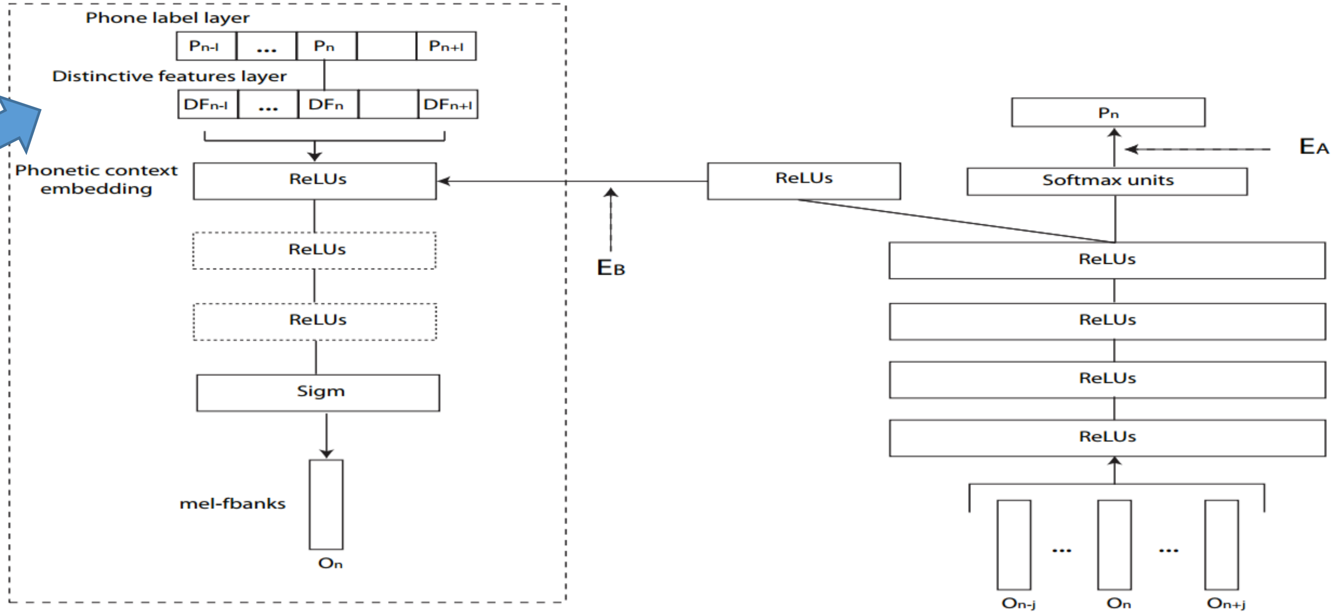
A Brief Review for Selected Papers in Interspeech16

Zhiyuan Tang, Speech Group

Feature extraction/Representation learning

consonant, voiced, unvoiced, fricative, nasal, stop
approximant, affricate, labial, dental, alveolar, lateral
post-alveolar, palatal, velar, glottal, syllabic, flapping
vowel, diphtong, nasalized, r-merged, close, close-mid
mid, open-mid, open, front, central, back, long, short
close2, close-mid2, mid2, open-mid2, open2
front2, central2, back2, long2, short2, silence



Paper: Phonetic Context Embeddings for DNN-HMM Phone Recognition

Feature extraction/Representation learning

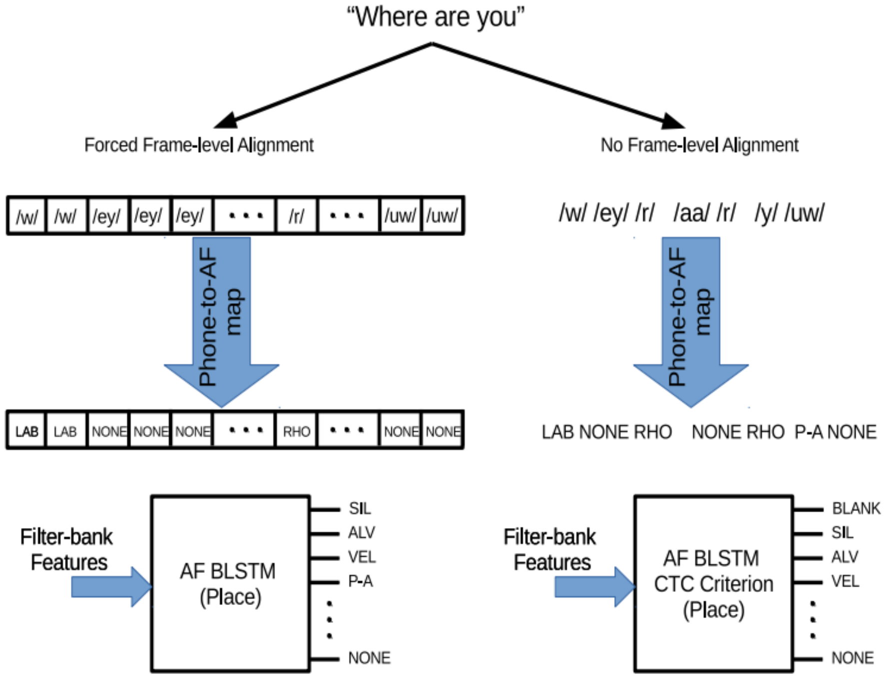


Figure 1: Building "Place" Articulatory Classifier

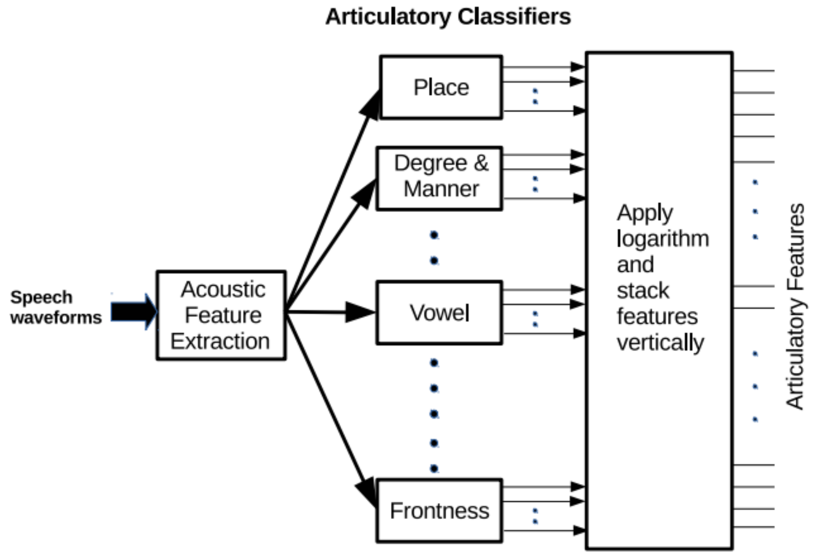
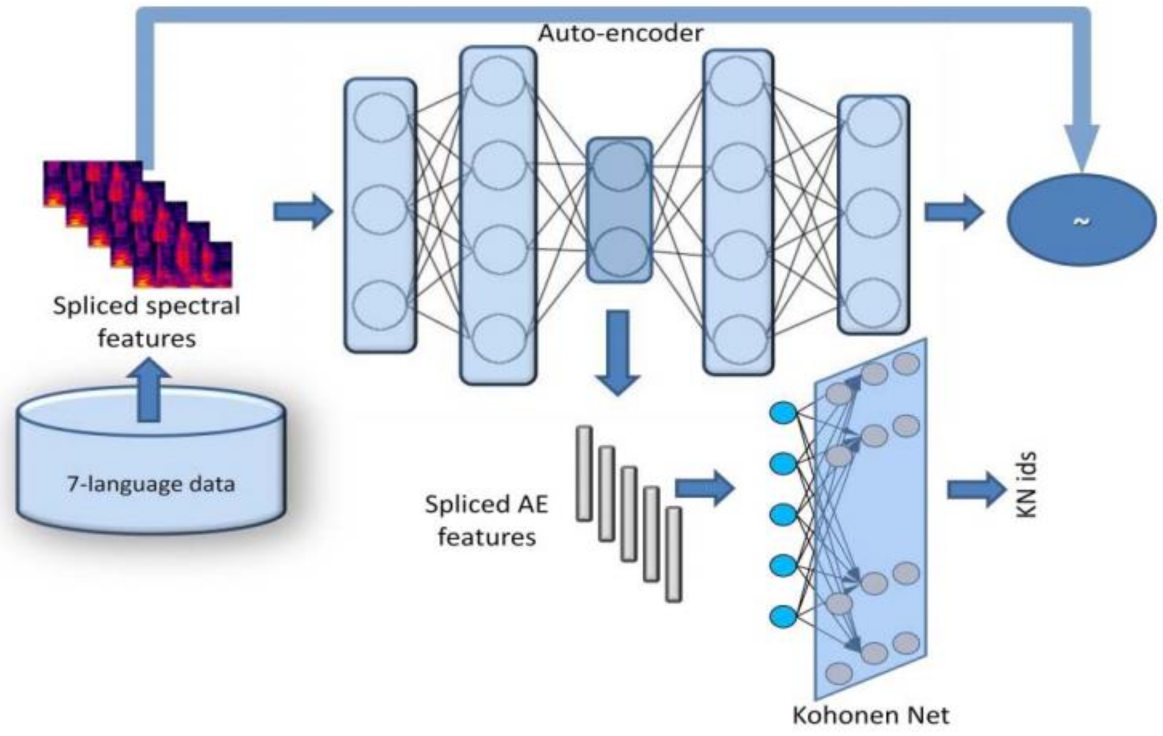


