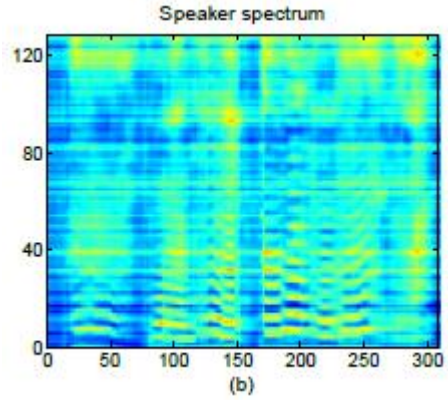
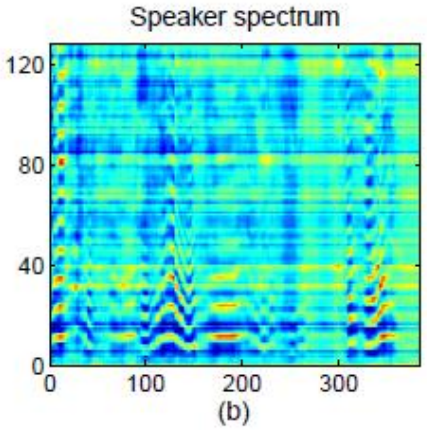
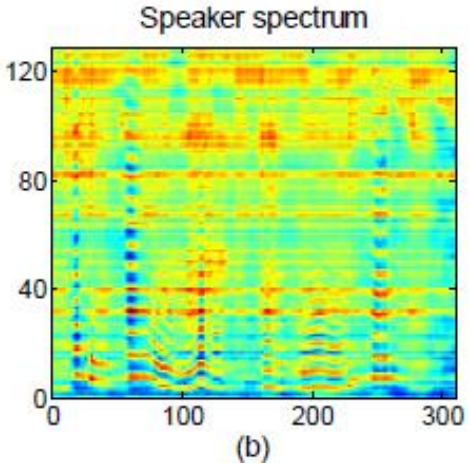


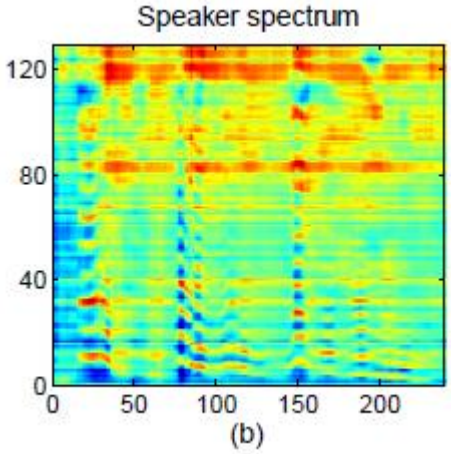
Figure 5: An example of spectrum reconstruction from the linguistic, speaker and emotion factors.

