Deep Learning for Speech Recognition

Hung-yi Lee

Outline

- Conventional Speech Recognition
- How to use Deep Learning in acoustic modeling?
- Why Deep Learning?
- Speaker Adaptation
- Multi-task Deep Learning
- New acoustic features
- Convolutional Neural Network (CNN)
- Applications in Acoustic Signal Processing

Conventional Speech Recognition

Machine Learning helps

This is a structured learning problem.

Evaluation function: $\mathrm{F}(X, W) = P(W|X)$

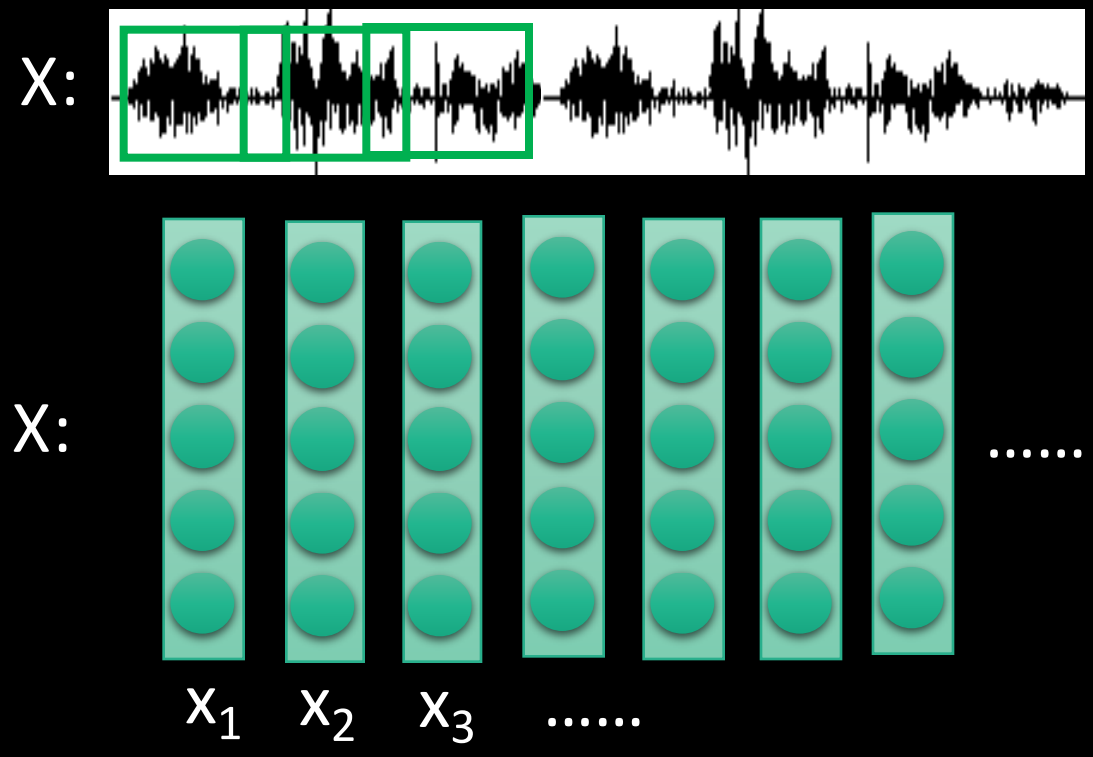
Inference: $\tilde{W} = arg \max_W F(X, W) = arg \max_W P(W|X)$

$$= arg \max_W \frac{P(X|W)P(W)}{P(X)} = arg \max_W P(X|W)P(W)$$

$P(X|W)$: Acoustic Model, $P(W)$: Language Model

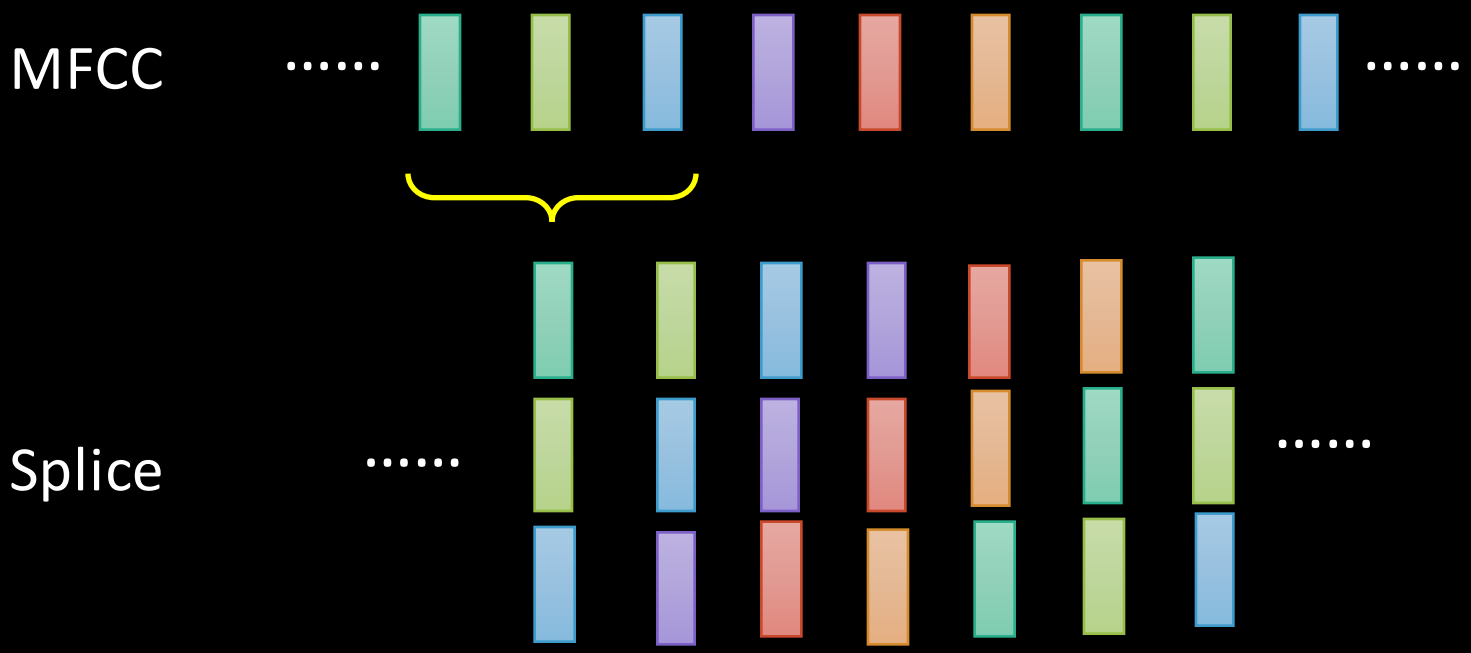
Input Representation

- Audio is represented by a vector sequence



Input Representation - Splice

- To consider some temporal information



Phoneme

- Phoneme: basic unit

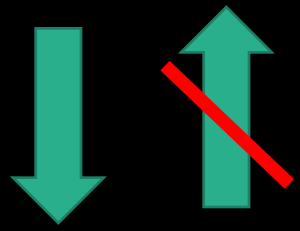
Each word corresponds to a sequence of phonemes.

```
divorce  /d ax v ao1 r s/
divorced  /d ax v ao1 r s t/
divorcee  /d ax v ao2 r s ey1/
do  /d uw1/
doctor  /d aa1 k t axr/
doctors  /d aa1 k t axr z/
doctrine  /d aa1 k t r ax n/
documented  /d aa1 k y uw m eh2 n t ix d/
documents  /d aa1 k y uw m ax n t s/
dodging  /d aa1 jh ix ng/
does  /d ah1 z/
doesn't  /d ah1 z en t/
dog  /d ao1 g/
dogmatically  /d ao g m ae1 t ih k ax l iy/
dogs  /d ao1 g z/
doin'  /d uw1 ix n/
```

Lexicon

what do you think

Lexicon



Different words can correspond to the same phonemes

hh w aa t d uw y uw th ih ng k

State

- Each phoneme correspond to a sequence of states

what do you think

Phone:

hh w aa t d uw y uw th ih ng k

Tri-phone:

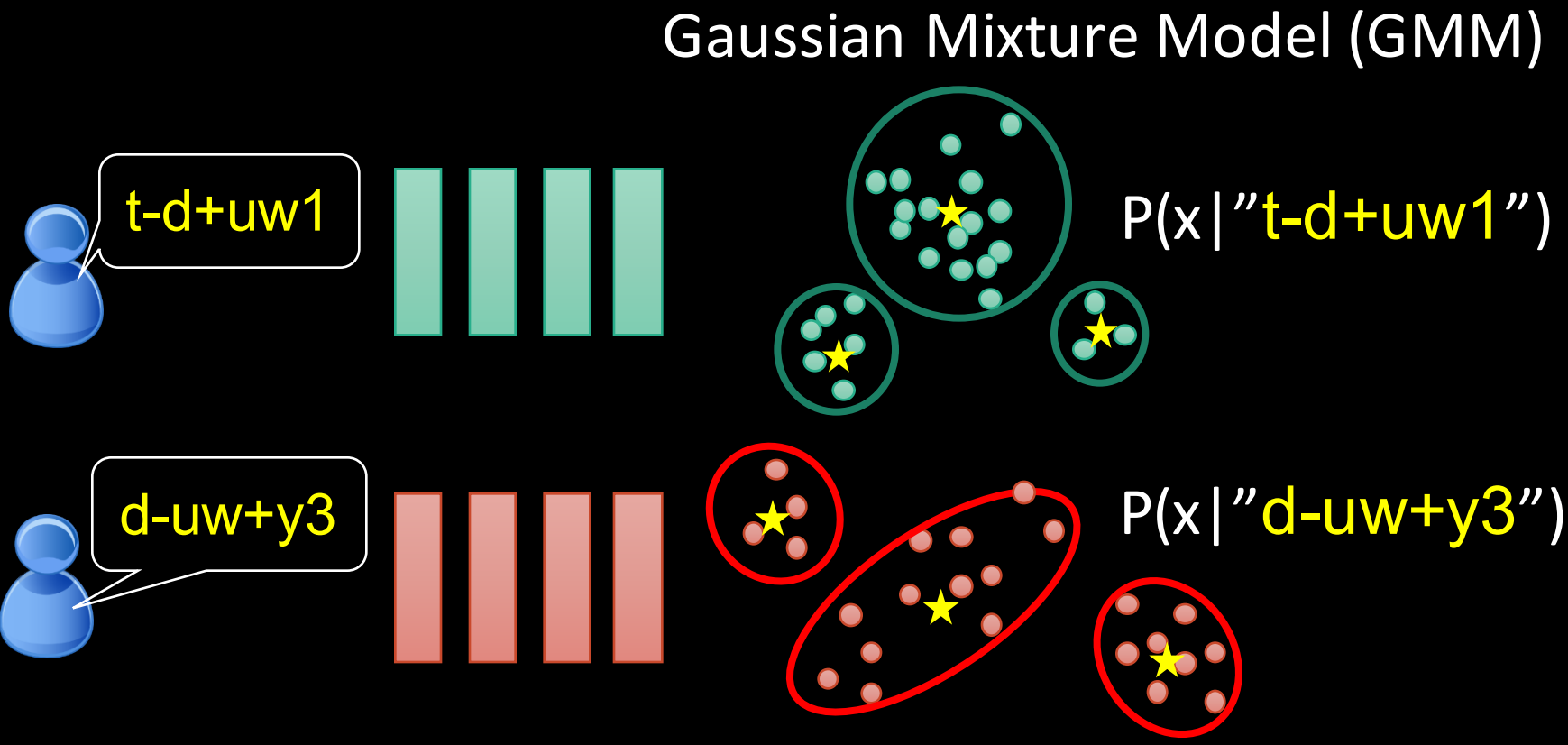
…… t-d+uw d-uw+y uw-y+uw y-uw+th ……

t-d+uw1 t-d+uw2 t-d+uw3 d-uw+y1 d-uw+y2 d-uw+y3

State:

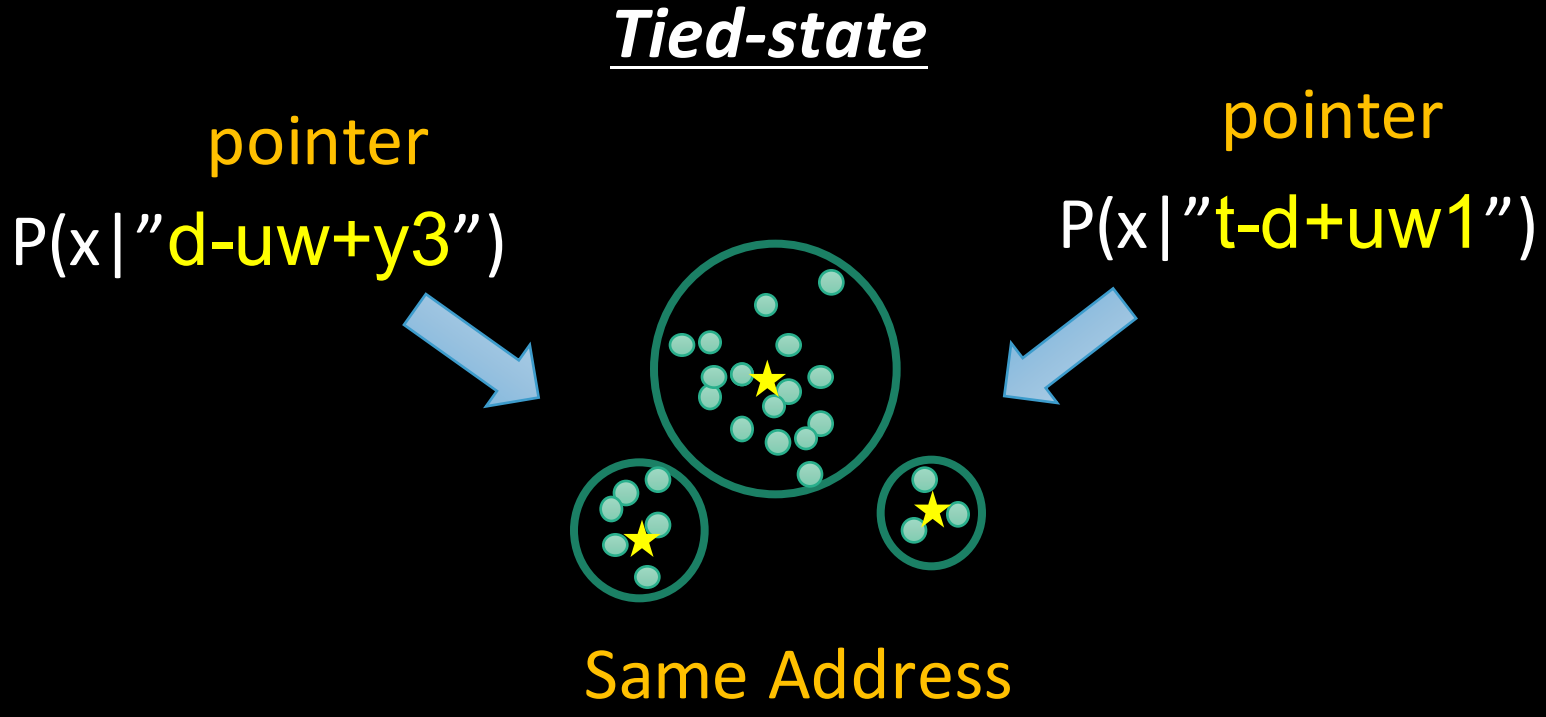
State

- Each state has a stationary distribution for acoustic features



State

- Each state has a stationary distribution for acoustic features

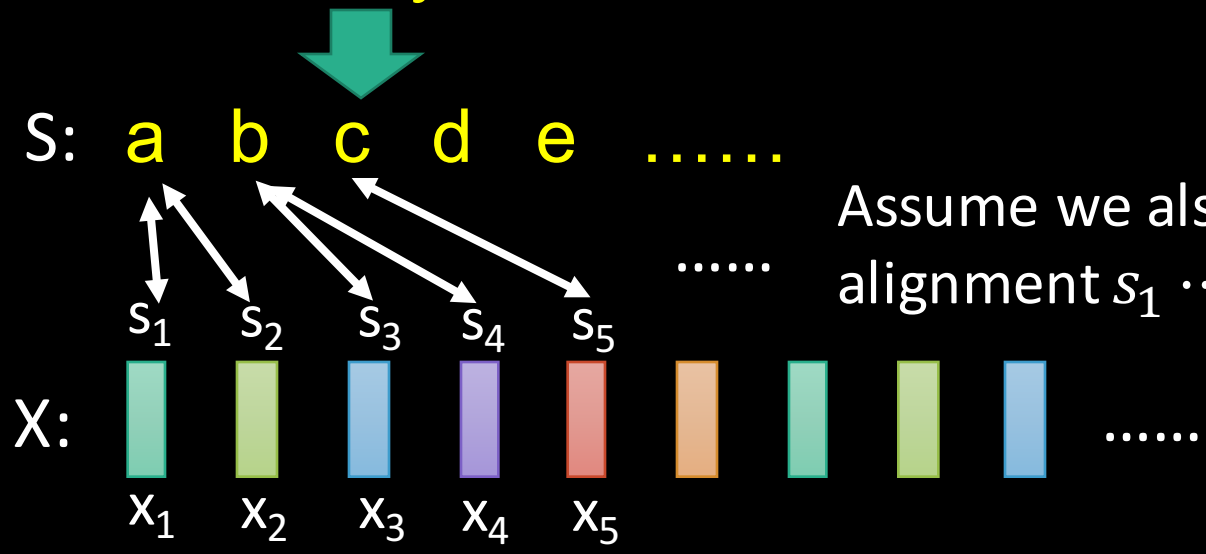


Acoustic Model

$$\tilde{W} = arg\max_{W} P(X|W)P(W)$$

P(X|W) = P(X|S)

W: what do you think?