Figure 2: Articulatory Feature Extractor

Paper: Articulatory Feature Extraction Using CTC to Build Articulatory Classifiers Without Forced Frame Alignments for Speech Recognition

Feature extraction/Representation learning



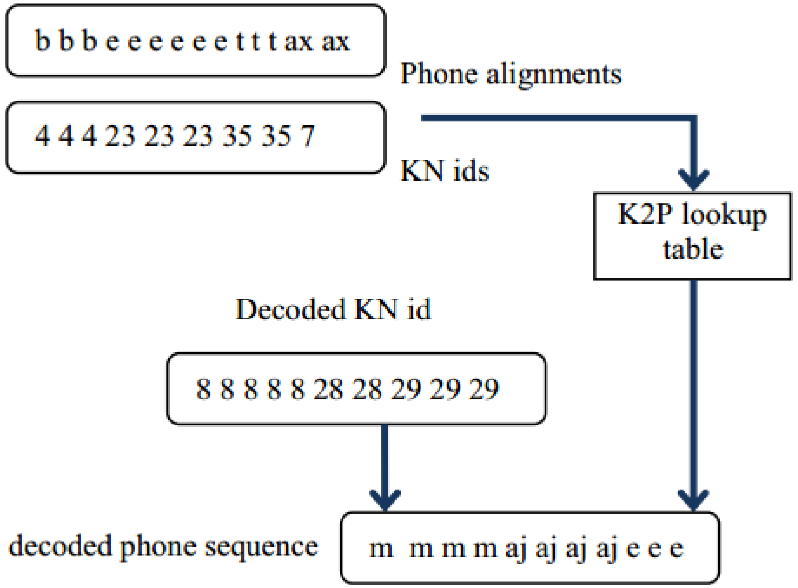


Figure 2: *Schematics of K2P lookup table creation and KN-id decoding.*

Paper: Unsupervised Learning of Acoustic Units Using Autoencoders and Kohonen Nets

Feature extraction/Representation learning

Mel features: $x \xrightarrow{\mathscr{F}} X \xrightarrow{|.|^2} |X|^2 \xrightarrow{Mel} Y = M|X|^2$

CLP feature: $x \xrightarrow{\mathscr{F}} X \xrightarrow{W} Y = WX \xrightarrow{|.|} |Y|$

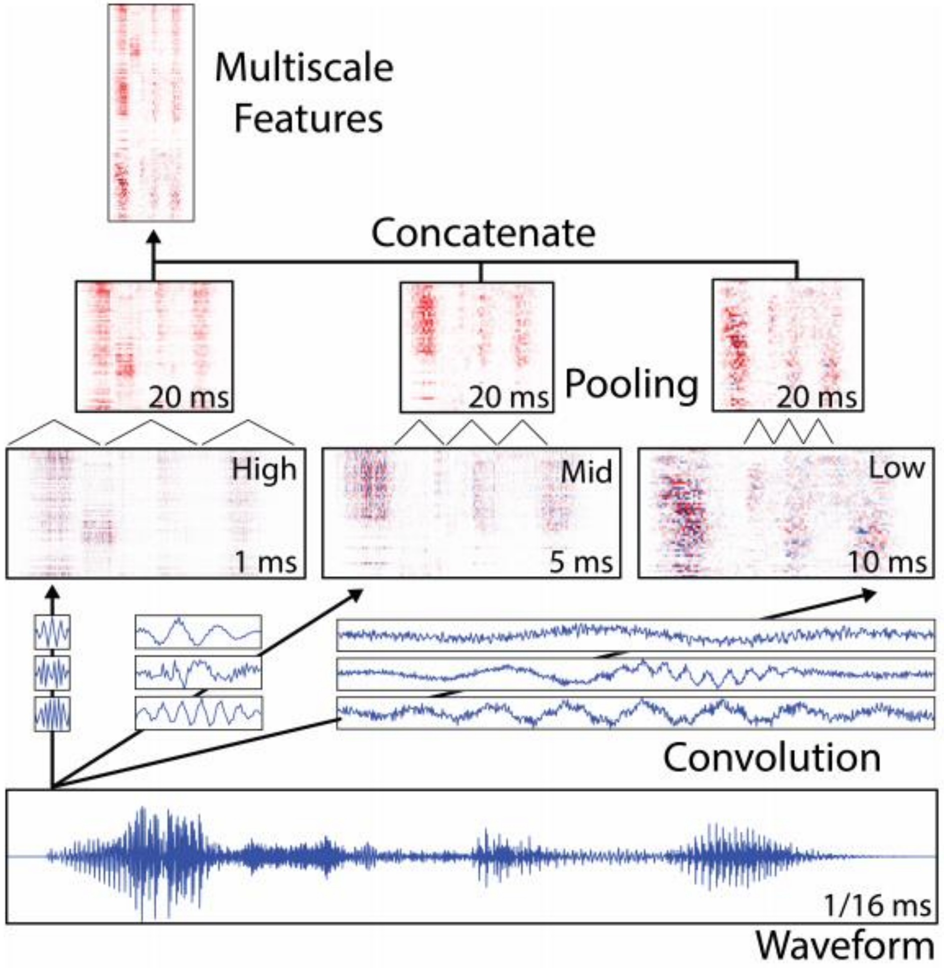
$$X = X_R + jX_I$$

$$W = W_R + jW_I$$

$$\begin{aligned} |Y| &= \left[\Re\{Y\}^2 + \Im\{Y\}^2\right]^{1/2} \\ \Re\{Y\} &= W_R X_R - W_I X_I \\ \Im\{Y\} &= W_R X_I + W_I X_R \end{aligned}$$

Paper: Complex Linear Projection (CLP): A Discriminative Approach to Joint Feature Extraction and Acoustic Modeling