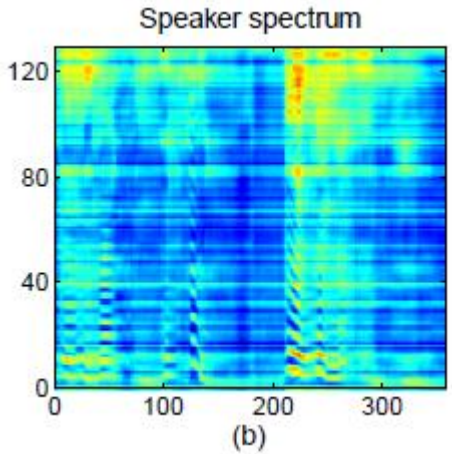
Benefit 1: A novel voiceprint toolkit

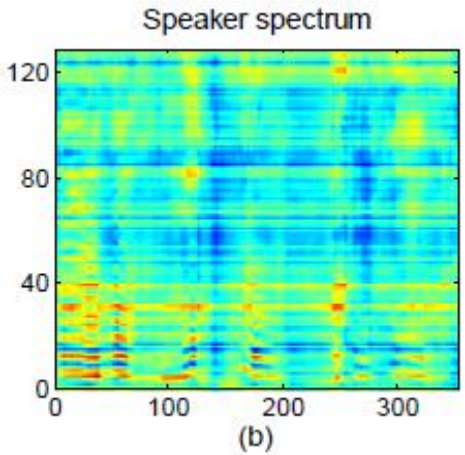












Benefit 2: A parsimonious voice encoder

- Encoder different factors separately
- Can choose what to transfer

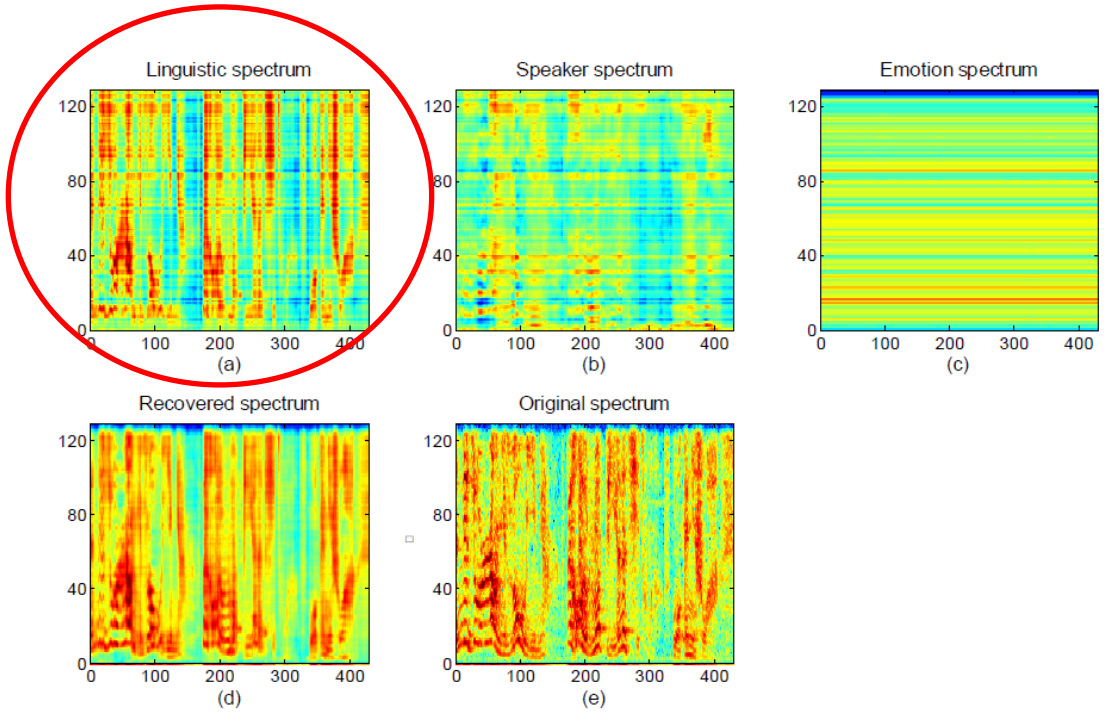


Figure 5: An example of spectrum reconstruction from the linguistic, speaker and emotion factors.