Assume we also know the alignment $s_1 \cdots s_T$.

$$P(X|S,h) = \prod_{t=1}^{T} \underbrace{P(s_t|s_{t-1})}_{\text{transition}} \underbrace{P(x_t|s_t)}_{\text{emission}}$$

Acoustic Model

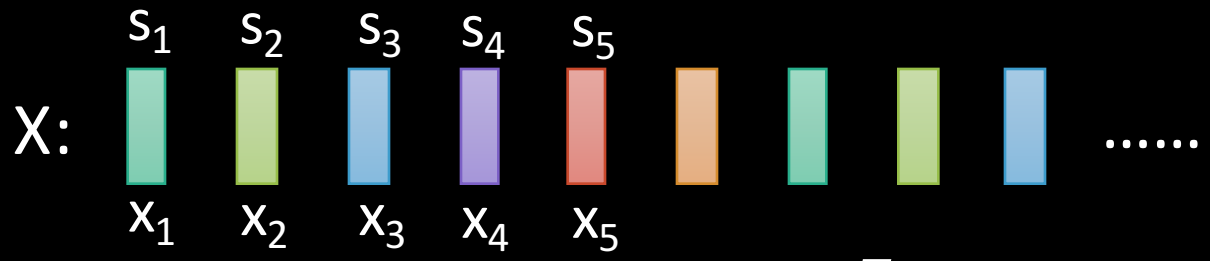
$$\tilde{W} = arg \max_{W} P(X|W) P(W)$$

P(X|W) = P(X|S)

W: what do you think?

S: a b c d e ……

Actually, we don't know the alignment.

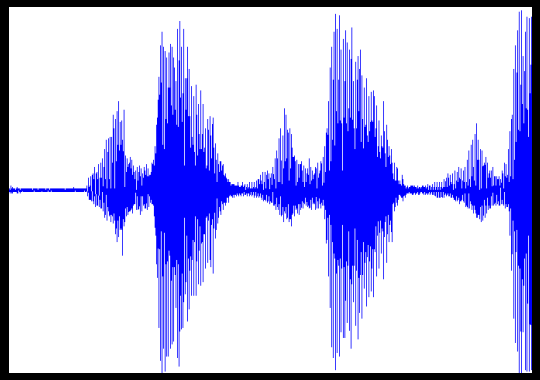


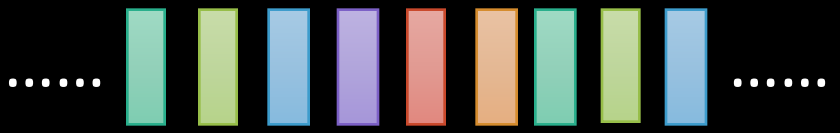
$$P(X|S) \approx \max_{s_1 \cdots s_T} \prod_{t=1}^{T} P(s_t|s_{t-1})\, P(x_t|s_t)$$

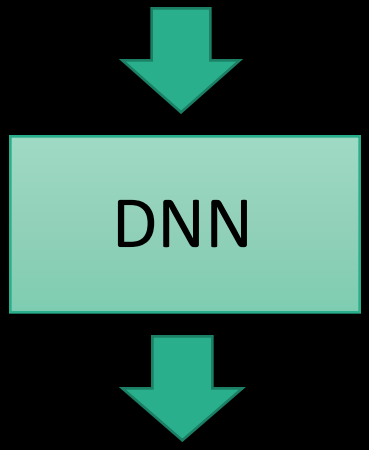
(Viterbi algorithm)

How to use Deep Learning?

People imagine ……



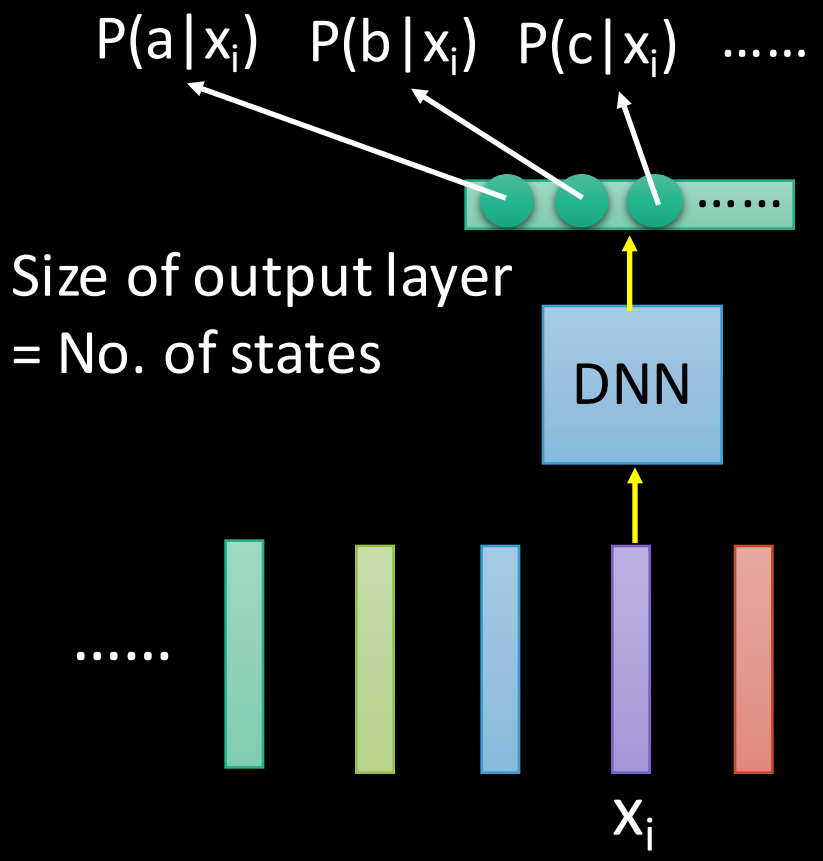




This can not be true!

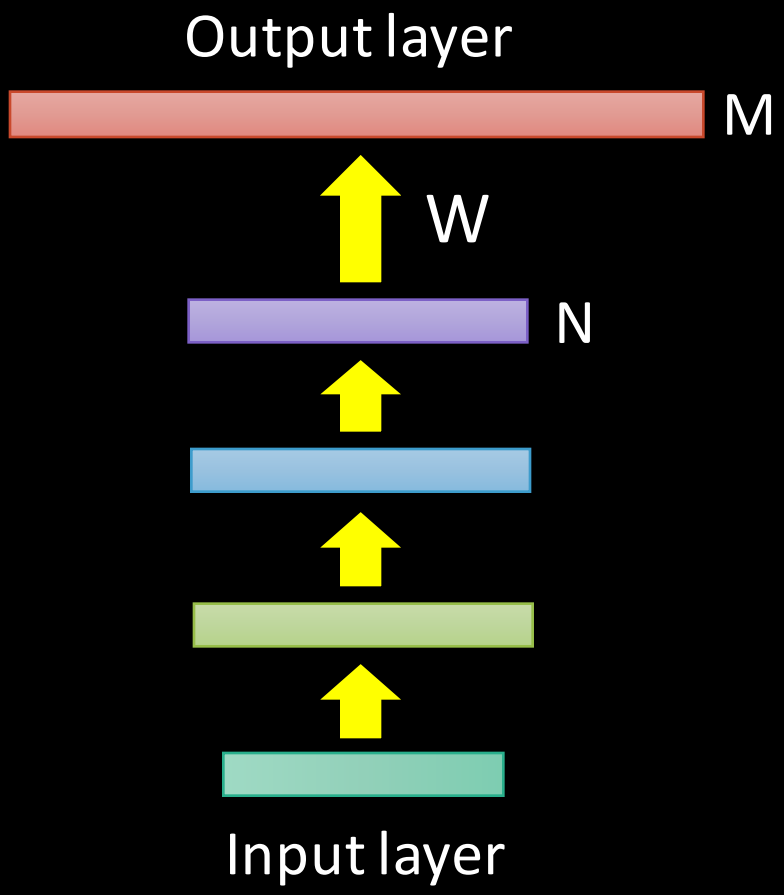
DNN can only take fixed length vectors as input.

What DNN can do is ……



- DNN input:
  One acoustic feature
- DNN output:
  Probability of each state

Low rank approximation



W: M X N

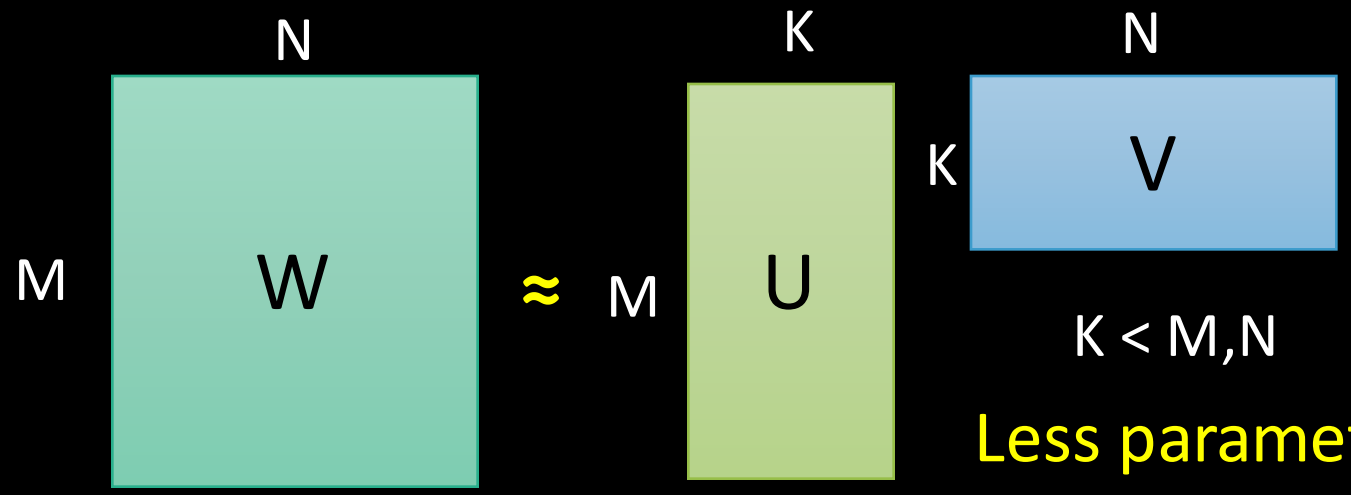
N is the size of the last hidden layer

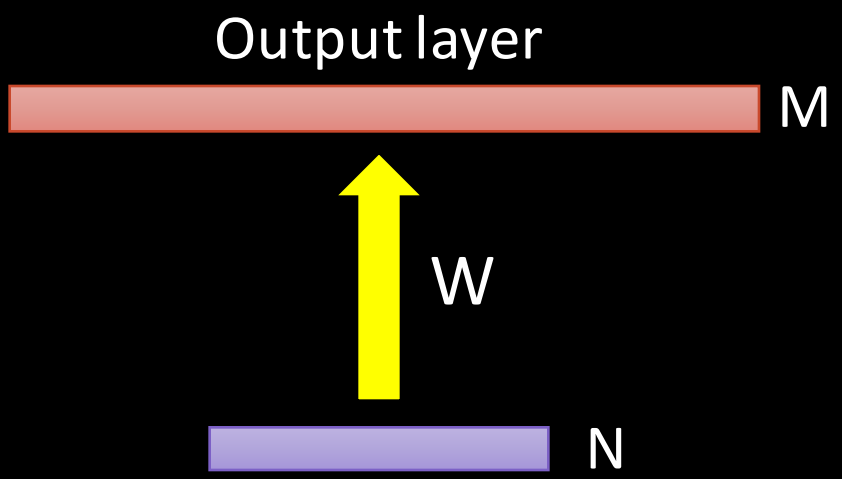
M is the size of output layer

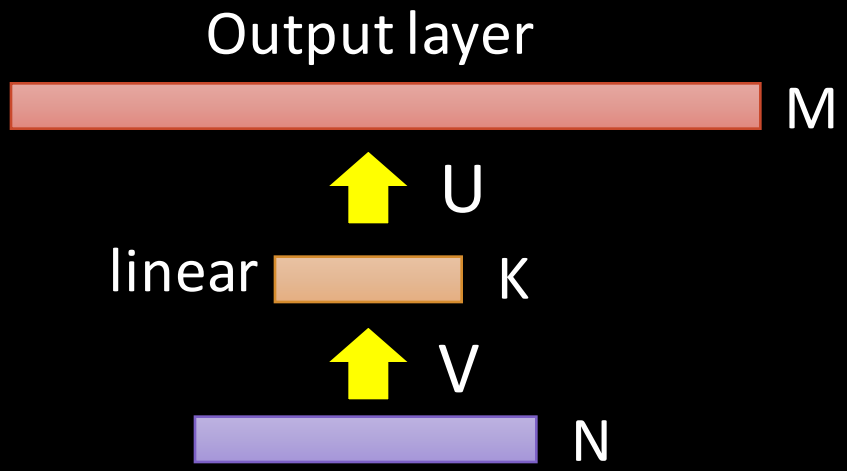
- Number of states

M can be large if the outputs are the ***states of tri-phone***.

Low rank approximation







How we use deep learning

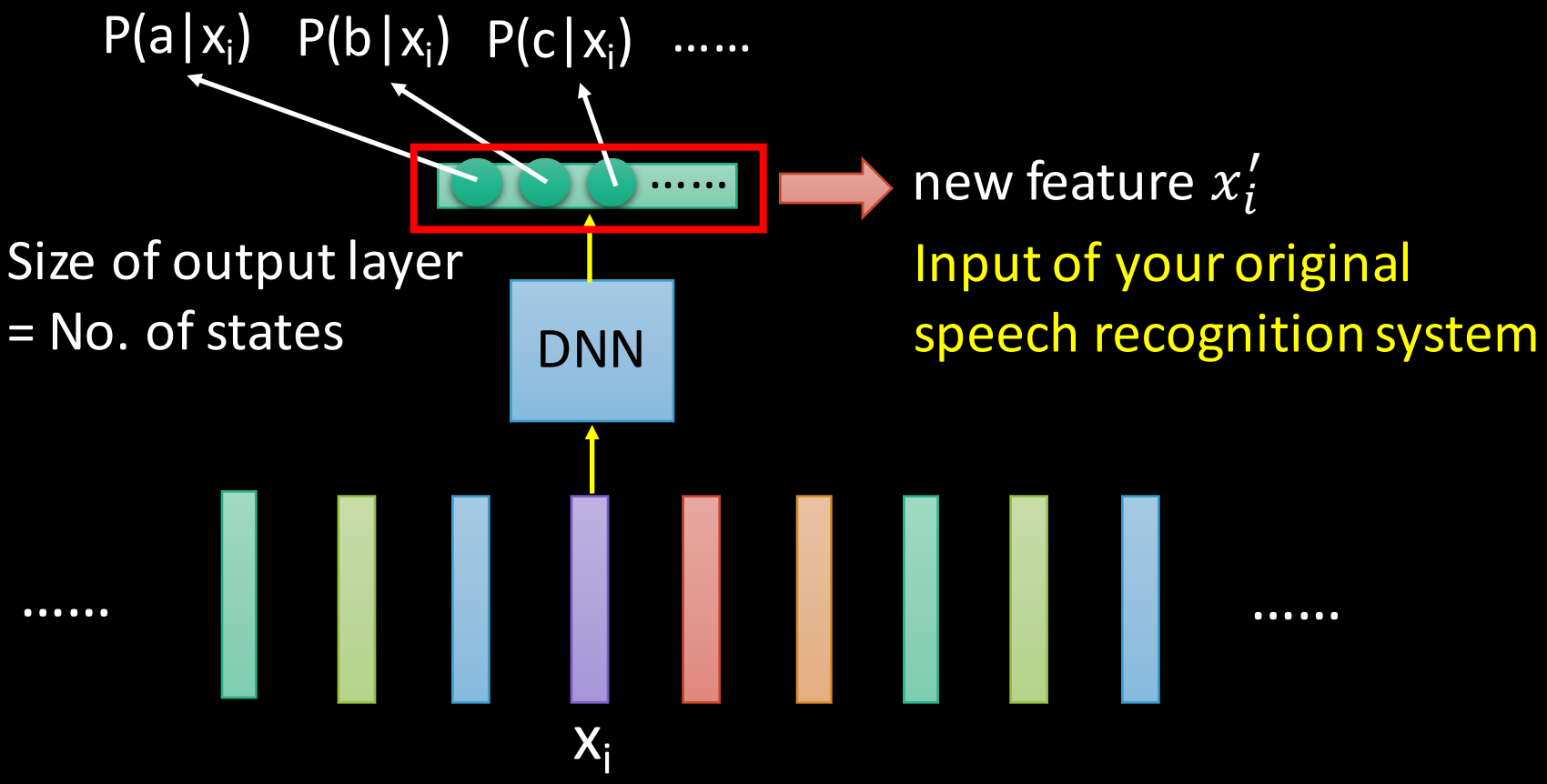
- There are three ways to use DNN for acoustic modeling
  - Way 1. Tandem
  - Way 2. DNN-HMM hybrid
  - Way 3. End-to-end

Efforts for exploiting deep learning

How to use Deep Learning?

Way 1: Tandem

Way 1: Tandem system



Last hidden layer or bottleneck layer are also possible.

How to use Deep Learning?