Feature extraction/Representation learning



Paper: Learning Multiscale Features Directly from Waveforms

Feature extraction/Representation learning

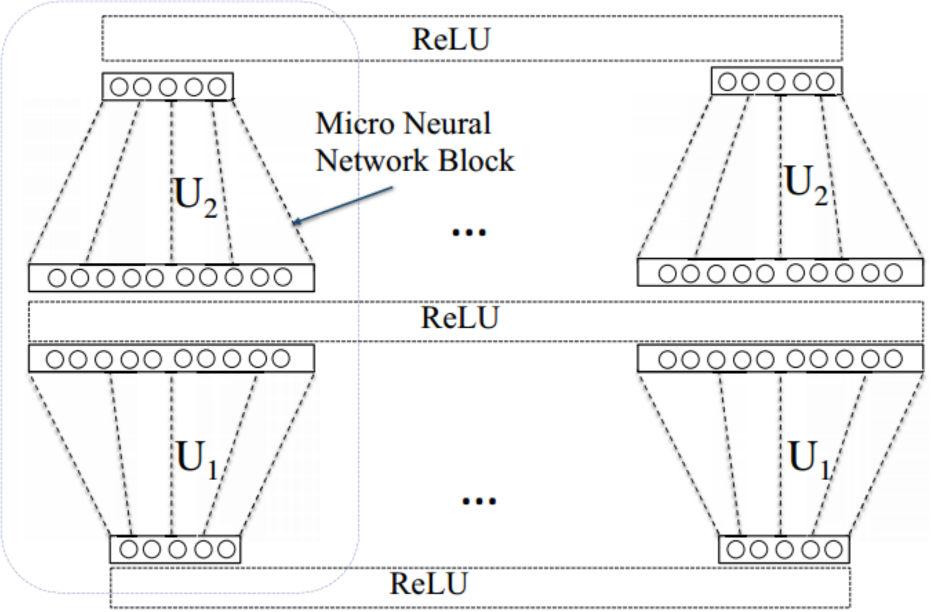


Figure 1: *Proposed NIN nonlinearity*

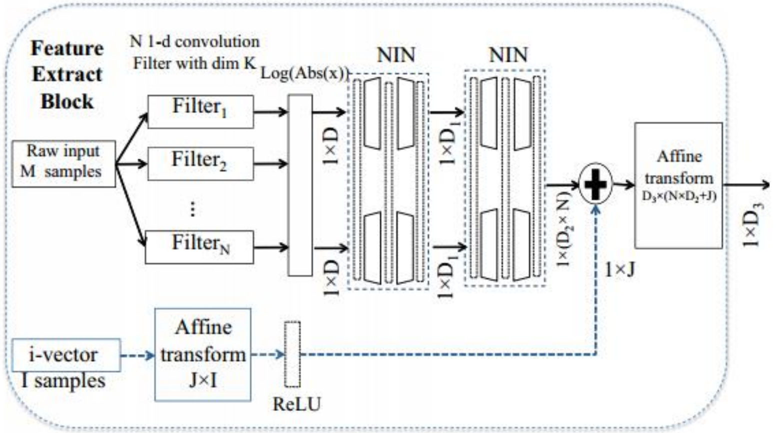


Figure 3: *Raw waveform feature extraction Block.*

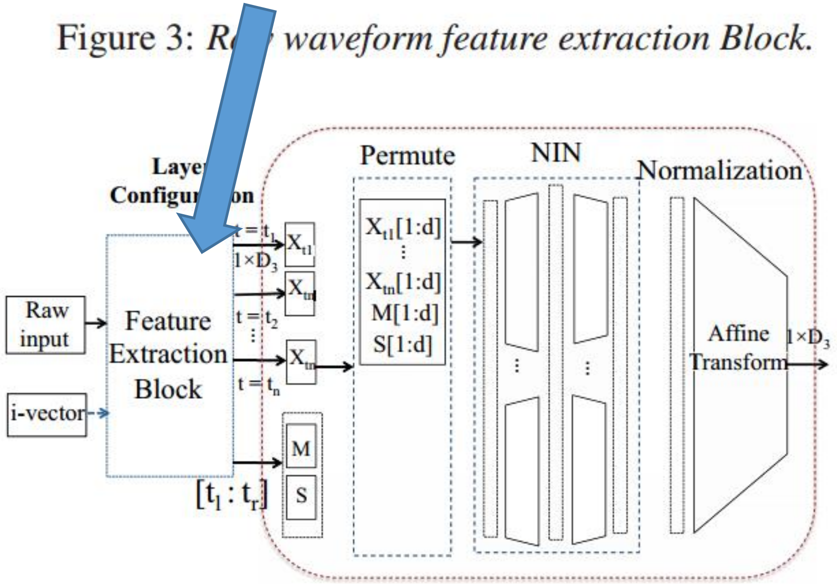
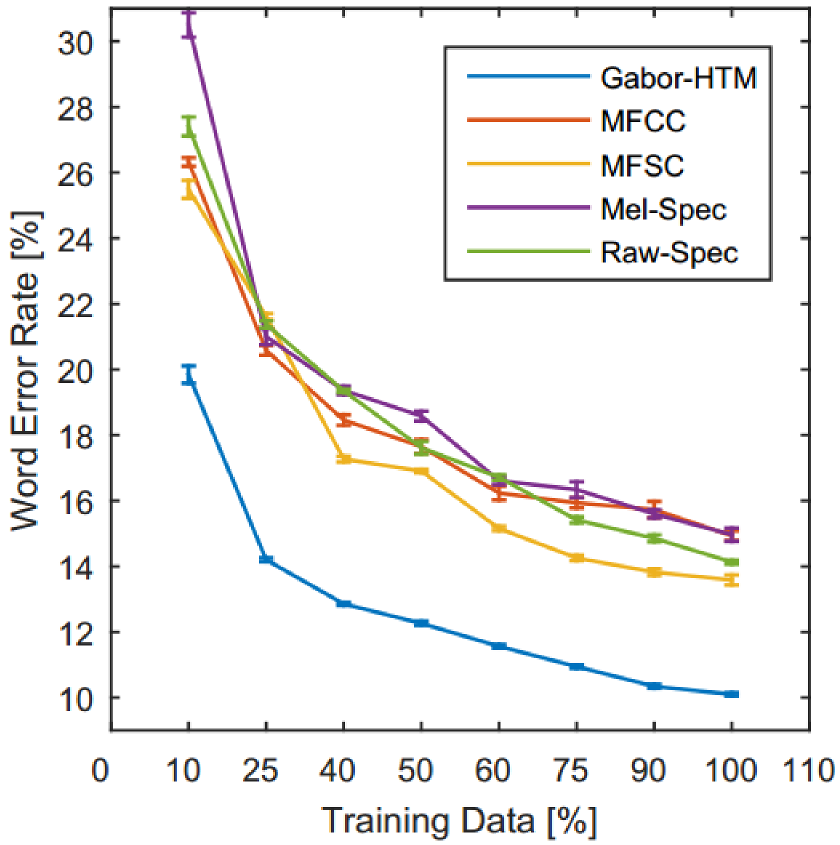


Figure 4: *Layer configuration in raw waveform classification block.*

Paper: Acoustic modelling from the signal domain using CNNs

Feature extraction/Representation learning

1. the amplitude spectrogram (Raw-Spec) obtained by applying FFT,
2. Mel-spectrogram (Mel-spec),
3. log-Mel-spectrogram(MFSC),
4. MFCC (plus deltas and double deltas) / Gabor filterbank (GBFB).



Paper: Why do ASR Systems Despite Neural Nets Still Depend on Robust Features

Feature extraction/Representation learning

Convolutional layers:
1. frequency
2. time-frequency LSTMs,
3. grid LSTMs (LDNN**)**
4. ReNet LSTMs.

Paper: Modeling Time-Frequency Patterns with LSTM vs. Convolutional Architectures for LVCSR Tasks

Network architecture

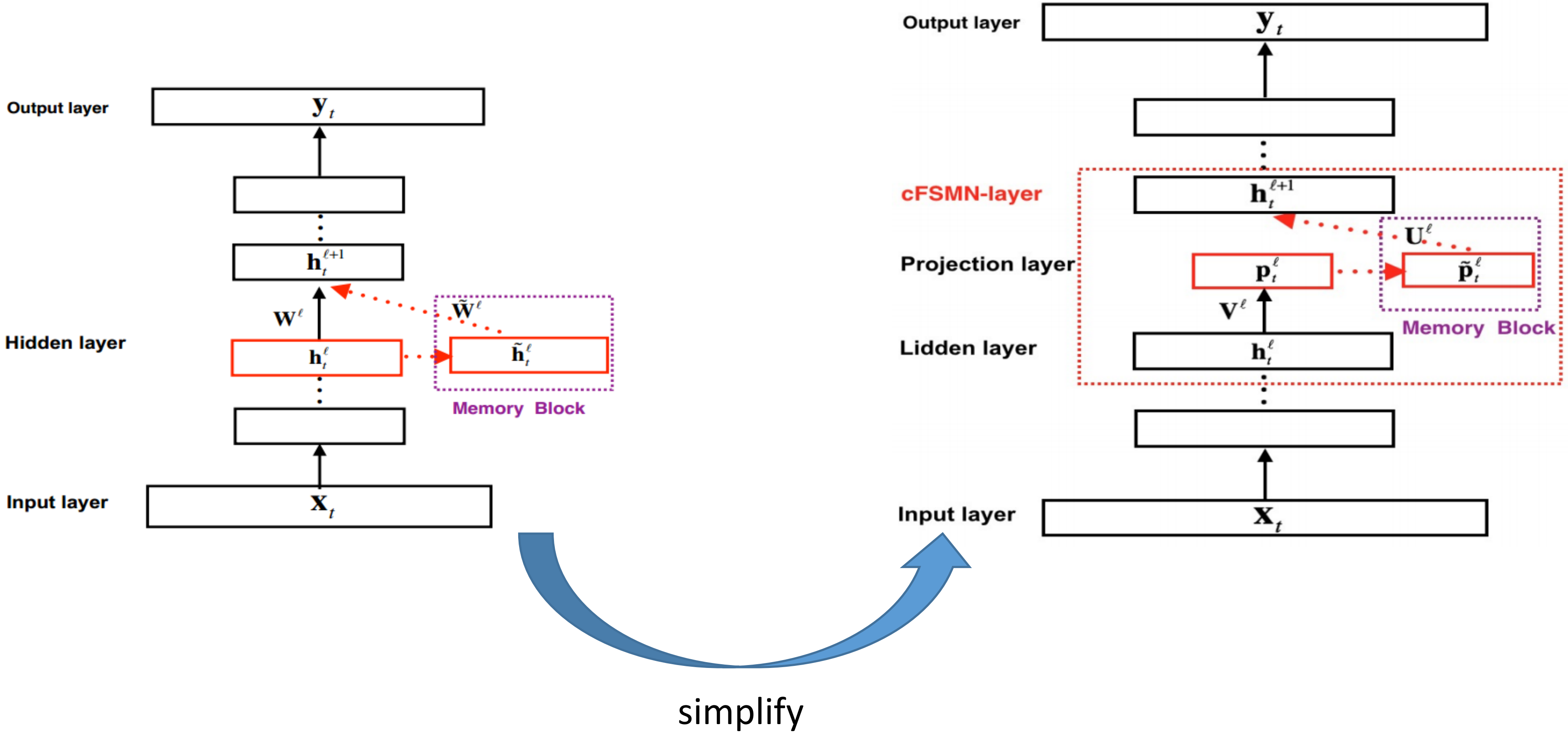
$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$
$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$
$$\overline{h}_t = \sigma(W_h x_t + U_h (r_t \odot h_{t-1}) + b_h)$$
$$h_t = (1 - z_t) h_{t-1} + z_t \overline{h}_t$$