Benefit 3: A flexible way for voice transform and emotional converter

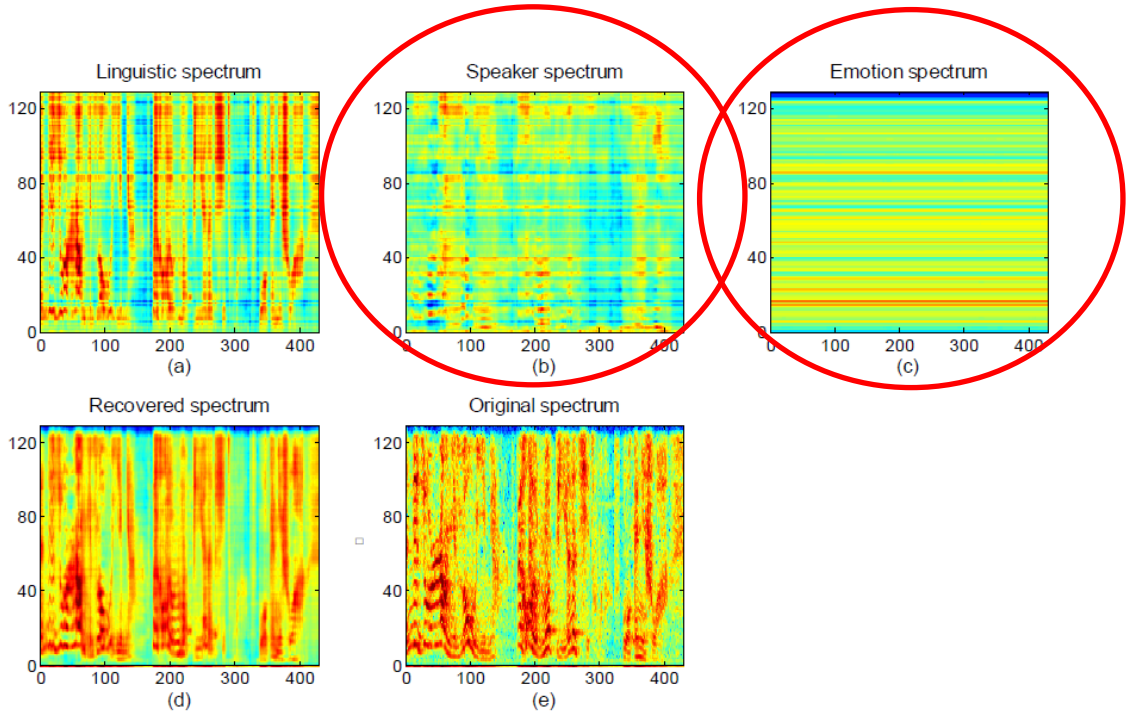


Figure 5: An example of spectrum reconstruction from the linguistic, speaker and emotion factors.

Benefit 4: A possible way for audio watermarking

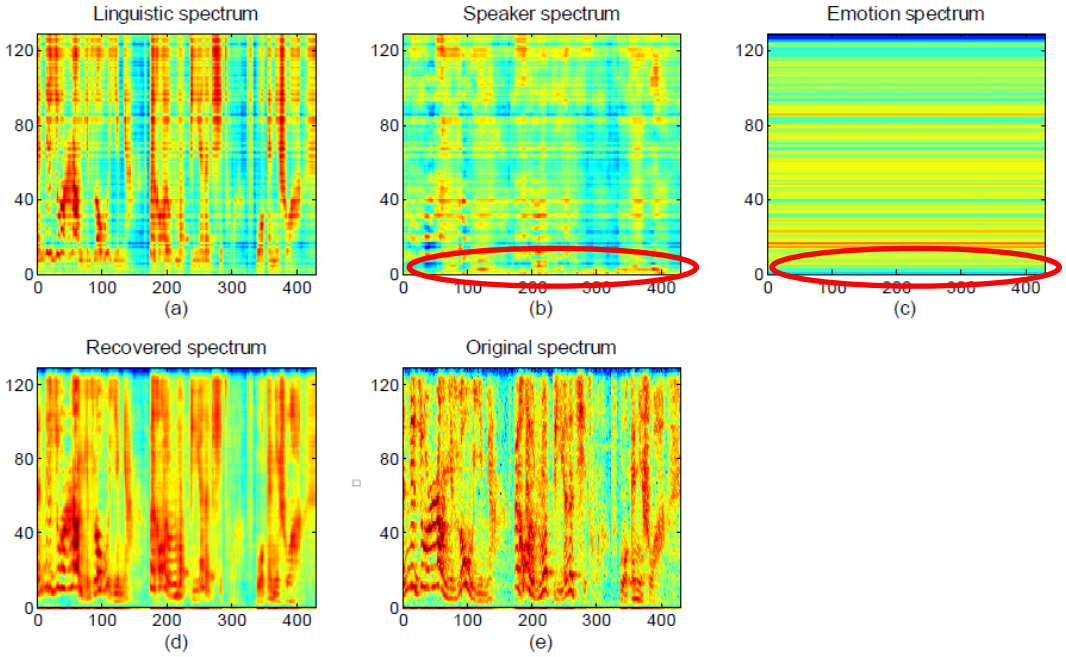


Figure 5: An example of spectrum reconstruction from the linguistic, speaker and emotion factors.

Benefit 5: A possible way to separate speaker

- Use both continuity of speech and speaker spectrum
- But first need to detect if overlapped is found
- Mixed speaker recognition?