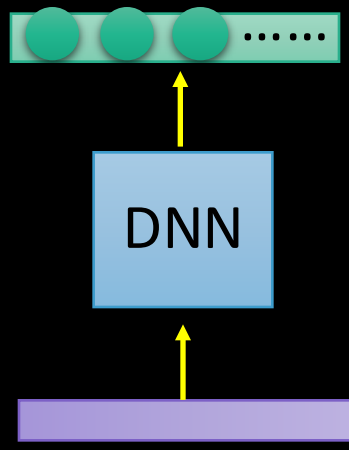
Way 2: DNN-HMM hybrid

Way 2: DNN-HMM Hybrid

$$\tilde{W} = arg \max_{W} P(W|X) = arg \max_{W} P(X|W)P(W)$$

$$P(X|W) \approx \max_{s_1 \cdots s_T} \prod_{t=1}^{T} P(s_t|s_{t-1}) \underline{P(x_t|s_t)}$$

From DNN



$P(s_t|x_t)$

$$P(x_t|s_t) = \frac{P(x_t, s_t)}{P(s_t)}$$

$$= \frac{P(s_t|x_t)P(x_t)}{P(s_t)}$$

Count from training data

Way 2: DNN-HMM Hybrid

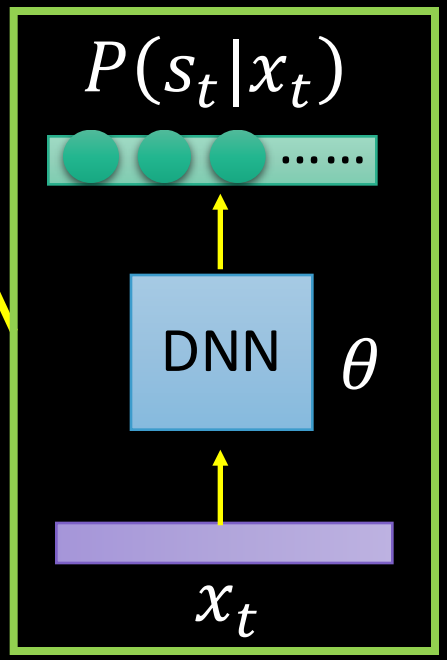
$$P(X|W) \approx \max_{s_1 \cdots s_T} \prod_{t=1}^{T} P(s_t|s_{t-1})\, P(x_t|s_t)$$

$$\cancel{P(X|W)} \approx \max_{s_1 \cdots s_T} \prod_{t=1}^{T} P(s_t|s_{t-1}) \frac{P(s_t|x_t)}{P(s_t)}$$

$$P(X|W;\theta)$$

From original HMM

Count from training data



This assembled vehicle works …….

Way 2: DNN-HMM Hybrid

- ***Sequential Training***

$$\widetilde{W} = arg\max_{W} P(X|W;\theta)P(W)$$

Given training data $(X_1, \widehat{W}_1), (X_2, \widehat{W}_2), \cdots (X_r, \widehat{W}_r), \cdots$

Find-tune the DNN parameters θ such that

$P(X_r|\widehat{W}_r;\theta)P(\widehat{W}_r)$ ⇒ increase

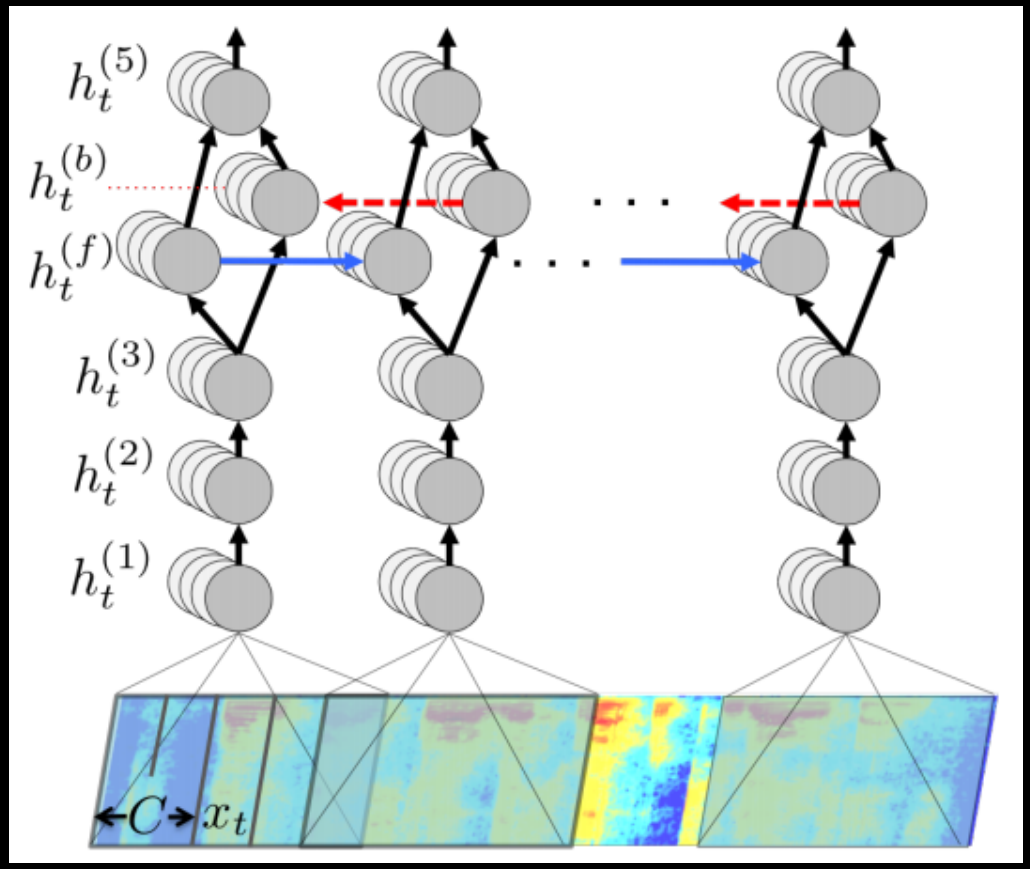
$P(X_r|W;\theta)P(W)$ ⇒ decrease

(W is any word sequence different from $\widehat{W}_r$)

How to use Deep Learning?

Way 3: End-to-end

Way 3: End-to-end - Character



Input: acoustic features (spectrograms)

Output: characters (and space)
+ null (~)

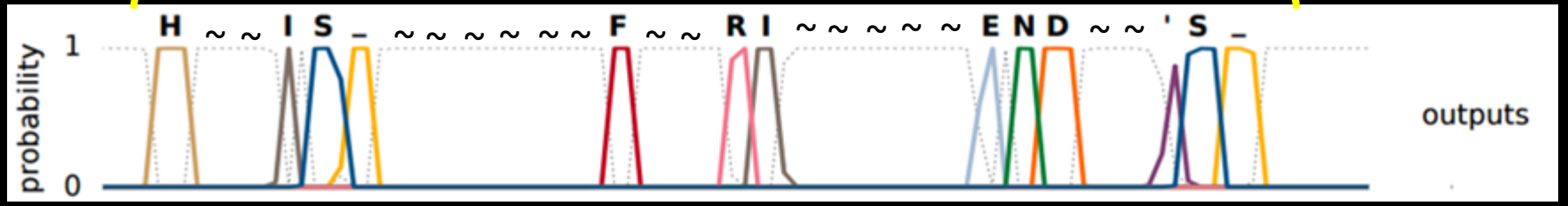
No phoneme and lexicon

(No OOV problem)

A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, A. Ng "Deep Speech: Scaling up end-to-end speech recognition", arXiv:1412.5567v2, 2014.

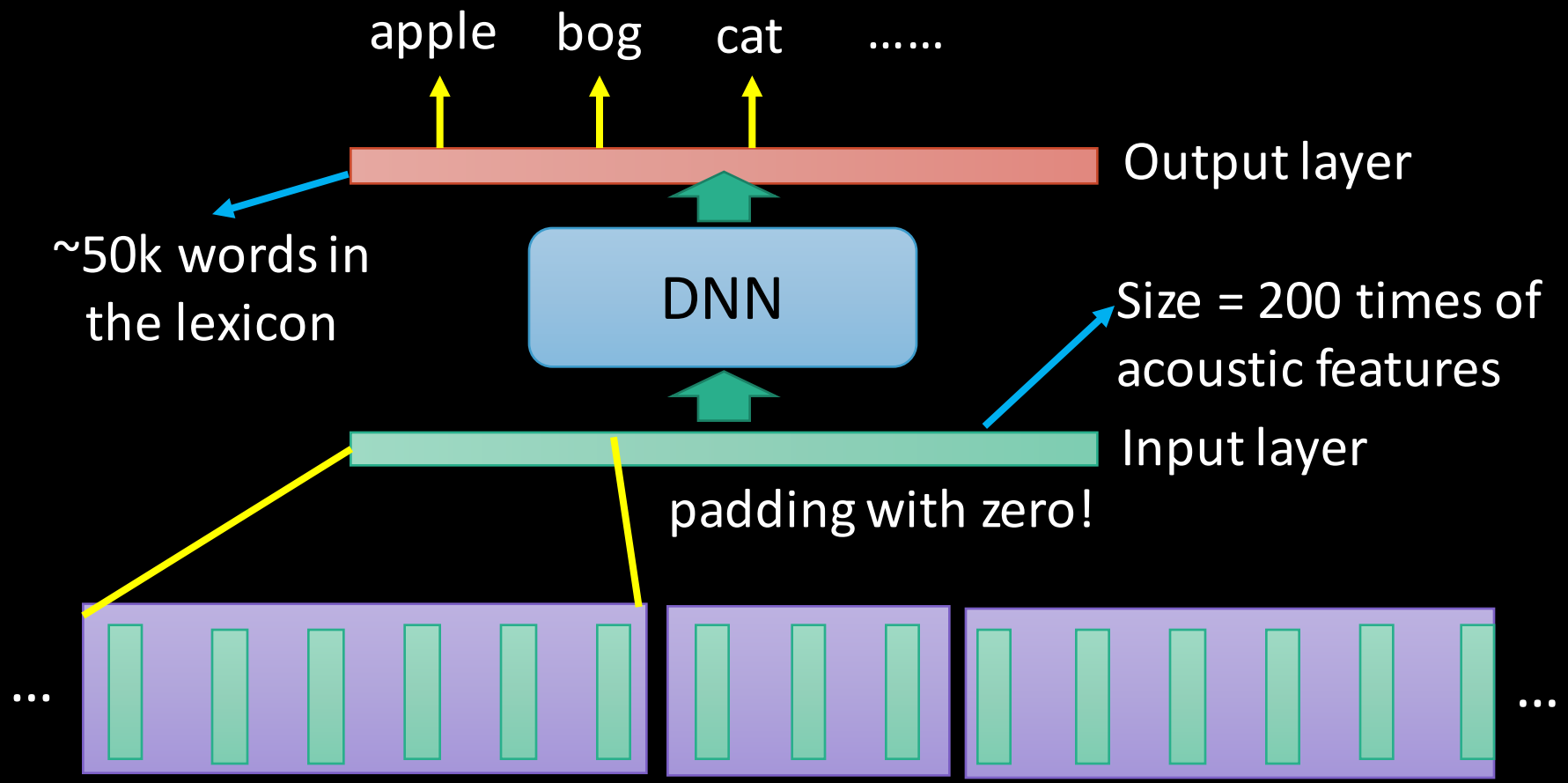
Way 3: End-to-end - Character

HIS FRIEND'S



Graves, Alex, and Navdeep Jaitly. "Towards end-to-end speech recognition with recurrent neural networks." *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. 2014.

Way 3: End-to-end – Word?



Use other systems to get the word boundaries

Ref: Bengio, Samy, and Georg Heigold. "Word embeddings for speech recognition.“, *Interspeech*. 2014.

Why Deep Learning?

Deeper is better?

Seide, Frank, Gang Li, and Dong Yu. "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks." *Interspeech*. 2011.

- Word error rate (WER)

multiple layers

Deeper is Better

LxN	DBN-PT (%)
1×2k	24.2
2×2k	20.4
3×2k	18.4
4×2k	17.8
5×2k	17.2
7×2k	17.1

Deeper is better?

Seide, Frank, Gang Li, and Dong Yu. "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks." *Interspeech*. 2011.

- Word error rate (WER)

multiple layers | 1 hidden layer

Deeper is Better

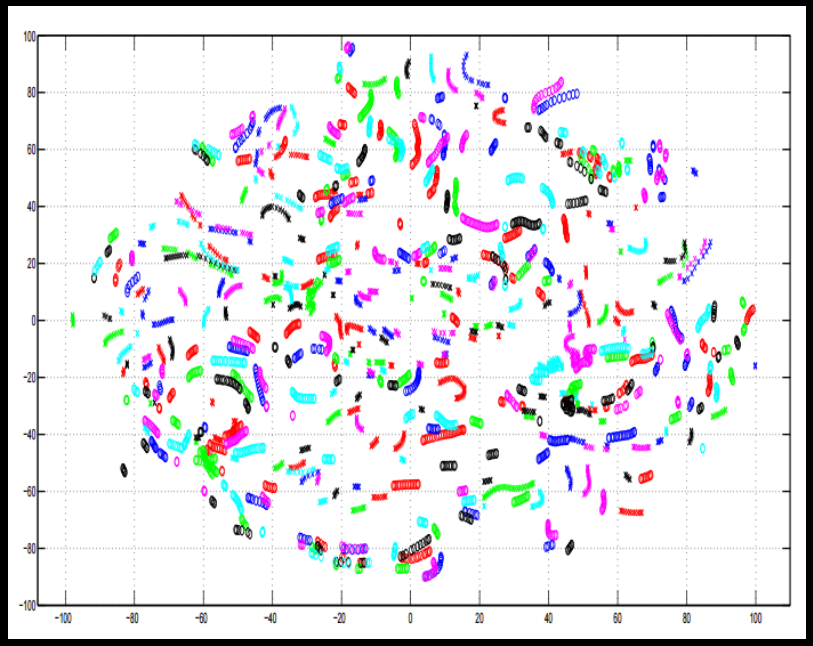
LxN	DBN-PT (%)	1xN	DBN-PT (%)
1×2k	24.2		
2×2k	20.4		
3×2k	18.4		
4×2k	17.8		
5×2k	17.2	1×3,772	22.5
7×2k	17.1	1×4,634	22.6
		1×16K	22.1

For a fixed number of parameters, a deep model is clearly better than the shallow one.

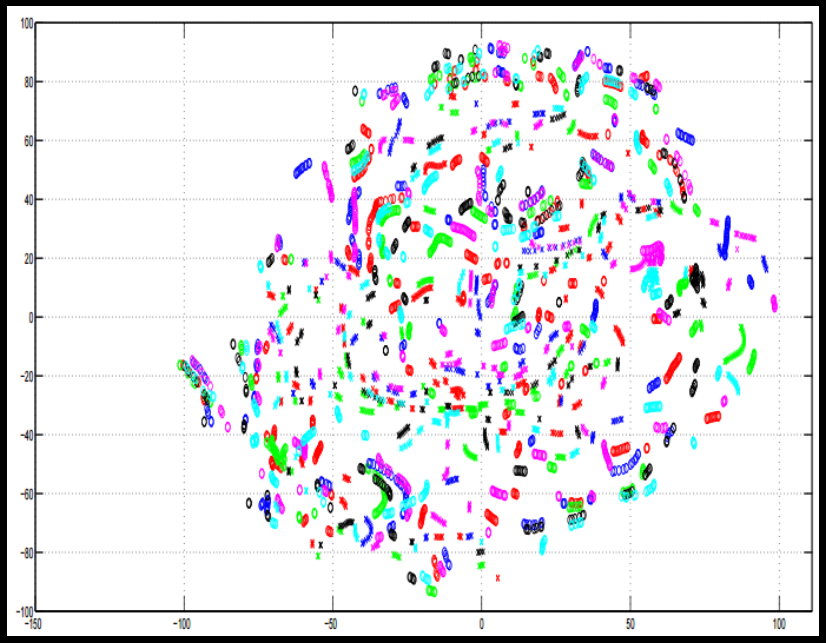
What does DNN do?

A. Mohamed, G. Hinton, and G. Penn, "Understanding how Deep Belief Networks Perform Acoustic Modelling," in ICASSP, 2012.

- Speaker normalization is automatically done in DNN



Input Acoustic Feature (MFCC)

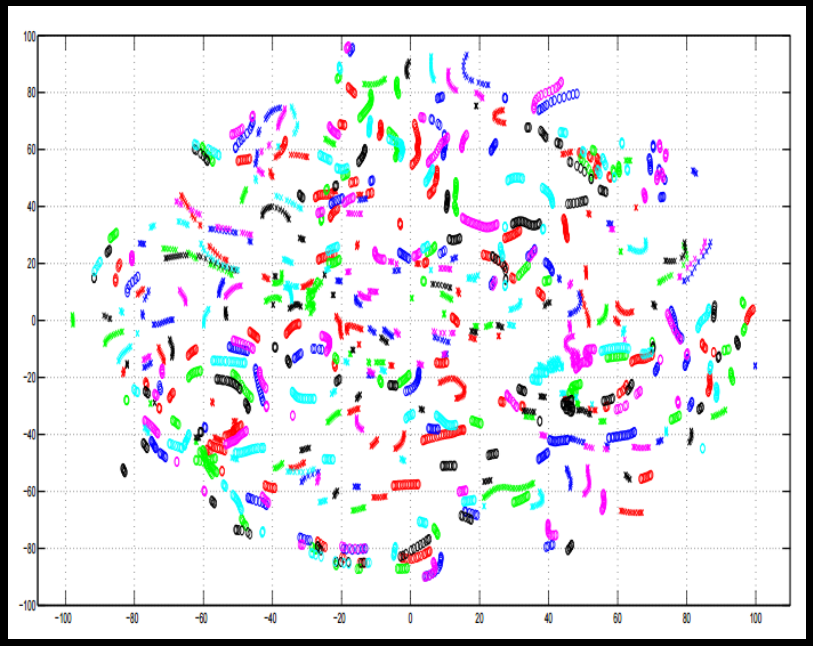


1-st Hidden Layer

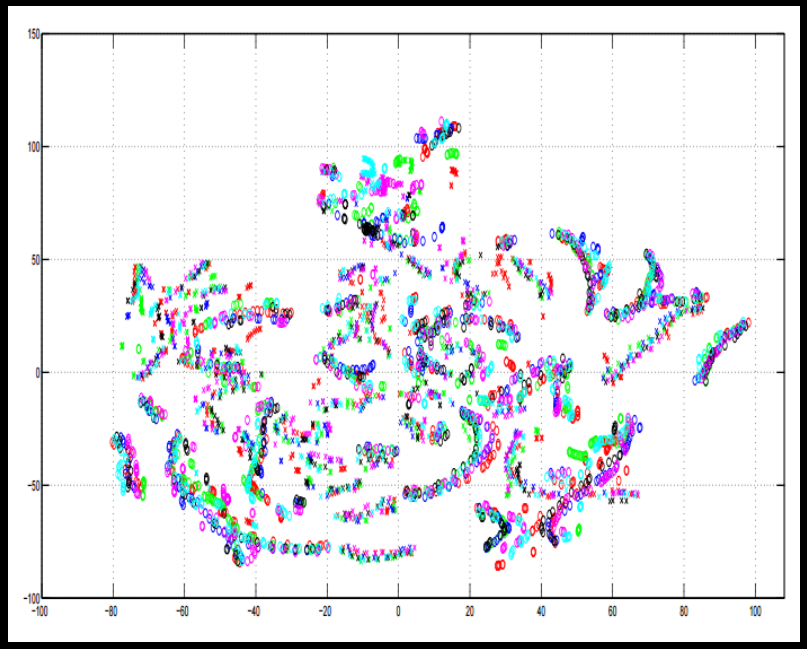
What does DNN do?