$$z_t = \sigma(W_z * x_t + \mathrm{pool}(U_z * h_{t-1}) + b_z)$$
$$r_t = \sigma(W_r * x_t + \mathrm{pool}(U_r * h_{t-1}) + b_r)$$
$$\overline{h}_t = \sigma(W_h * x_t + \mathrm{pool}(U_h * (r_t \odot h_{t-1})) + b_h)$$
$$h_t = (1 - z_t) h_{t-1} + z_t \overline{h}_t$$

GRU + convolution + pooling

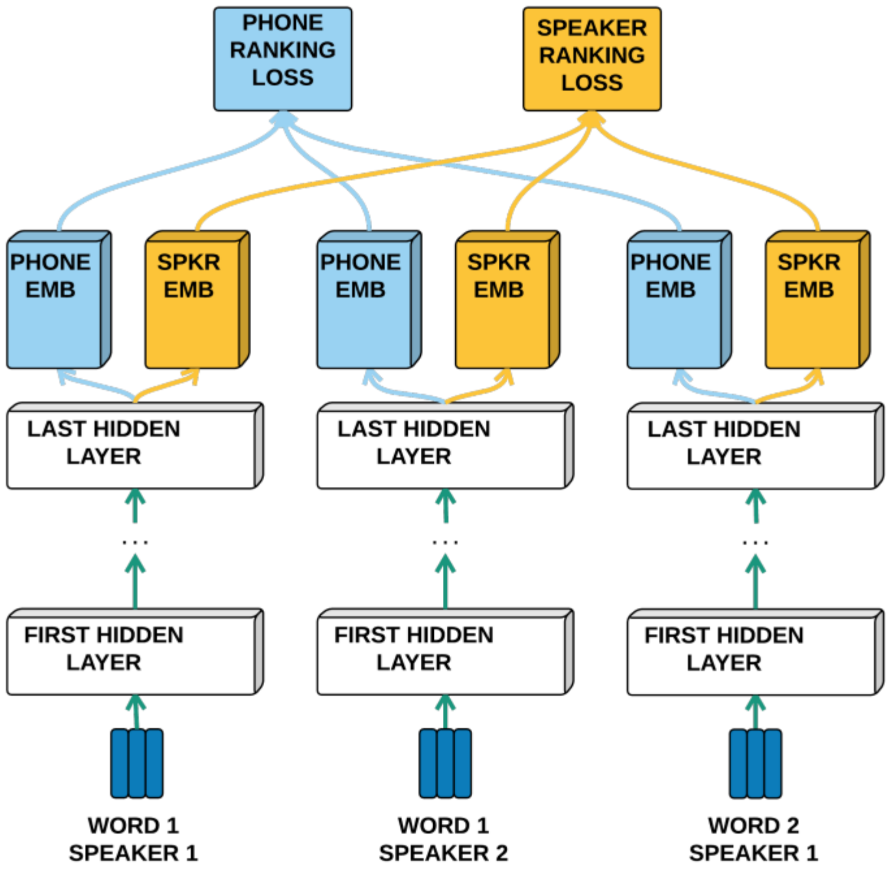
Paper: Acoustic Modeling Using Bidirectional Gated Recurrent Convolutional Units

Network architecture



Paper: Compact Feedforward Sequential Memory Networks for Large Vocabulary Continuous Speech Recognition

Network architecture



asymmetry:
speaker benefited consistently,
not for phone task.

Paper: Joint Learning of Speaker and Phonetic Similarities with Siamese Networks

End-to-End

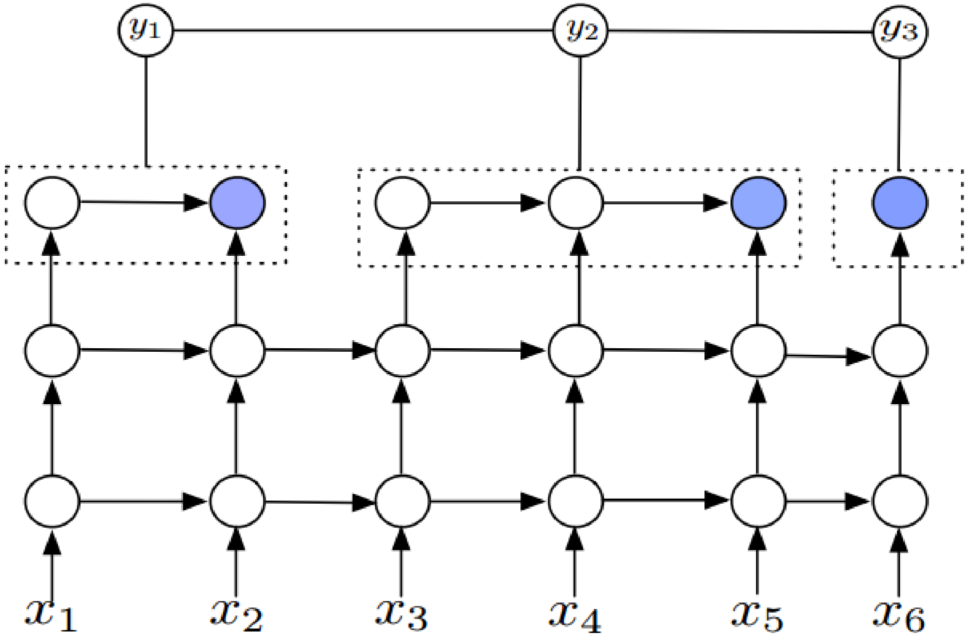
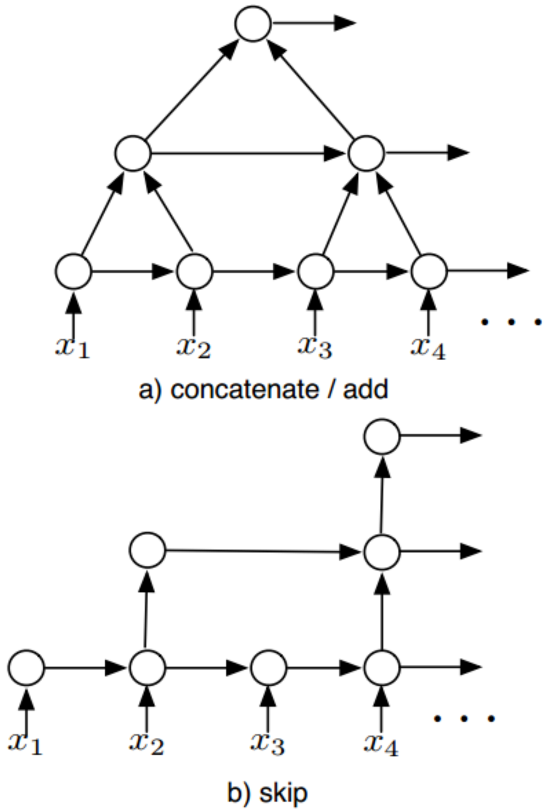


Figure 1: Segmental RNN using a first-order CRF. The coloured circles denote the segment embedding vector $\mathbf{h}^j_{d_j}$ in Eq.(7). Using bi-directional RNNs is straightforward.



segmental RNN = RNN encoder + segmental CRF

Paper: Segmental Recurrent Neural Networks for End-to-end Speech Recognition

End-to-End

CNNs with CTC without recurrent:

(1) more layers, which results in more nonlinearities and larger input receptive fields for units in the top layers;

(2) reasonably **large context** windows, which help the model to capture the spatial/temporal relations of input sequences in reasonable time-scales;

(3) the **Maxout** unit, which has more functional freedoms comparing to ReLU and parametric ReLU.