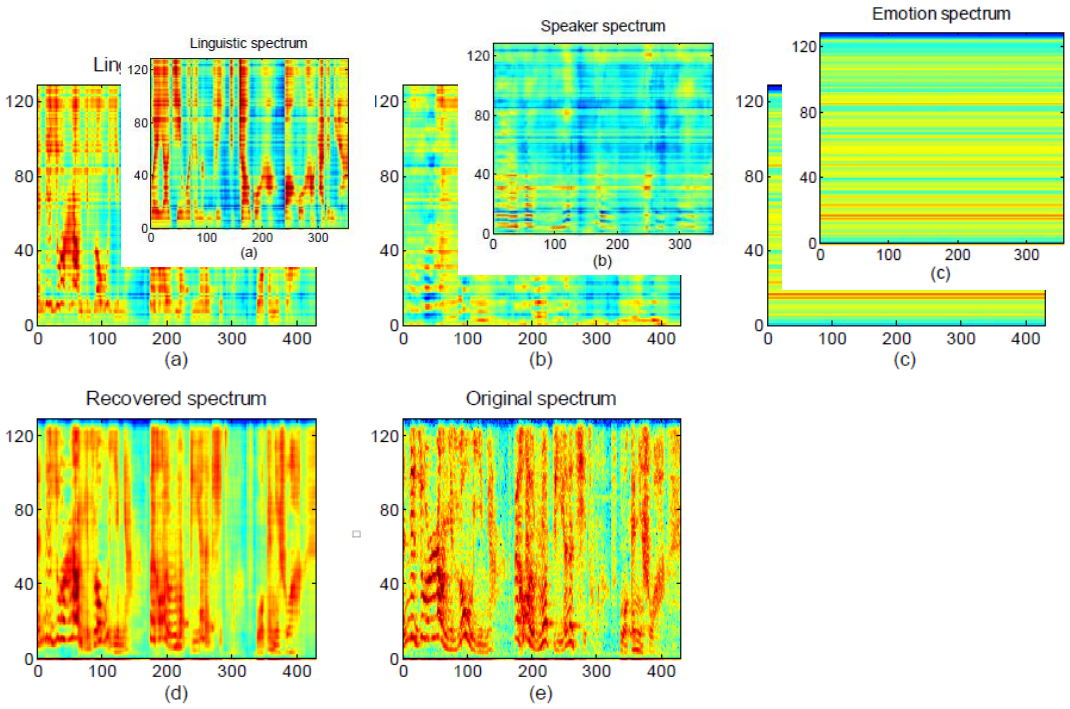
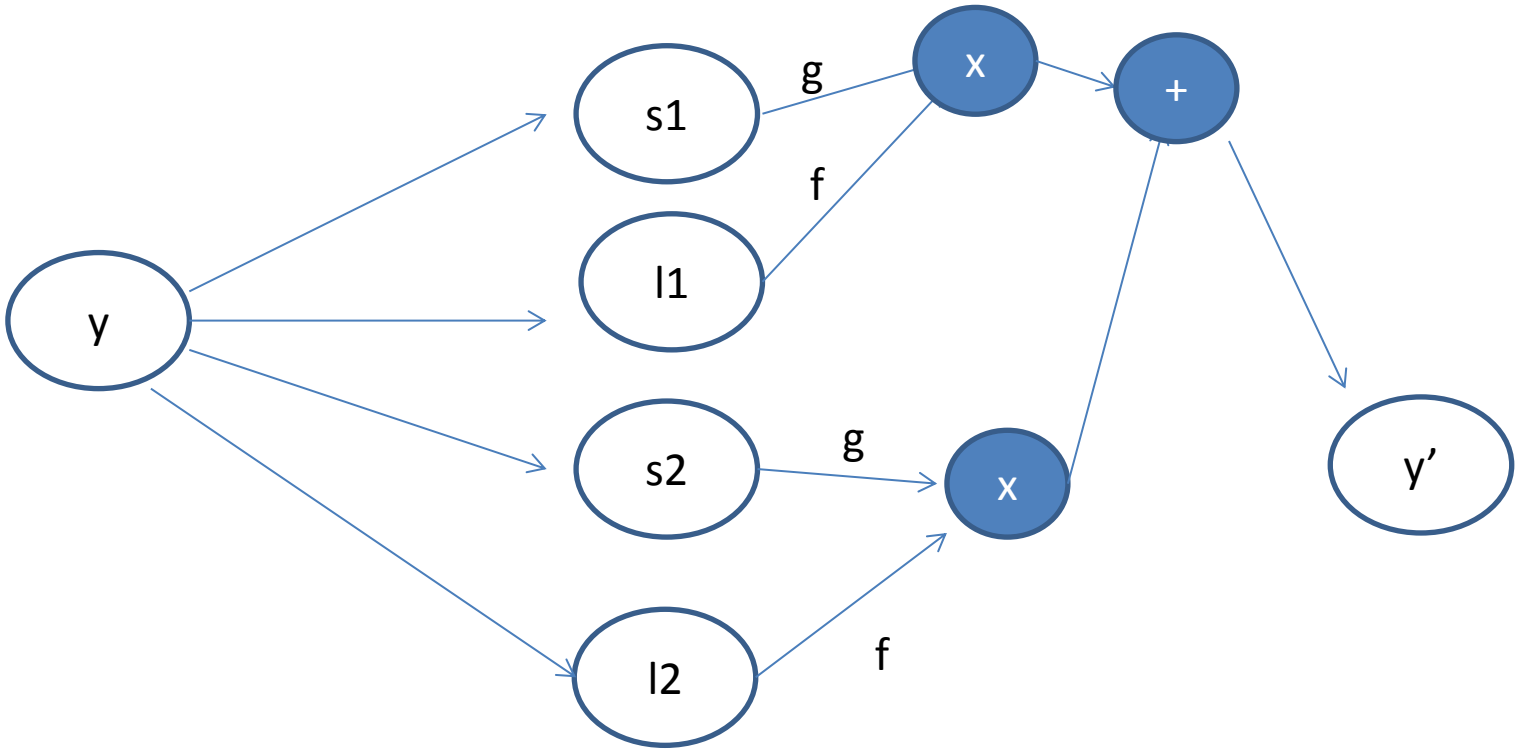