A. Mohamed, G. Hinton, and G. Penn, "Understanding how Deep Belief Networks Perform Acoustic Modelling," in ICASSP, 2012.

- Speaker normalization is automatically done in DNN



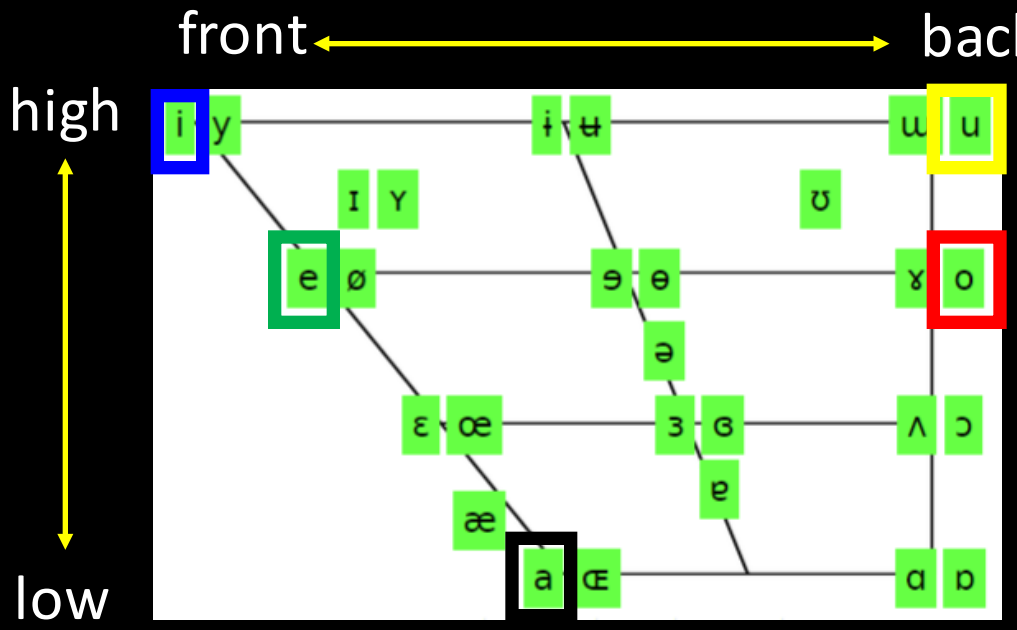
Input Acoustic Feature (MFCC)



8-th Hidden Layer

What does DNN do?

- In ordinary acoustic models, all the states are modeled independently
  - Not effective way to model human voice



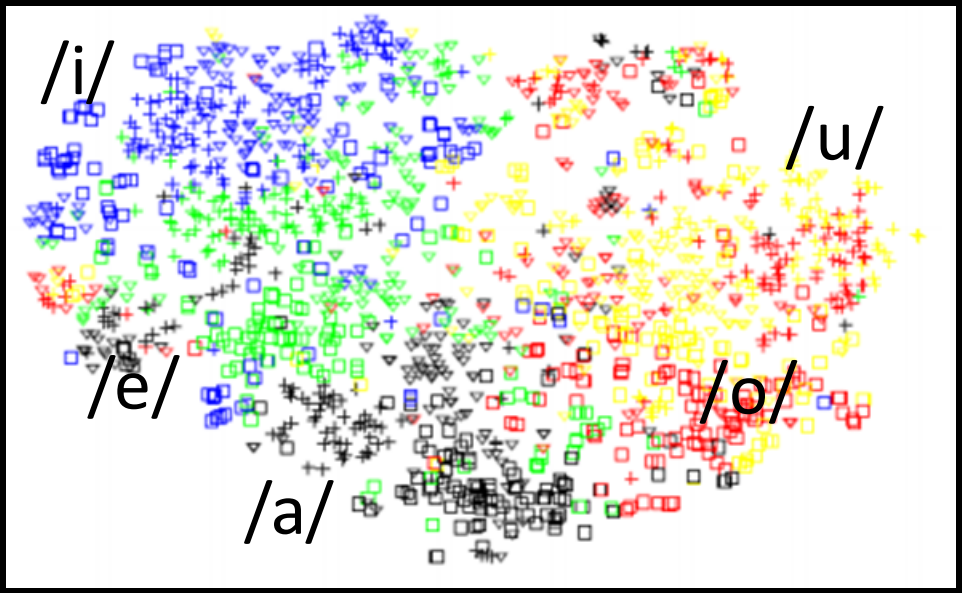
The sound of vowel is only controlled by a few factors.

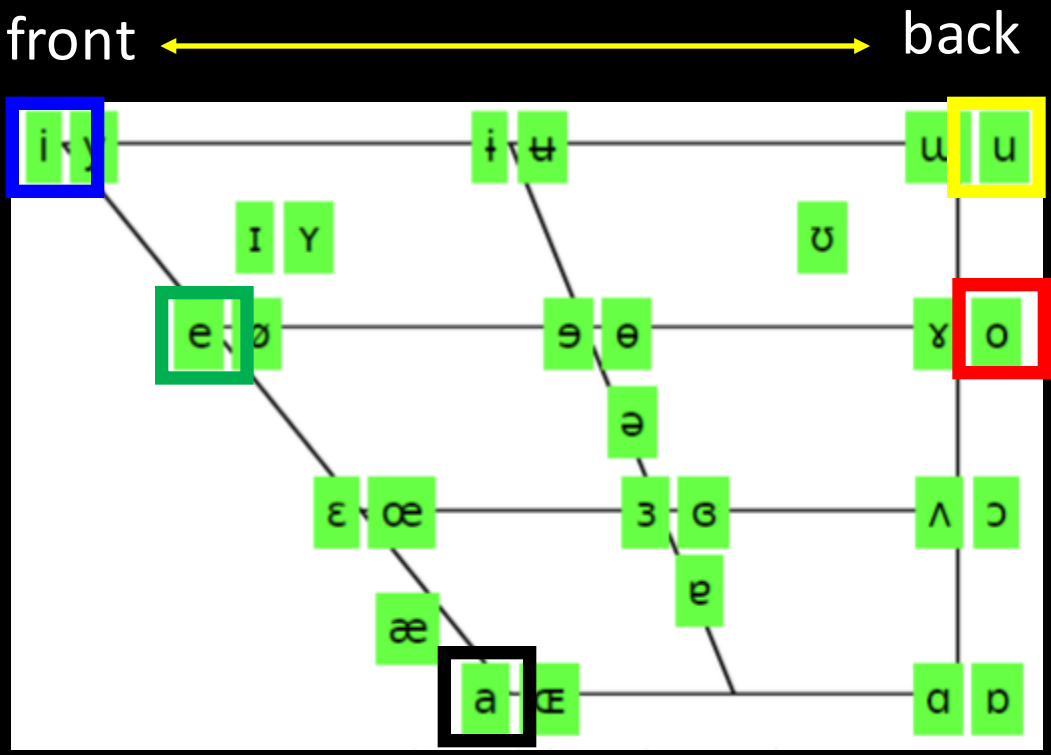
http://www.ipachart.com/

What does DNN do?

Vu, Ngoc Thang, Jochen Weiner, and Tanja Schultz. "Investigating the Learning Effect of Multilingual Bottle-Neck Features for ASR." *Interspeech*. 2014.

Output of hidden layer reduce to two dimensions





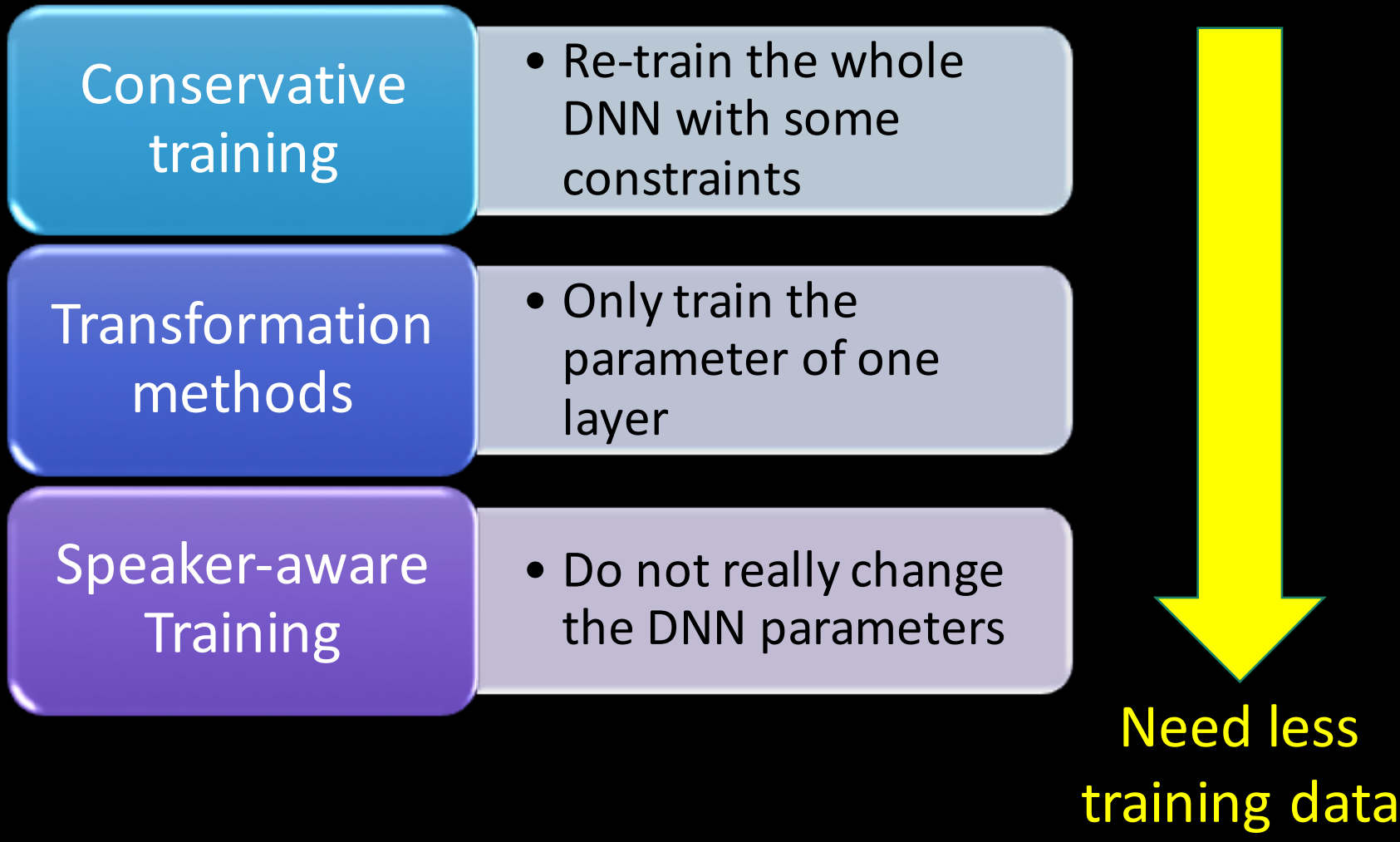
- The lower layers detect the manner of articulation
- All the states share the results from the same set of detectors.
- Use parameters effectively

Speaker Adaptation

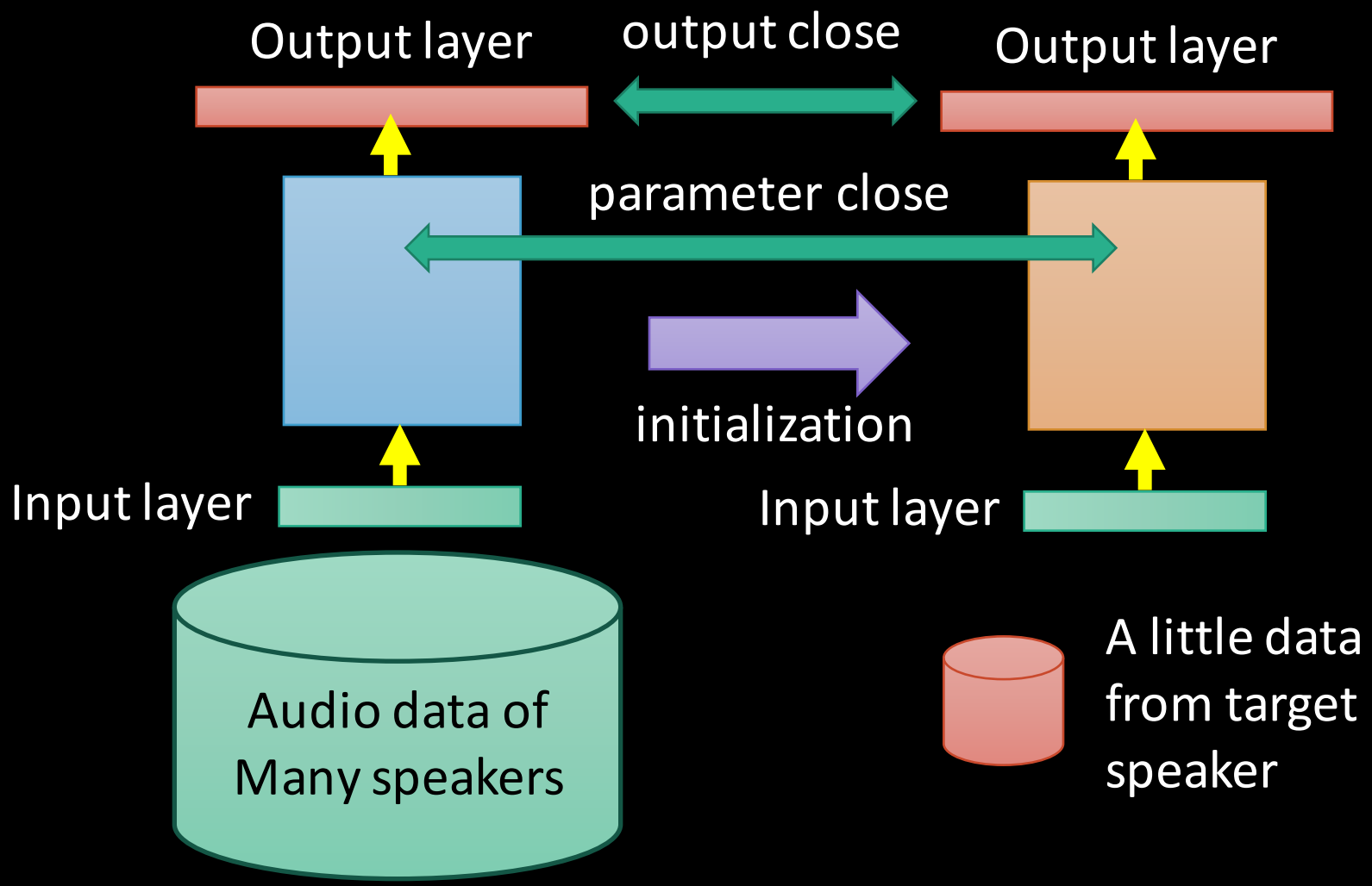
Speaker Adaptation

- Speaker adaptation: use different models to recognition the speech of different speakers
  - Collect the audio data of each speaker
- A DNN model for each speaker
  - Challenge: limited data for training
    - Not enough data for directly training a DNN model
    - Not enough data for just fine-tune a speaker independent DNN model