Paper: Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks

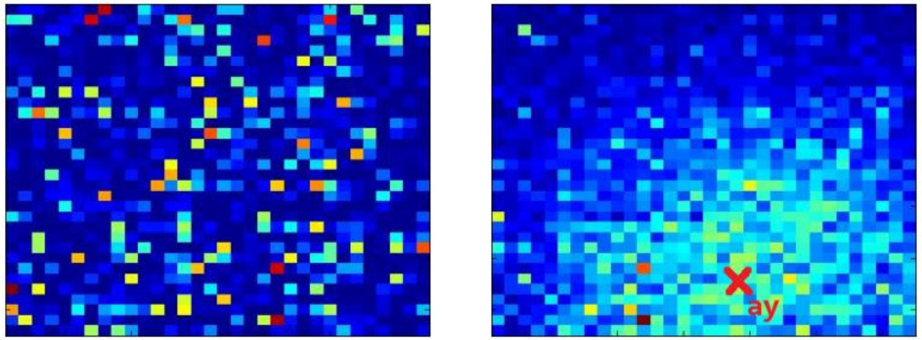
Regularization/Adaptation

$$\mathcal{R}_{st}(\boldsymbol{\theta}) = \frac{1}{T}\sum_t \sum_l \sum_i g(\mathbf{s}_i, \hat{\mathbf{s}}_{p_t}) \log \left(\frac{g(\mathbf{s}_i, \hat{\mathbf{s}}_{p_t})}{\bar{h}_i^{(l)}(\mathbf{x}_t)} \right)$$

$$\mathcal{R}_L(\boldsymbol{\alpha}^{(s)}) = \frac{1}{2T^{(s)}} \sum_l \sum_i \sum_j \left(q_{ij} \sum_{t \in \mathbb{I}^{(s)}} f_{ij}\left(\mathbf{x}_t; \boldsymbol{\alpha}^{(s)}\right) \right)$$

stimulation term: encourage the DNN activations in a region (prior) to be similar:

1. regularization, similar activations are grouped together in the network-grid, a phone (or grapheme) dependent prior distribution is defined over the normalized activation function outputs;
2. speaker adaptation, grouping based on functional similarities on speaker-dependent scaling factors.

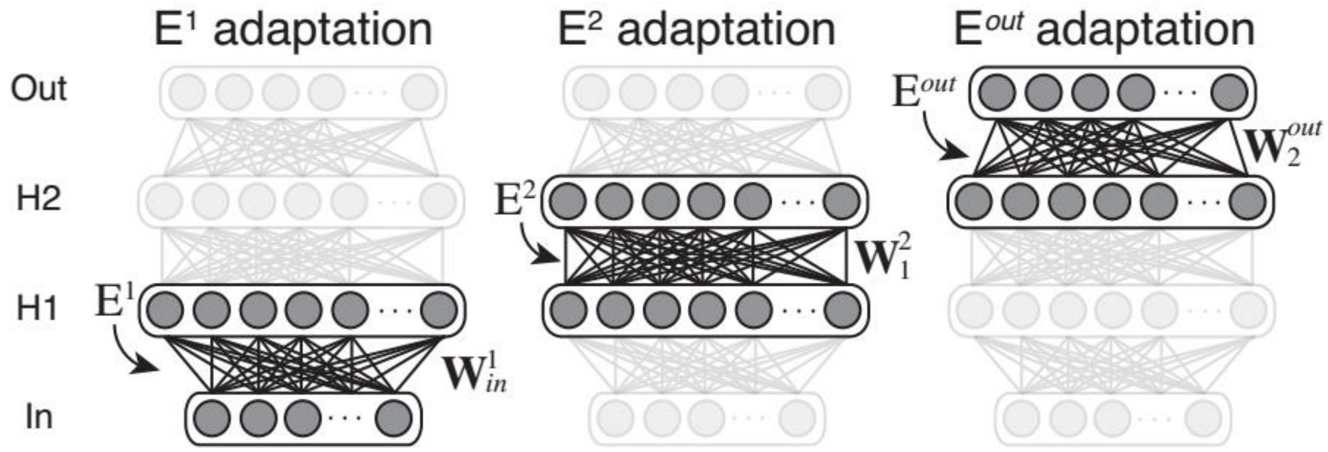


(a) Unstimulated (b) Stimulated $\eta_{st} = 0.05$

position nearer, pattern more similar

Paper: Stimulated Deep Neural Network for Speech Recognition

Regularization/Adaptation

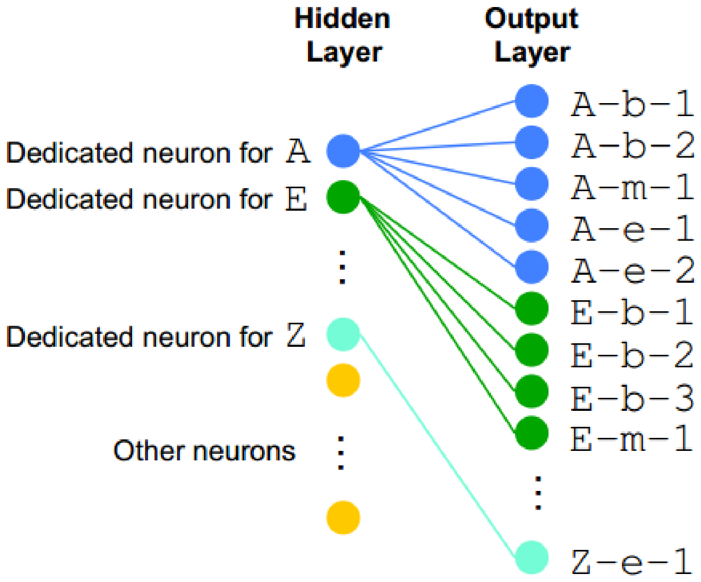


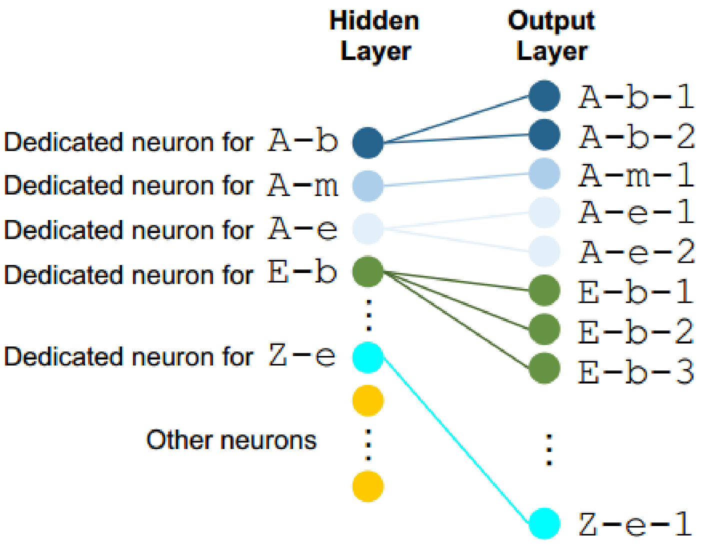
$$E^\ell = \sum_{i,j=1}^{N} \left(\frac{1}{T} \sum_{\tau=1}^{T} y_{i\tau}^\ell y_{j\tau}^\ell - c_{Rij}^\ell \right)^2 .$$

co-activation statistics of each layer describes the relationship of activations, during adaptation, minimize the distance between new co-activation and the previous one (**unsupervised**), this can filter noise

Paper: Adaptation of Neural Networks Constrained by Prior Statistics of Node Co-Activations