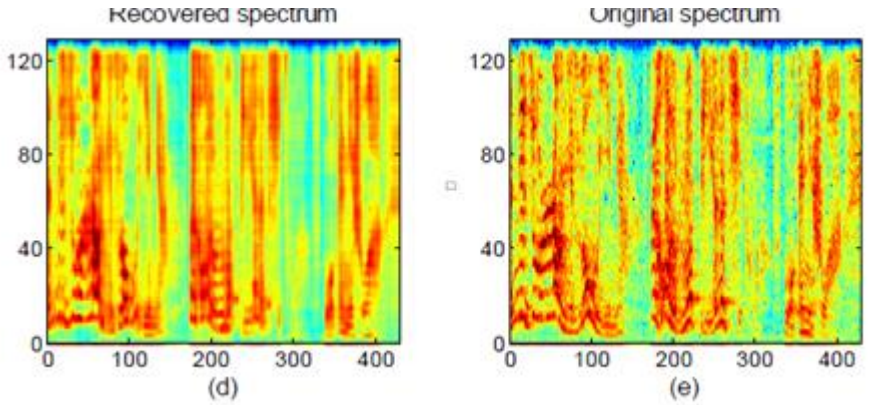
Figure 5: An example of spectrum reconstruction from the linguistic, speaker and emotion factors.

A possible structure



Benefit 6: A possible way for noise reduction

- DNN automatically removed channel Gaussian noise
- But how other noises?



Conclusions

- Speech signals are composed of many factors, seems very complex
- However we can factorize linguistic and speaker factors by single-task deep learning.
- A cascaded approach can be used to factorize speech signals one by one, which opens the door to decipher speech signals.
- Speech signals can be recovered by the deep factors by a recovery deep neural network. This provides a lot of opportunities for us to manipulate the speech signal from a very novel perspective. Previous source+channel, now phone+speaker! This is more task-oriented.
- Speech signal, once it can be factorized, seems very simple now. Many things can be done now.
- What makes it possible? Deep learning + large data, which makes IDF possible, then CDF, then total factorization.