Categories of Methods

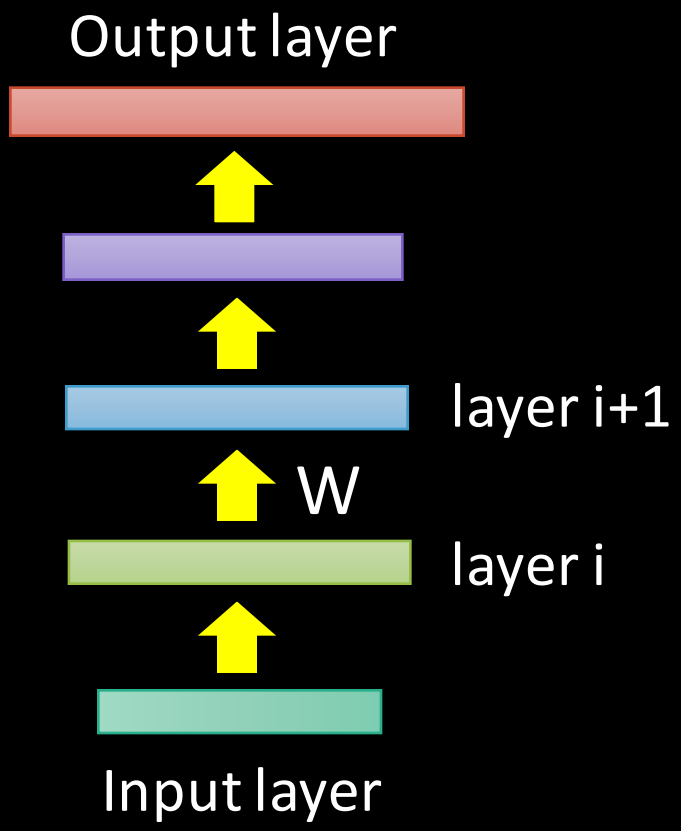


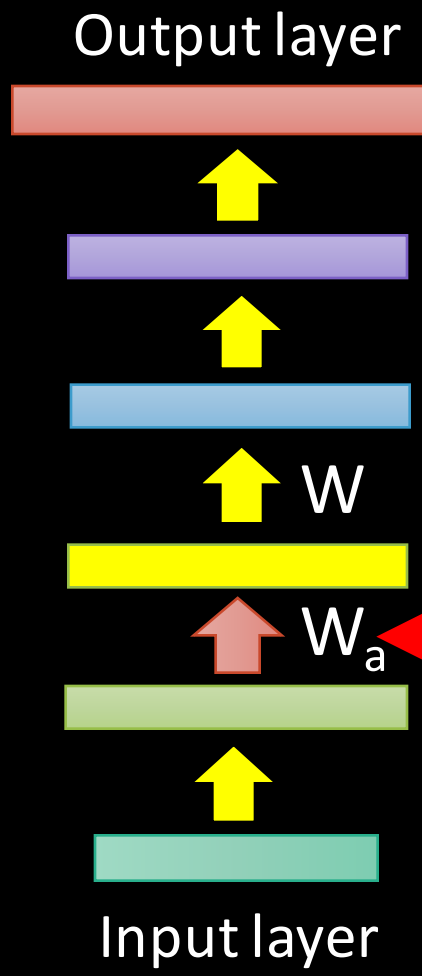
Conservative Training



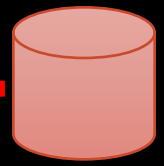
Transformation methods

Add an extra layer





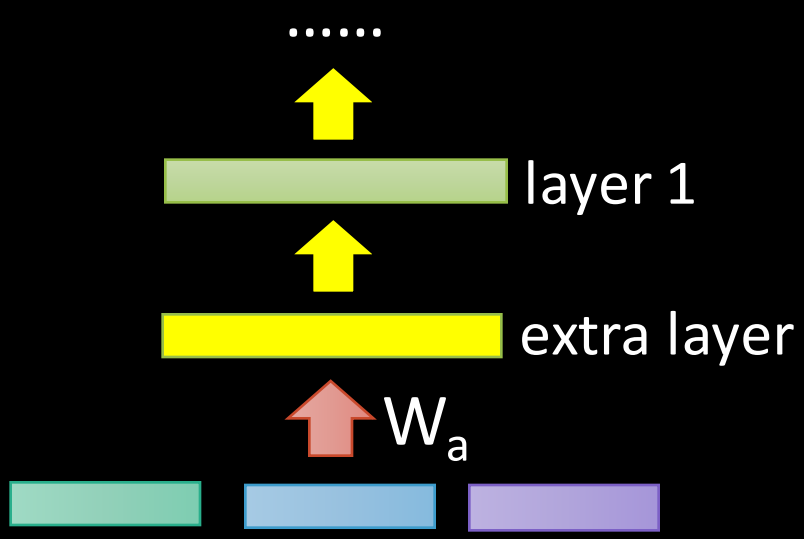
Fix all the other parameters

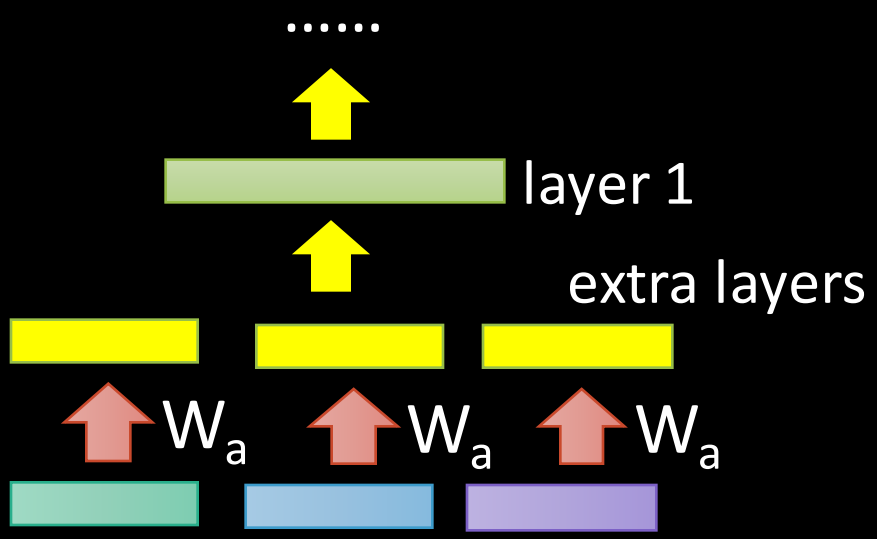


A little data from target speaker

Transformation methods

- Add the extra layer between the input and first layer
- With splicing



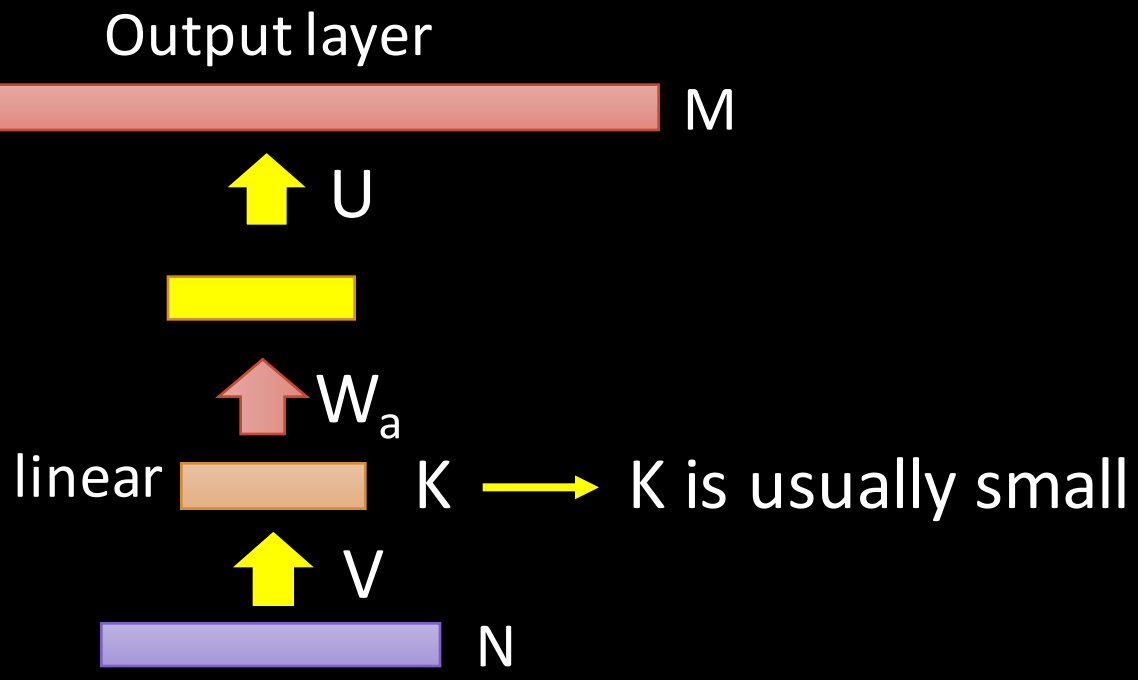


Larger W_a → More data

Smaller W_a → less data

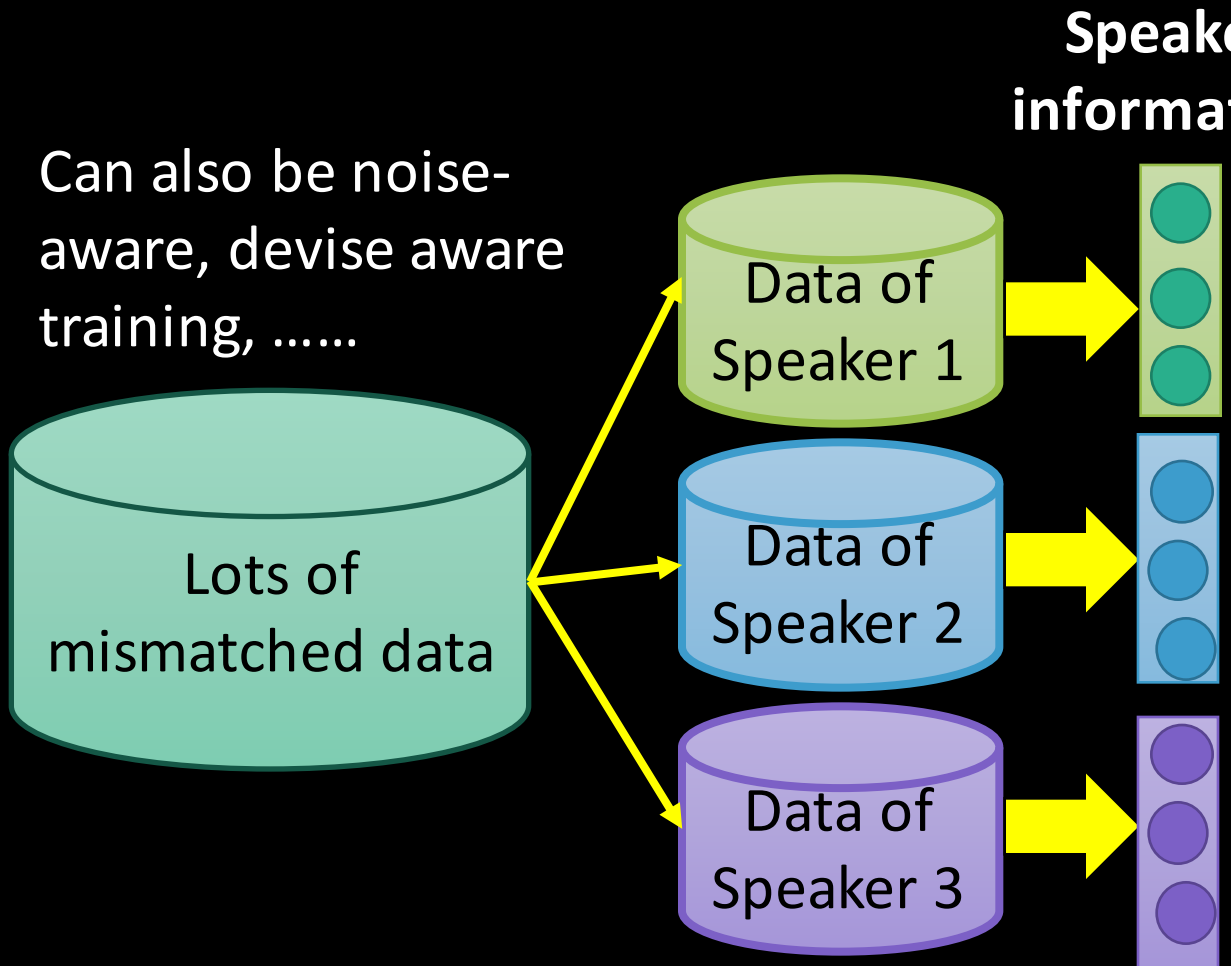
Transformation methods

- SVD bottleneck adaptation



Speaker-aware Training

Can also be noise-aware, devise aware training, ……

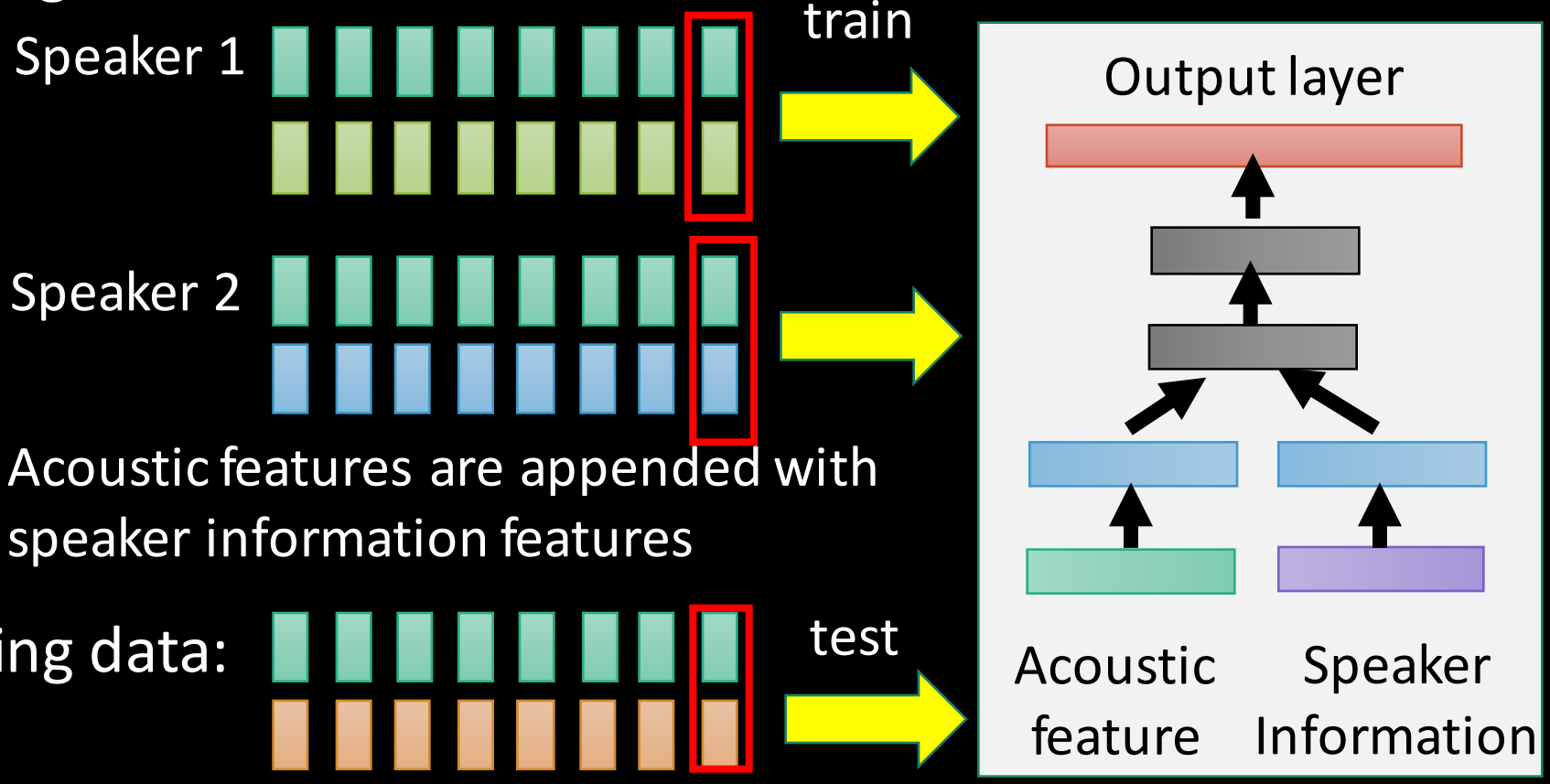


Fixed length low dimension vectors

Text transcription is not needed for extracting the vectors.