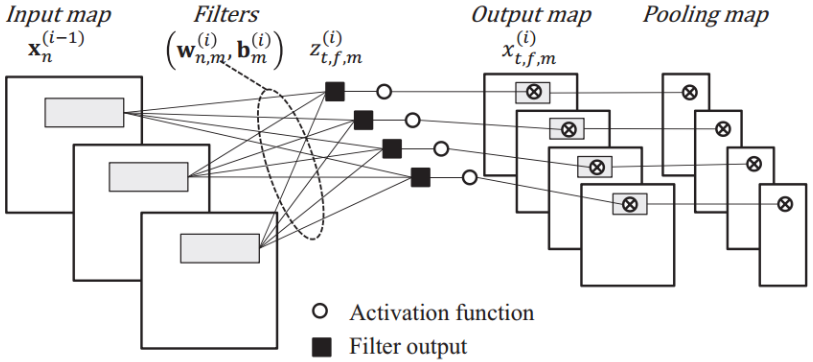
Regularization/Adaptation



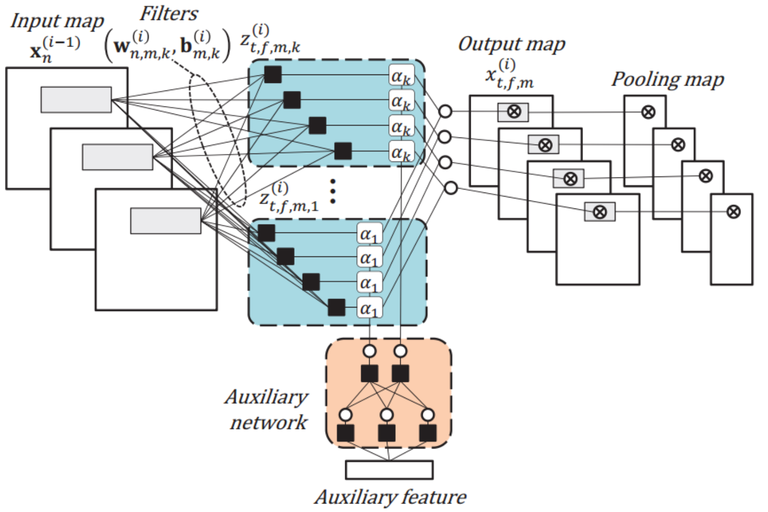


Paper: Improved Neural Network Initialization by Grouping Context-Dependent Targets for Acoustic Modeling

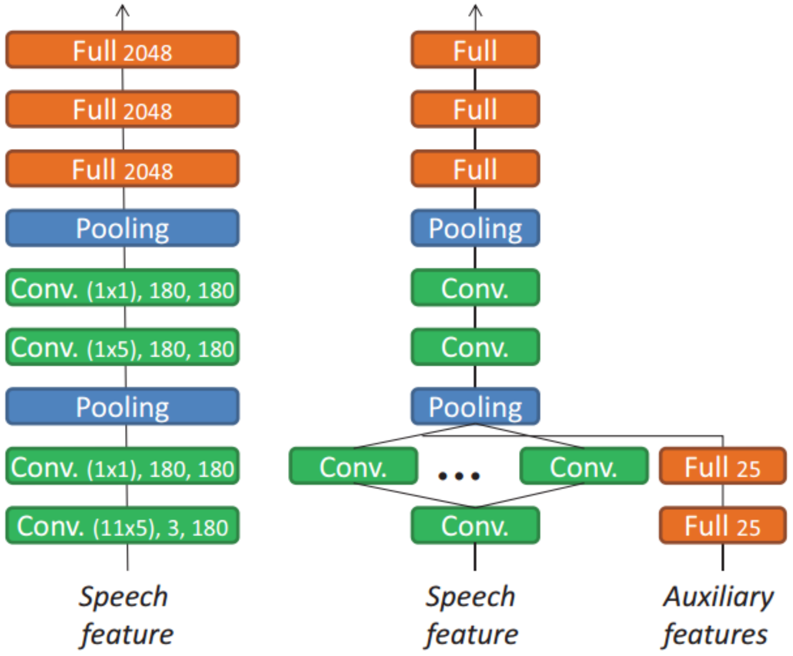
Regularization/Adaptation



(a) convolutional layer



(b) context adaptive convolutional layer



(a) NiN (b) CA-CNN (i=2)

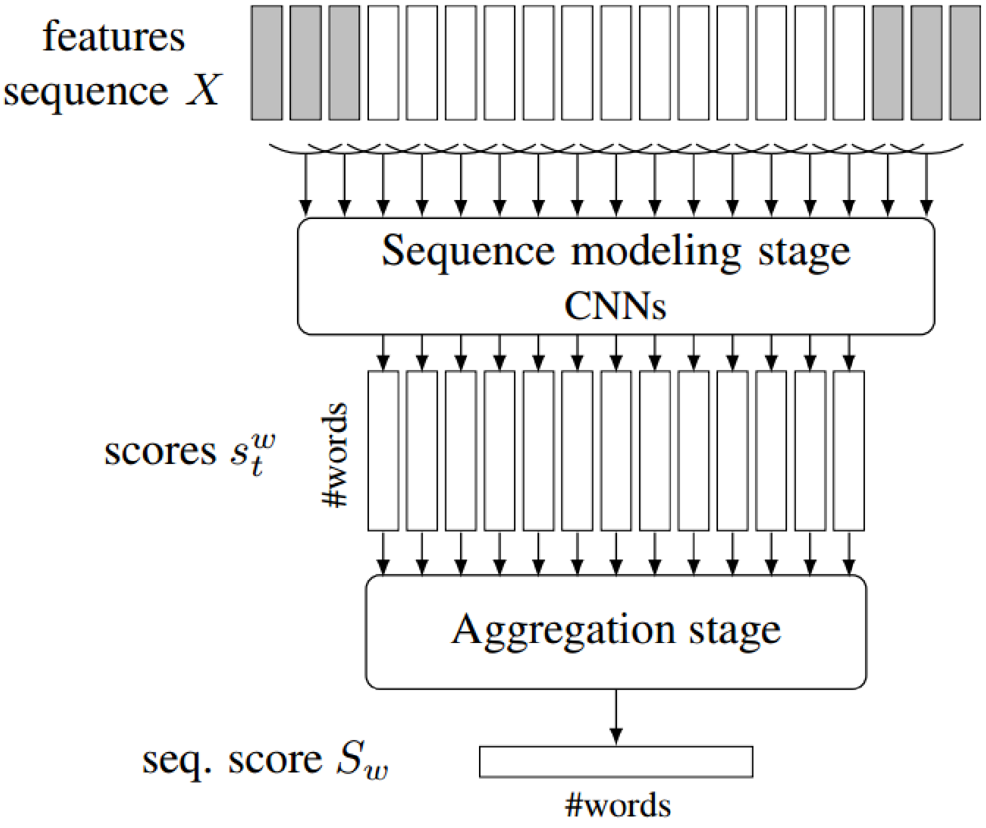
Paper: Context Adaptive Neural Network for Rapid Adaptation of Deep CNN Based Acoustic Models

Regularization/Adaptation

1. the learning hidden unit contributions (LHUC), each speaker puts a vector on activations of each layer,
2. subspace LHUC means, the vector can be computed from a low-dim vector (such as i-vector) multiplied by a matrix shared among speakers for one layer.

Paper: Subspace LHUC for Fast Adaptation of Deep Neural Network Acoustic Models

Attention



$$S_w^r(X) = \frac{1}{r}\log\left(\frac{1}{T}\sum_t \exp(r s_t^w(X))\right)$$

$$\mathcal{L}(S(X), y) = \sum_{w=1}^{|\mathcal{W}|} \log(1 + e^{-y_w S_w(X)})$$

Bag-of-word

Paper: Jointly Learning to Locate and Classify Words Using Convolutional Networks

Attention

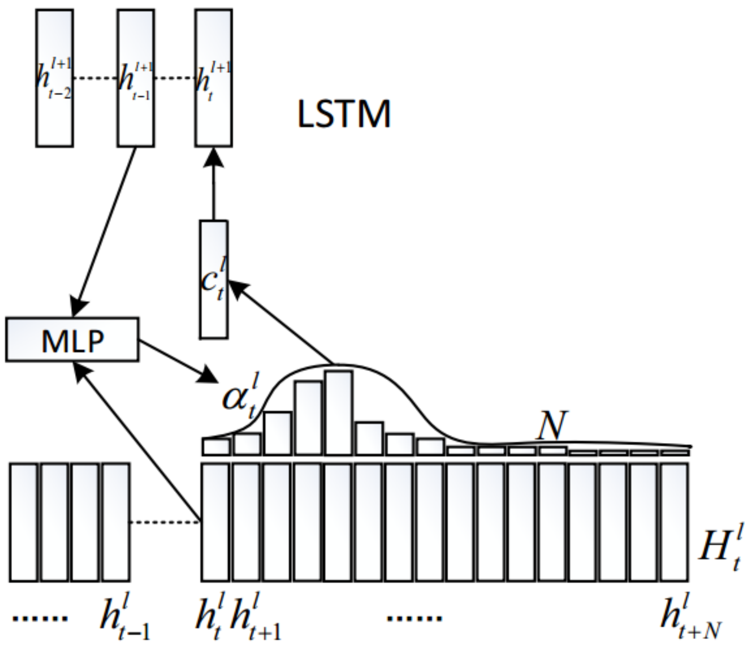


Figure 2: Attention-based LSTM architecture with future context size of N.