Speaker-aware Training

Training data:



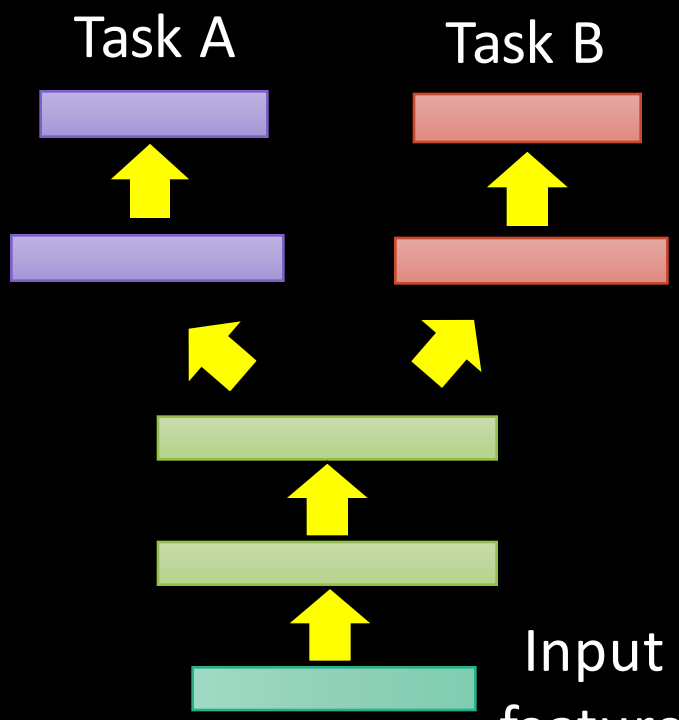
All the speaker use the same DNN model

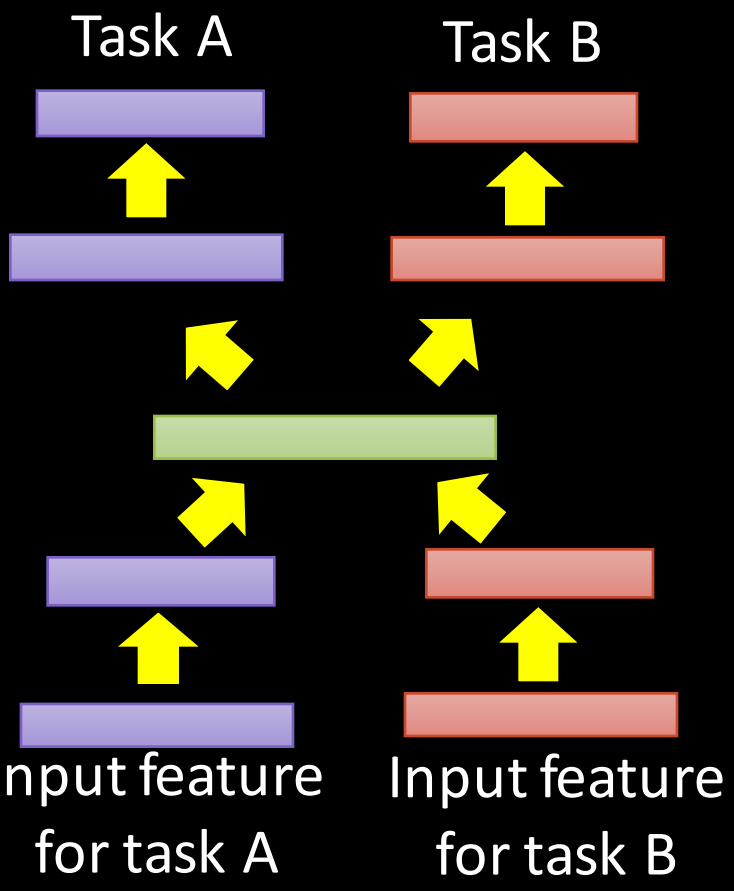
Different speaker augmented by different features

Multi-task Learning

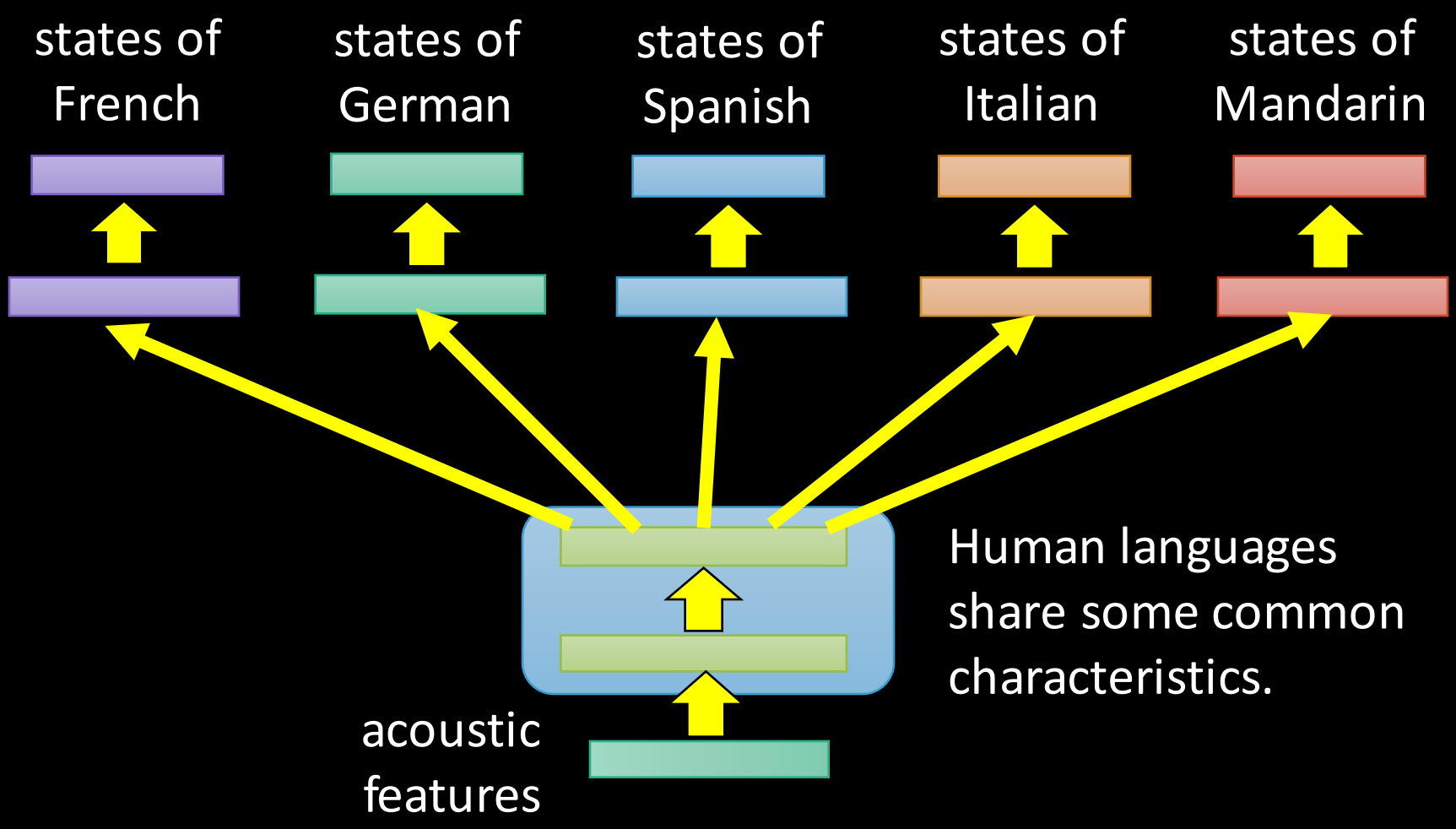
Multitask Learning

- The multi-layer structure makes DNN suitable for multitask learning

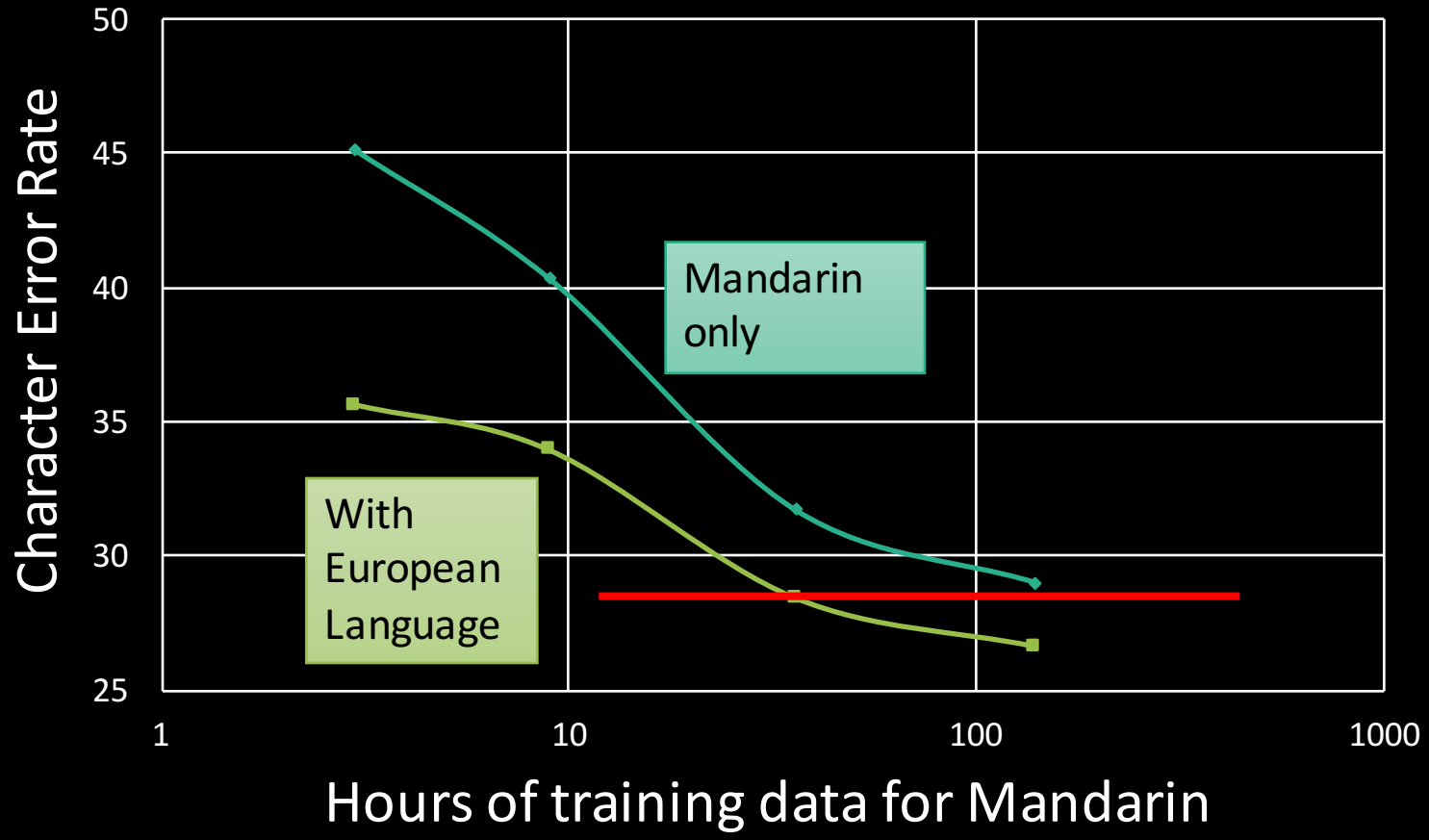




Multitask Learning - Multilingual

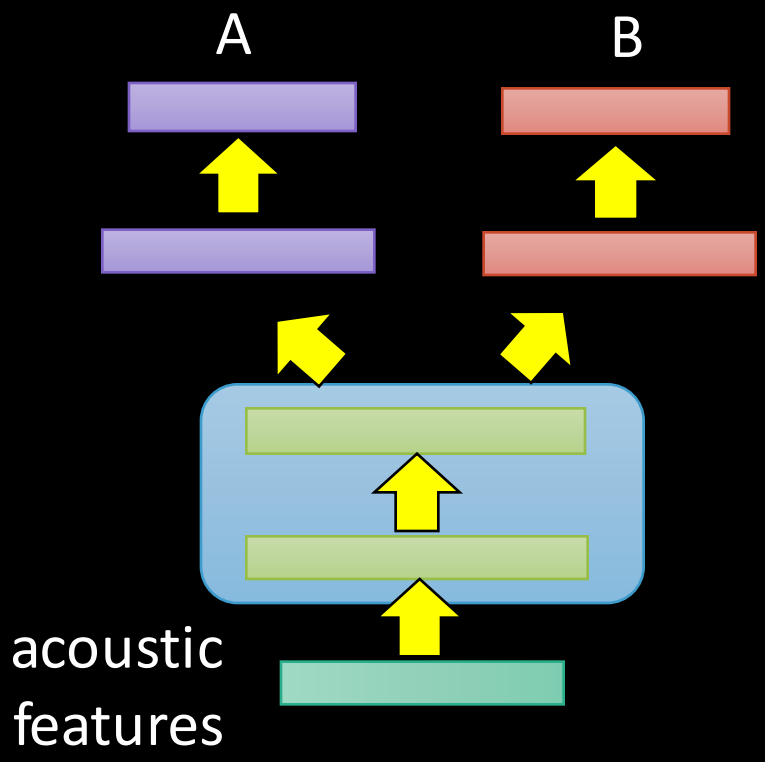


Multitask Learning - Multilingual



Huang, Jui-Ting, et al. "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers." *Acoustics, Speech and Signal Processing (ICASSP), 2013*

Multitask Learning - Different units



A= state B = phoneme

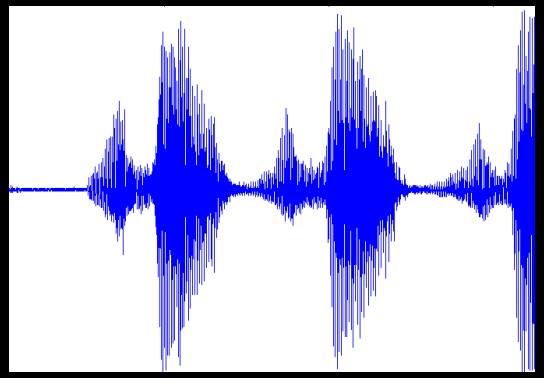
A= state B = gender

A= state B = grapheme (character)

Deep Learning for Acoustic Modeling

New acoustic features

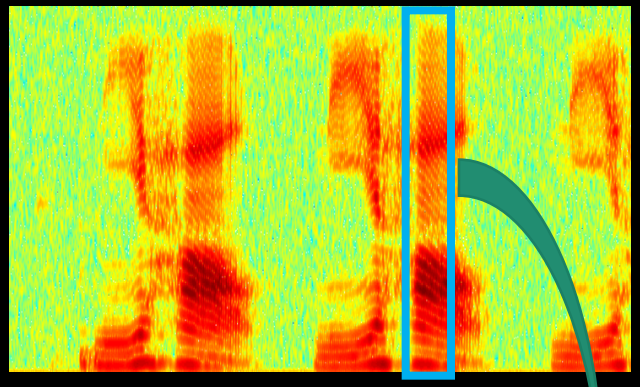
MFCC



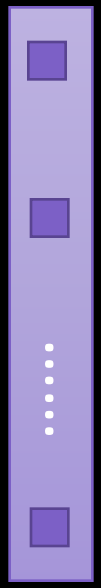
Waveform

DFT

spectrogram



filter bank



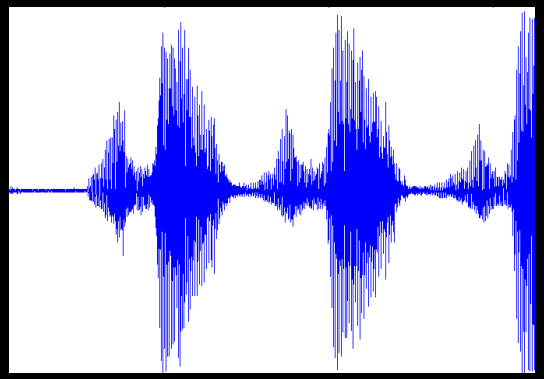
log

DCT

MFCC

Input of DNN

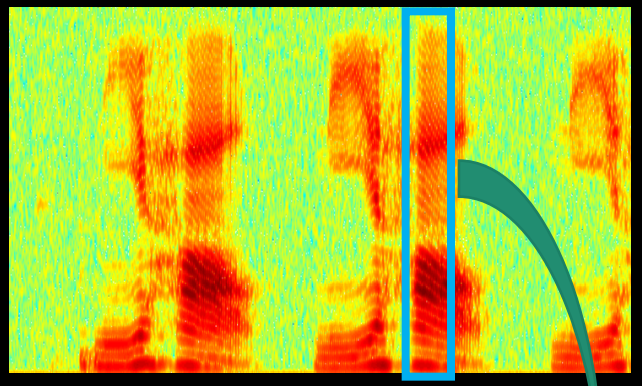
Filter-bank Output

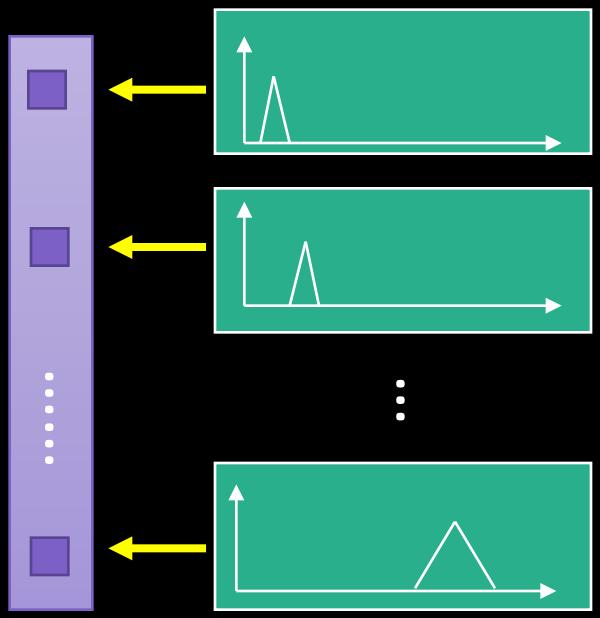


Waveform

DFT

spectrogram





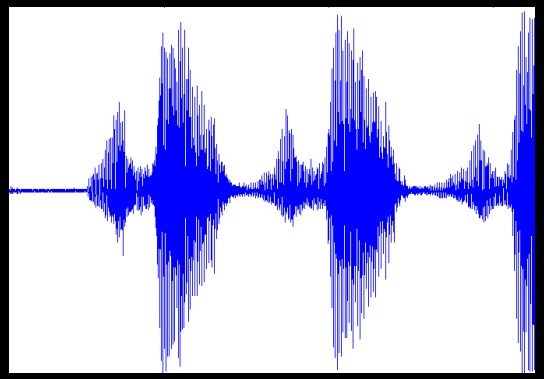
filter bank

log

Input of DNN

- Kind of standard now

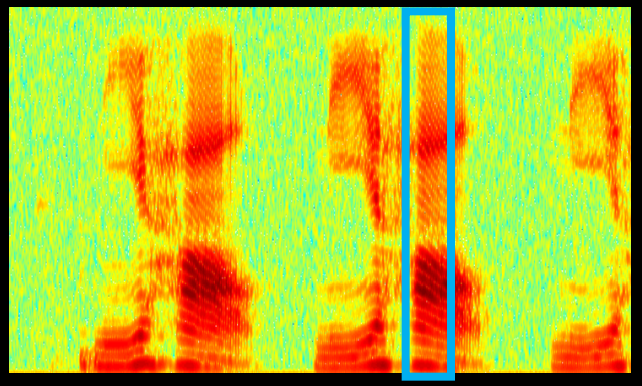
Spectrogram



Waveform

DFT

spectrogram

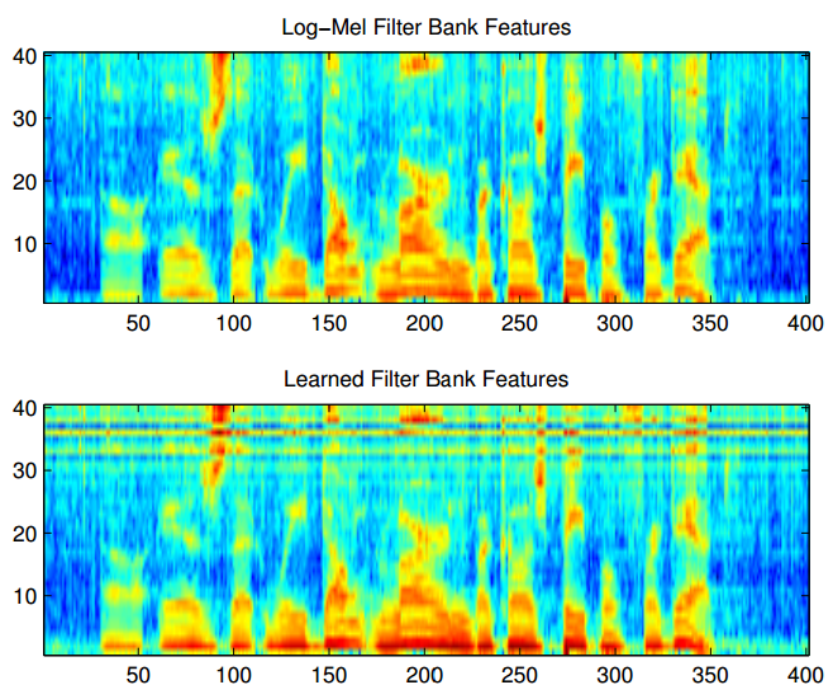


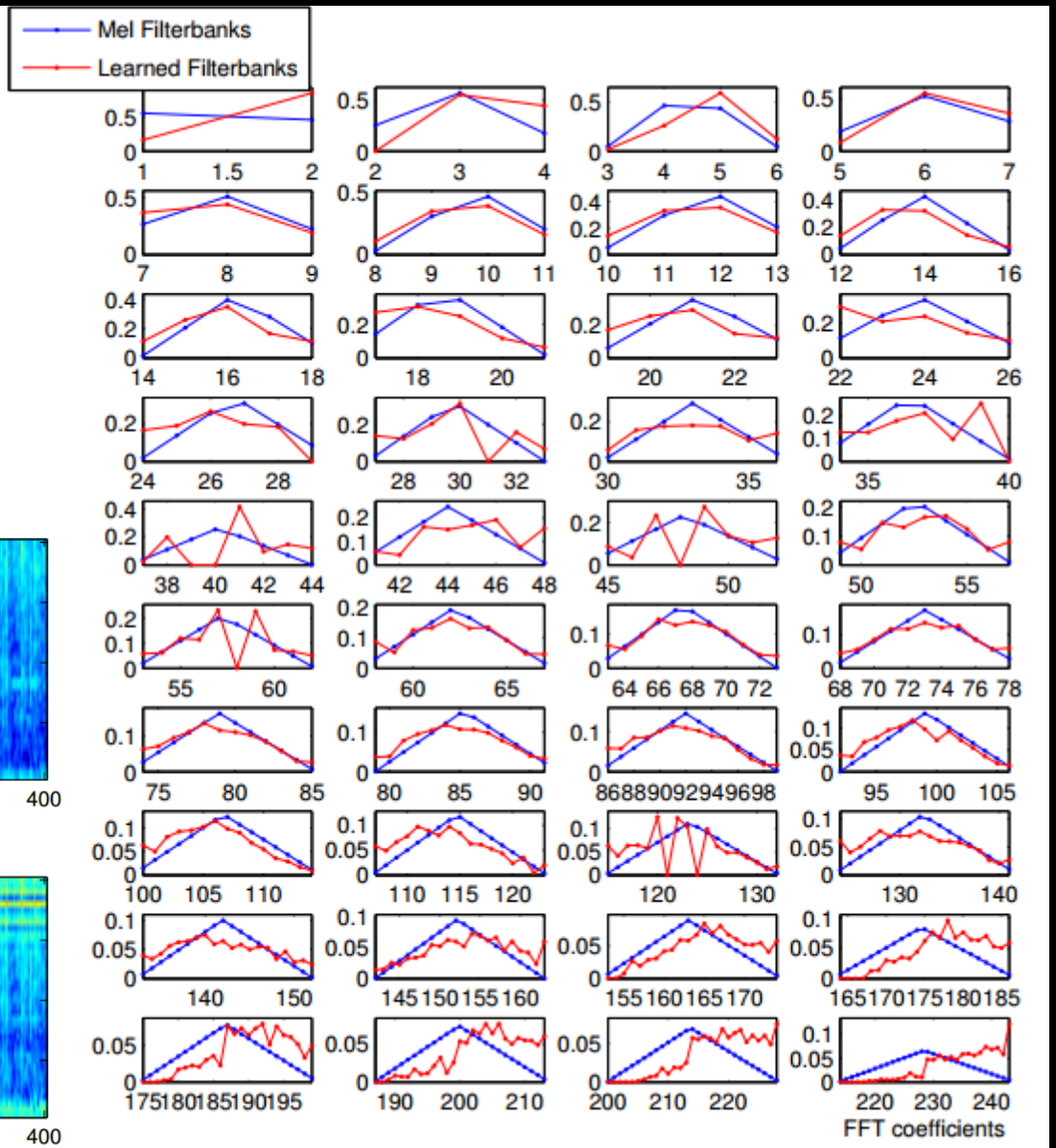
- common today

Input of DNN

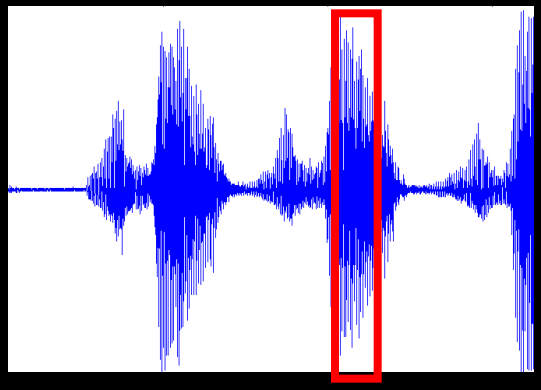
- 5% relative improvement over fitlerbank output
- Ref: ainath, T. N., Kingsbury, B., Mohamed, A. R., & Ramabhadran, B., "Learning filter banks within a deep neural network framework," In *Automatic Speech Recognition and Understanding (ASRU), 2013*

Spectrogram





Waveform?



Waveform

Input of DNN

- If success, no Signal & Systems ☺

- People tried, but not better than spectrogram yet

- Ref: Tüske, Z., Golik, P., Schlüter, R., & Ney, H., "Acoustic modeling with deep neural networks using raw time signal for LVCSR," In *INTERPSEECH 2014*

- Still need to take Signal & Systems …… ☺

Waveform?

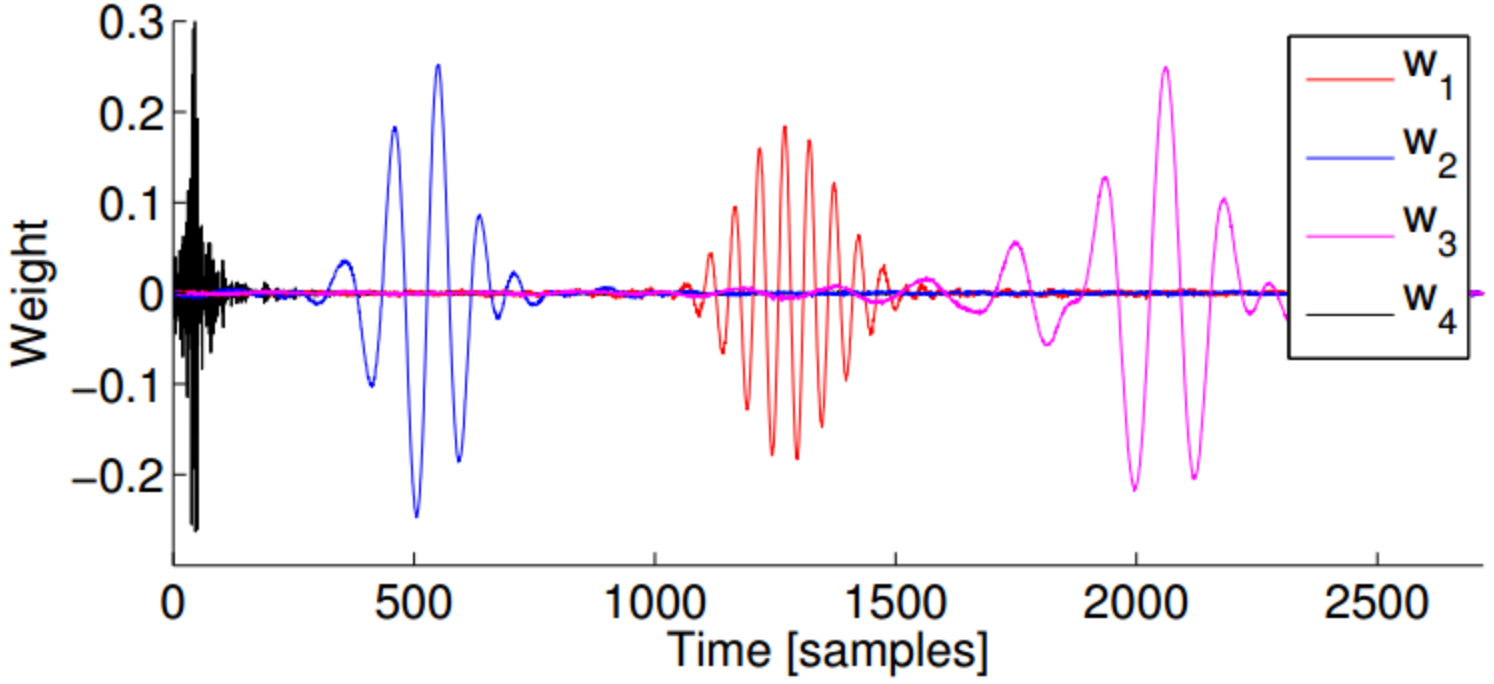
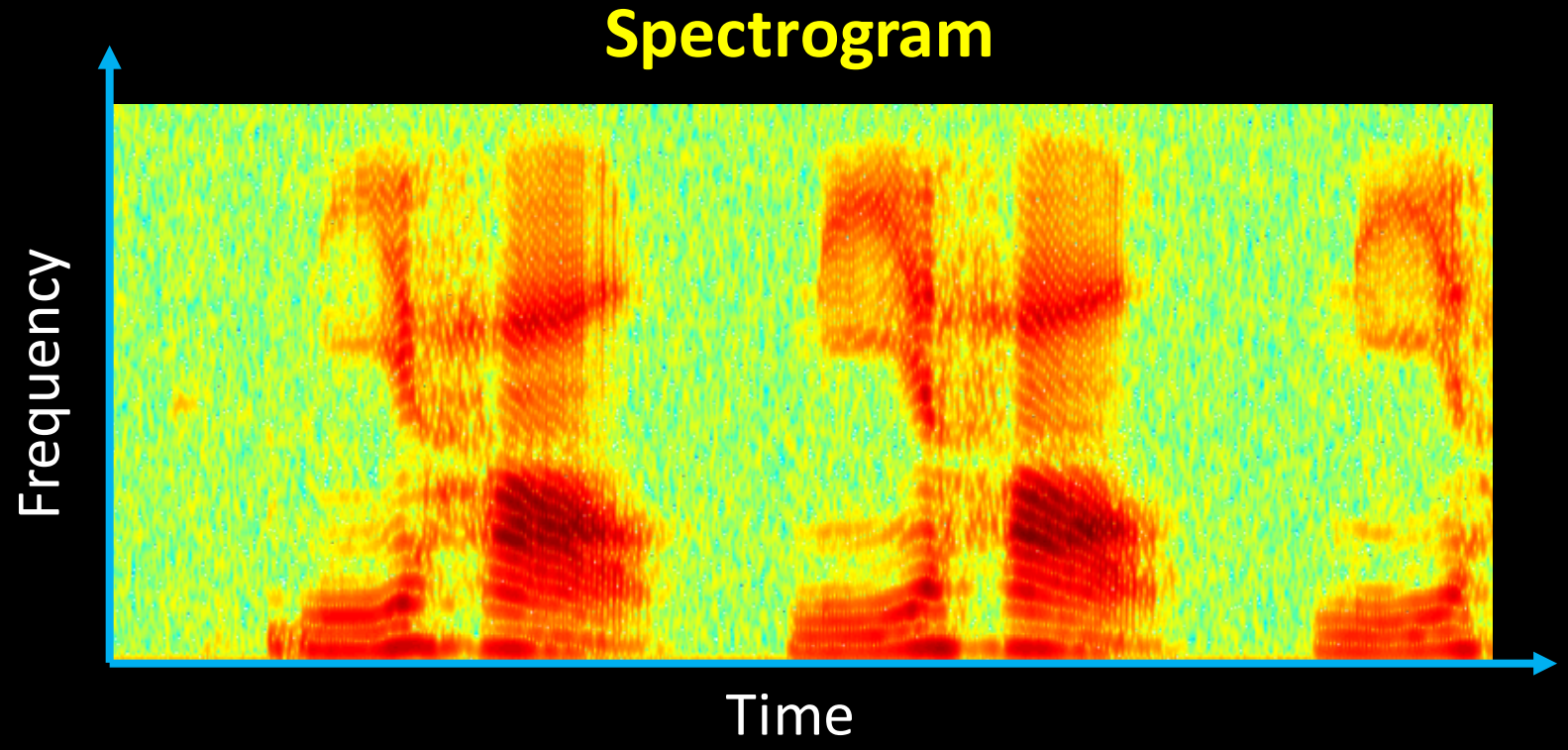


Figure 1: *Four rows from the first layer weight matrix trained on raw time signal. The time range corresponds to 17 frames of 10 ms* ($17 \cdot 10ms \cdot 16kHz = 2720$)

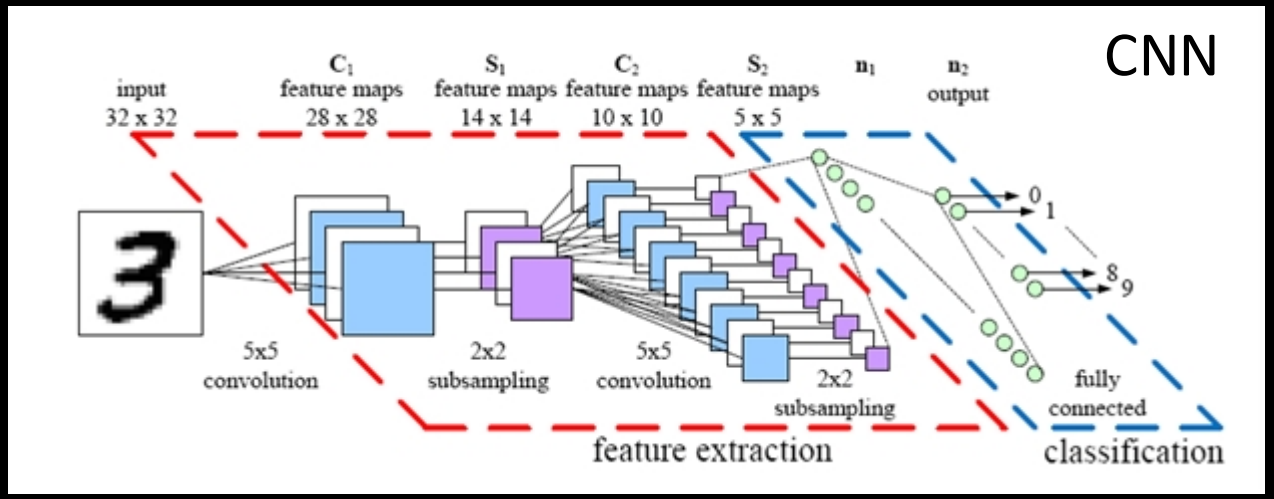
Convolutional Neural Network (CNN)