Paper: Future Context Attention for Unidirectional LSTM Based Acoustic Model

Attention

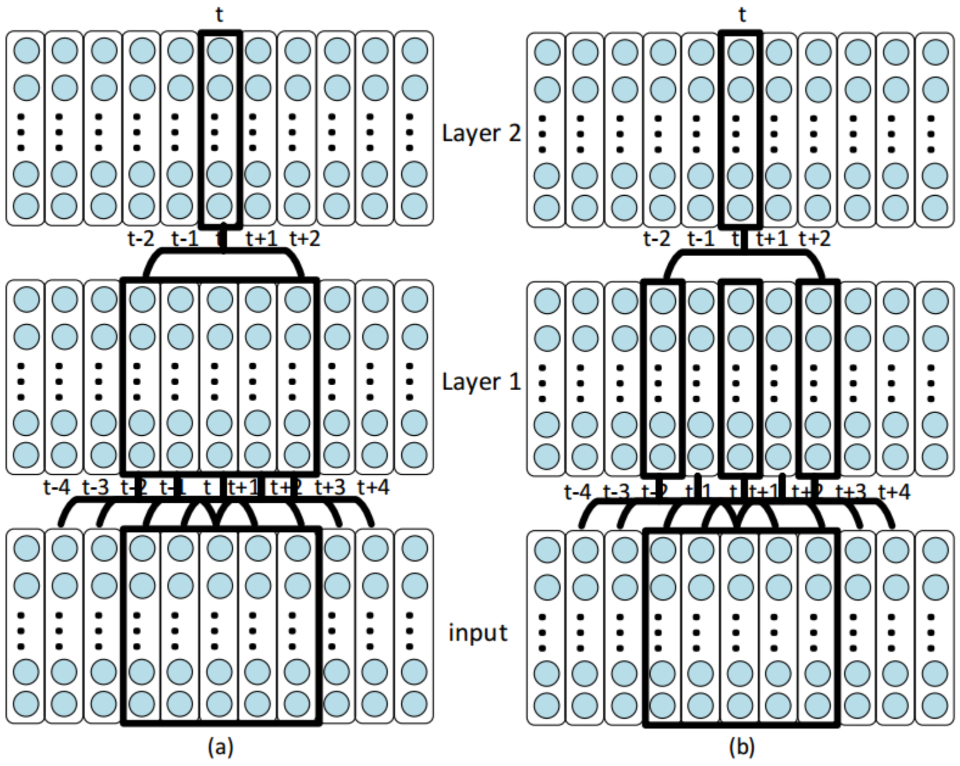


Figure 1: *Illustration of the layer-wise context expansion.*

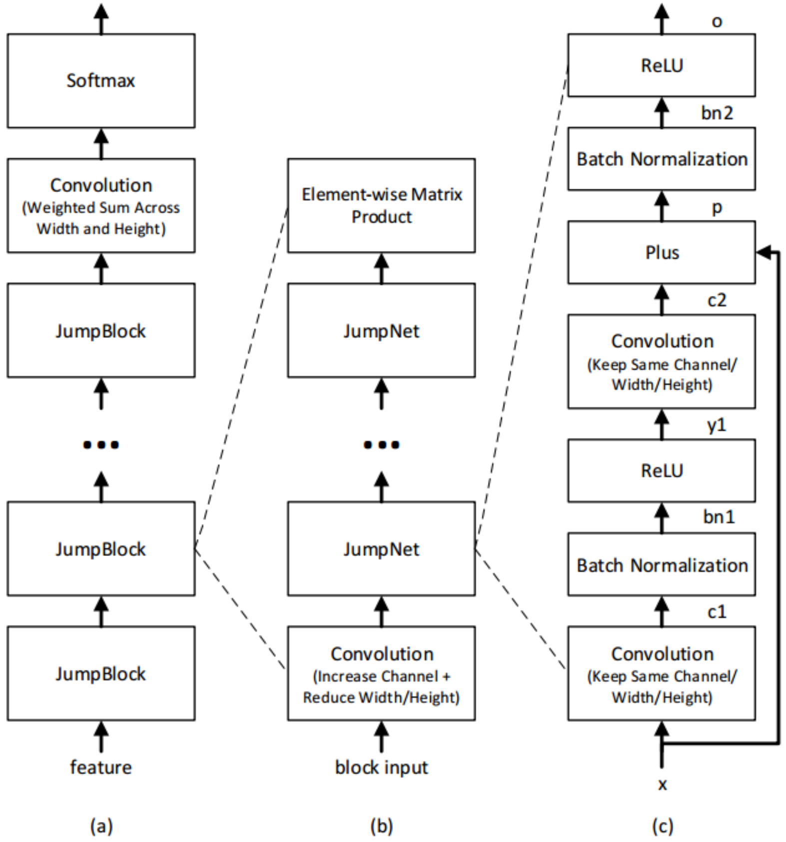


Figure 2: *The detailed diagram of the final model.*

1. with a same kernel;
2. the contribution of each lower frame is learned (**attention**), same for all channels;

Paper: Deep Convolutional Neural Networks with Layer-wise Context Expansion and Attention

Criterion

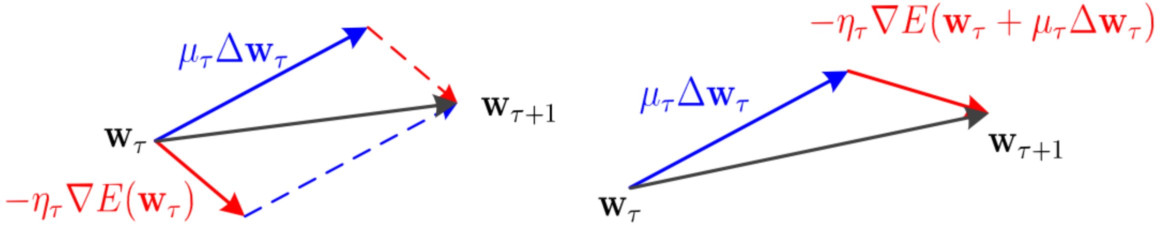


Figure 1: *Geometrical interpretation for SGD with momentum (left) and NAG (right).*

1. Hard: iteratively
2. Soft: $$\mathbf{w}_{\tau+1} = \mathbf{w}_\tau + \mu_\tau \left[\gamma + (1-\gamma)\mu_{\tau+1}\right] \Delta \mathbf{w}_\tau - \eta_\tau \left[\gamma + (1-\gamma)(1+\mu_{\tau+1})\right] \nabla E(\mathbf{w}_\tau)$$

Paper: Hybrid Accelerated Optimization for Speech Recognition

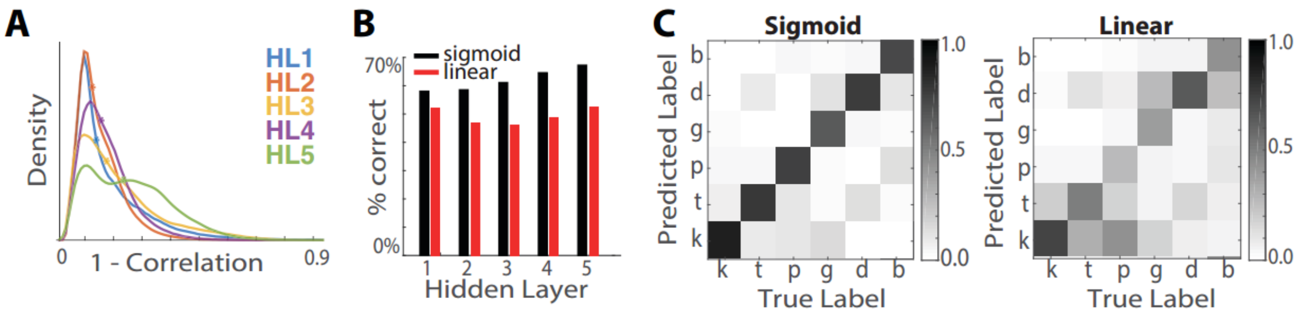
Criterion

minimize a weighted average of the **MMI** criterion and the **KL-divergence** between the student and teacher hypothesis posteriors, hypothesis level