CNN

- Speech can be treated as images



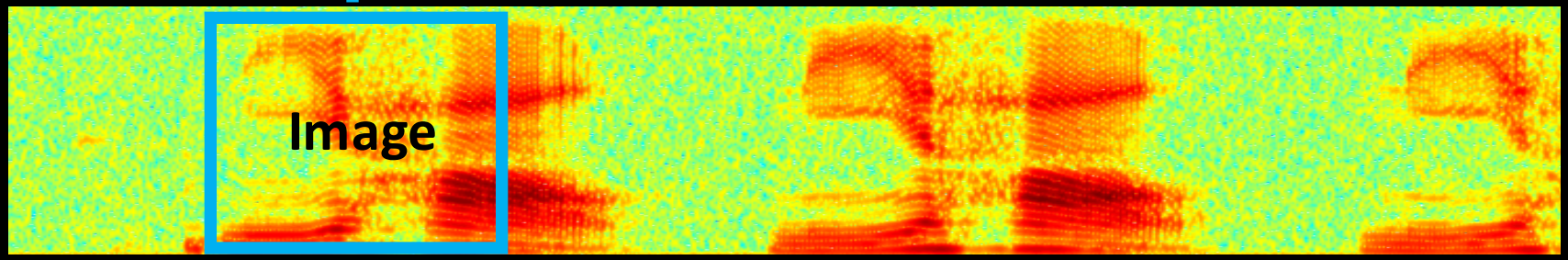
CNN



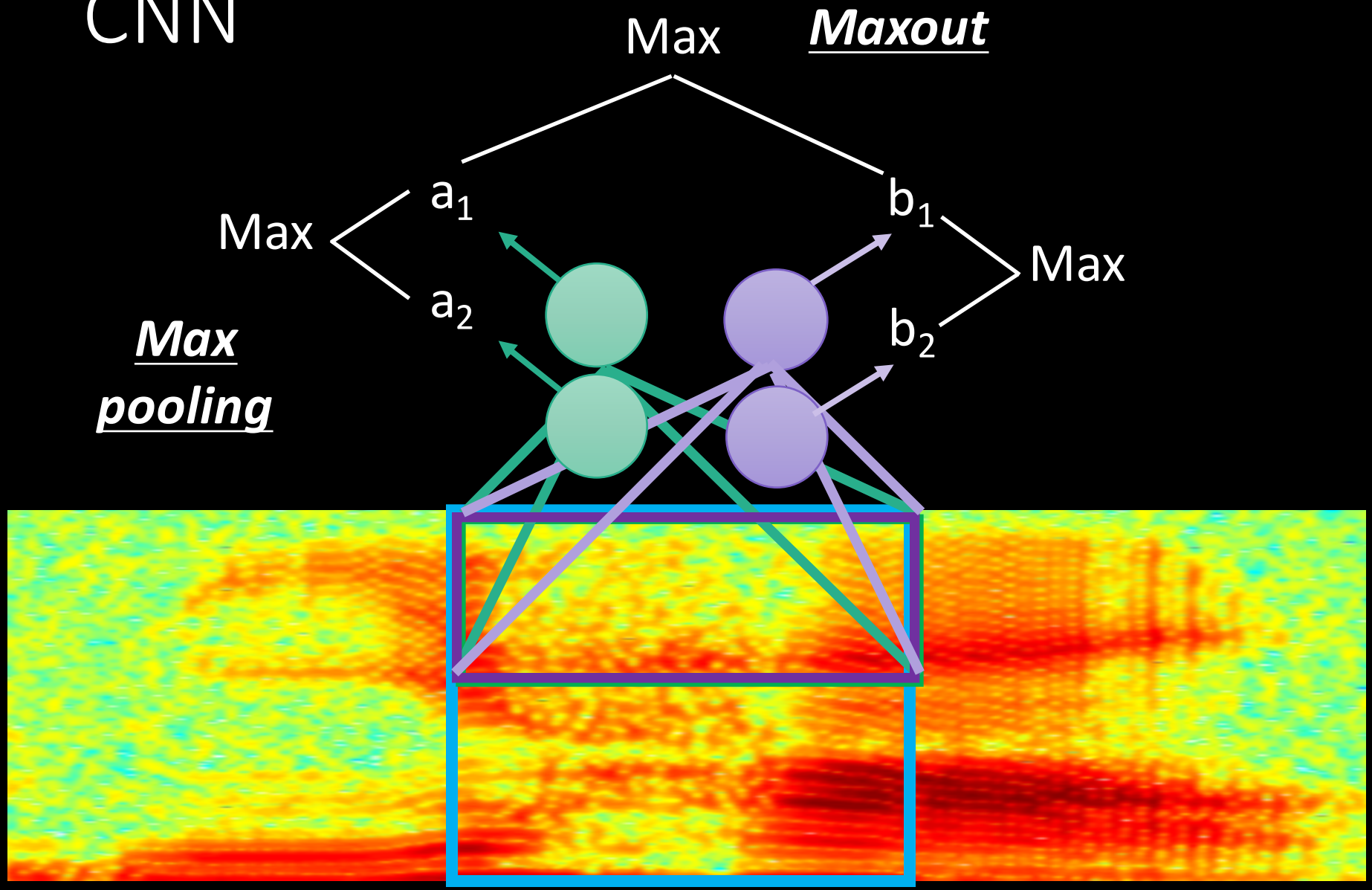
Probabilities of states

CNN

Replace DNN by CNN

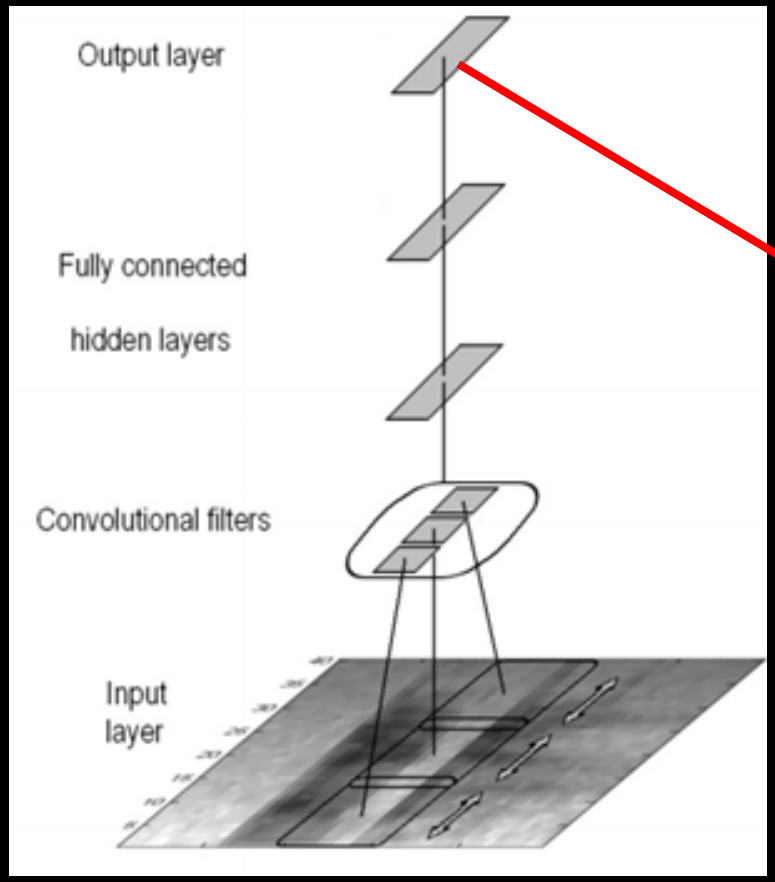


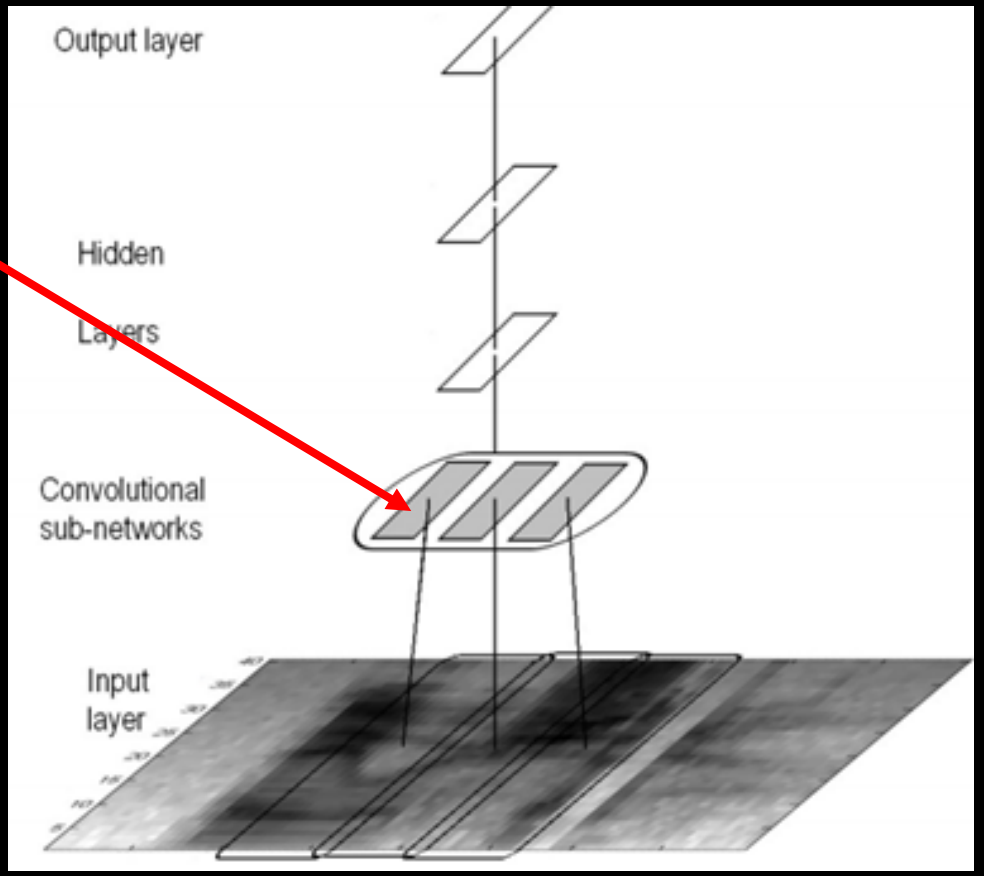
CNN



CNN

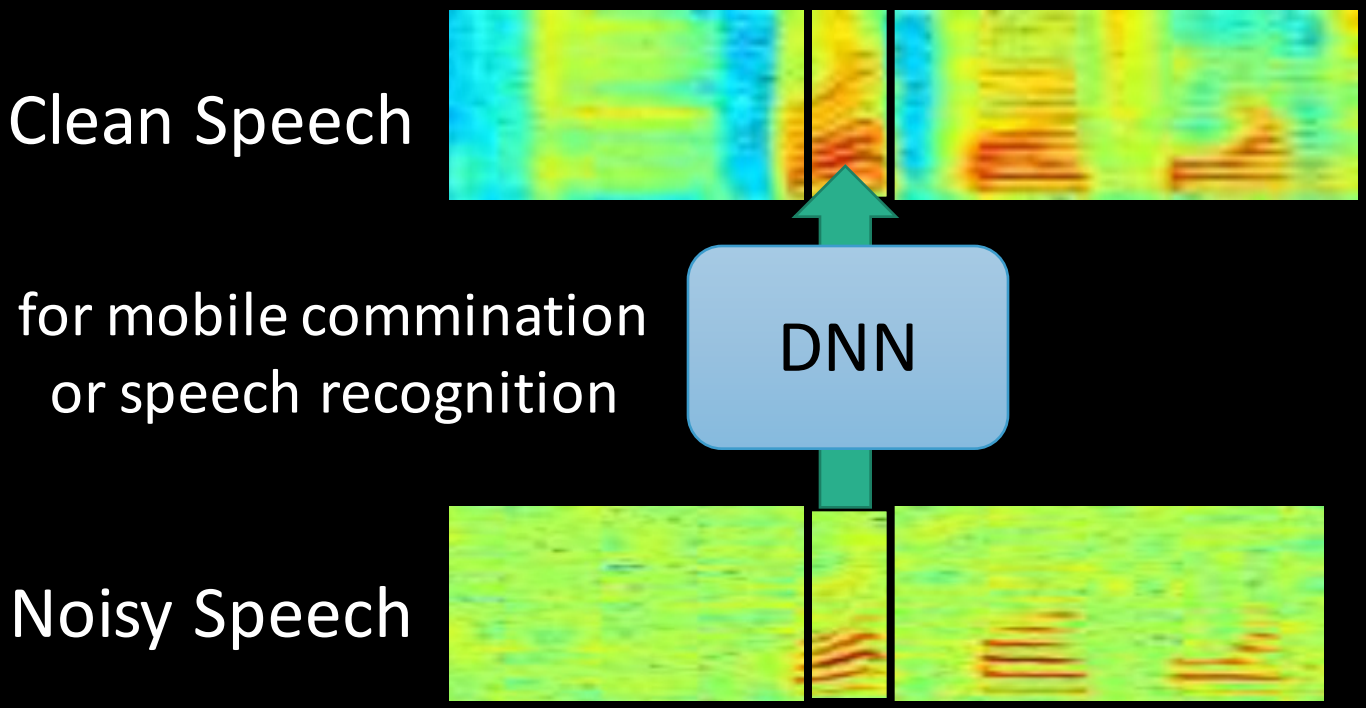
Tóth, László. "Convolutional Deep Maxout Networks for Phone Recognition", Interspeech, 2014.





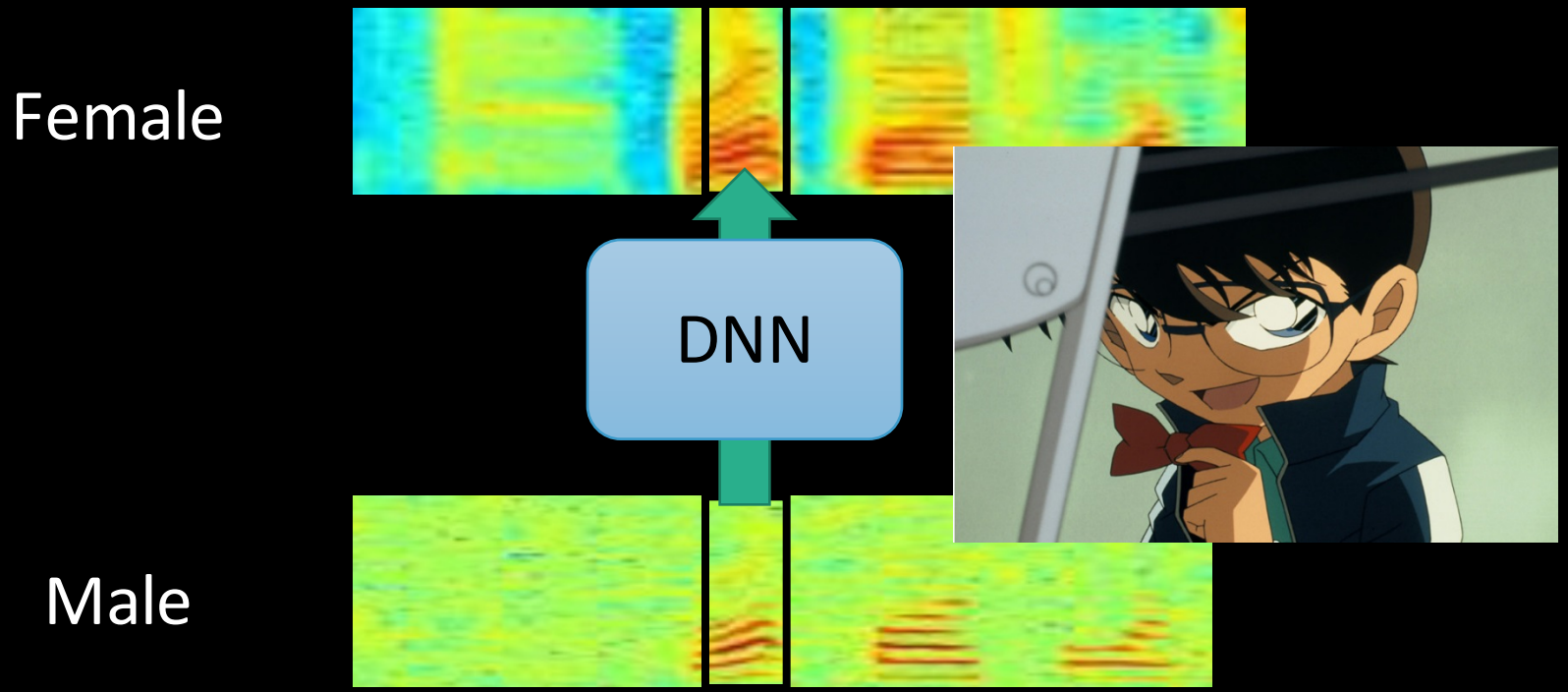
Applications in Acoustic Signal Processing

DNN for Speech Enhancement



- Demo for speech enhancement: http://home.ustc.edu.cn/~xuyong62/demo/SE_DNN.html

DNN for Voice Conversion



- Demo for Voice Conversion: http://research.microsoft.com/en-us/projects/vcnn/default.aspx

Concluding Remarks

Concluding Remarks

- Conventional Speech Recognition
- How to use Deep Learning in acoustic modeling?
- Why Deep Learning?
- Speaker Adaptation
- Multi-task Deep Learning
- New acoustic features
- Convolutional Neural Network (CNN)
- Applications in Acoustic Signal Processing

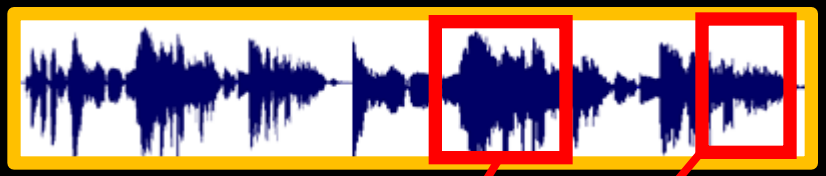
Thank you for
your attention!

More Researches related to Speech

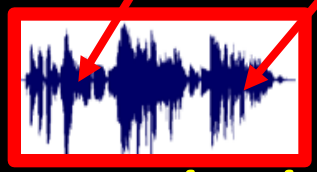
Find the lectures related to "deep learning"

Spoken Content Retrieval

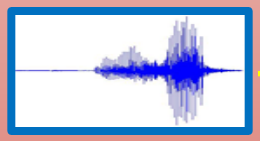
lecture recordings



Summary



Speech Summarization



Speech Recognition

你好

core techniques

Computer Assisted Language Learning

I would like to leave Taipei on November 2nd

Information Extraction

Hi

Hello

Dialogue