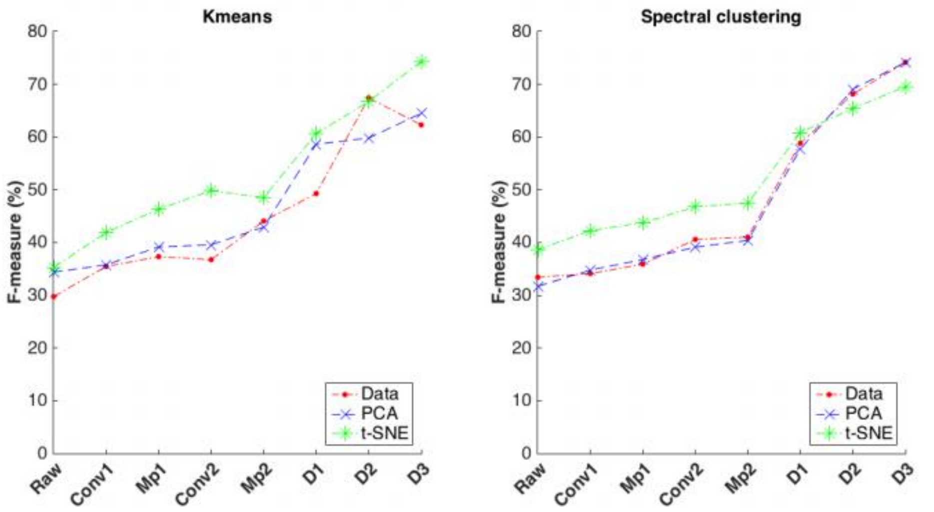
Paper: Sequence Student-Teacher Training of Deep Neural Networks

Visualization

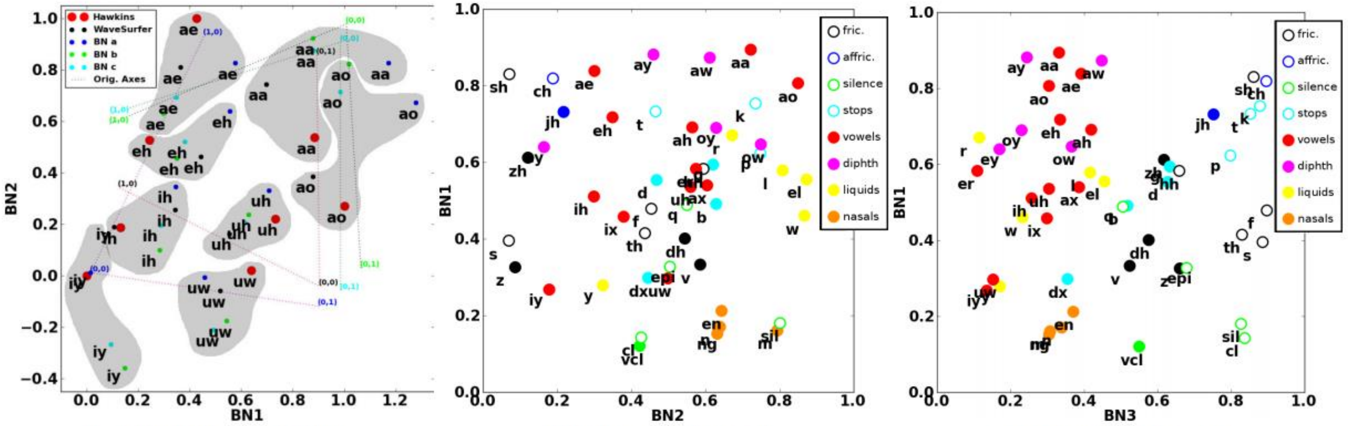
1.



2. linear and non-linear clustering; activation map



3. Phone to small-size BNF (3 to 9 neurons)



(a) Aligned BNFs and formants. (b) BNF (1, 2) MLP 3d/3. (c) BNF (1, 3) MLP 3d/3.

Paper: On the Role of Nonlinear Transformations in Deep Neural Network Acoustic Models

Inferring phonemic classes from CNN activation maps using clustering techniques

Interpretation of Low Dimensional Neural Network Bottleneck Features in Terms of Human Perception and Production

Compression

1. GMM, SGMM, DNN, memory bandwidth
2. quantization, 32-bit to 8-bit

Paper: Memory-Efficient Modeling and Search Techniques for Hardware ASR Decoders
On the efficient representation and execution of deep acoustic models

ASR system

- i-vector based Bottleneck features
- data augmentation (scale the waveform, change the speed)
- score fusion of different models
- rescore with LMs of 4-gram or/and LSTM

1. Improving English Conversational Telephone Speech Recognition
2. The IBM 2016 English Conversational Telephone Speech Recognition System

Highway

DNN:
(LSTM)

$$\mathbf{y} = H(\mathbf{x}, \mathbf{W_H}).$$

$$\mathbf{y} = H(\mathbf{x}, \mathbf{W_H}) \cdot \underline{T(\mathbf{x}, \mathbf{W_T})} + \mathbf{x} \cdot \underline{C(\mathbf{x}, \mathbf{W_C})}.$$

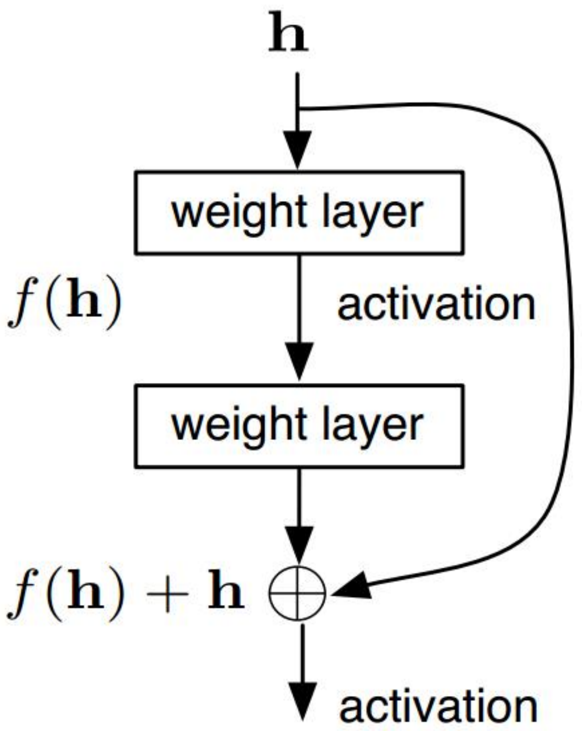
LSTM:

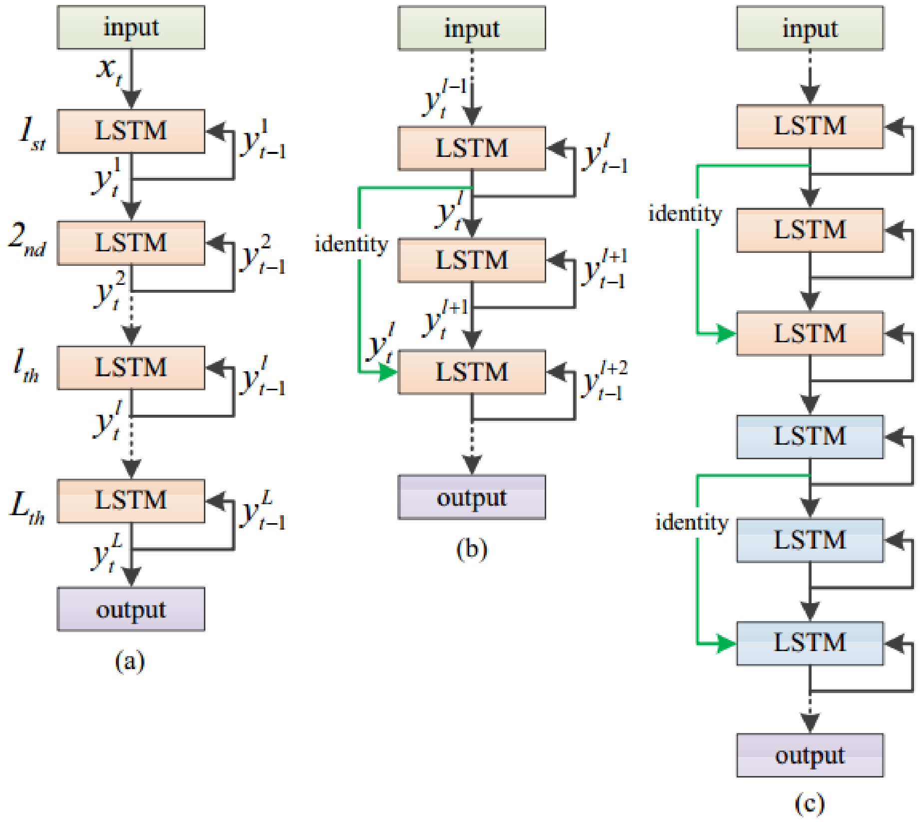
$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{mc}\mathbf{m}_{t-1} + \mathbf{b}_c)$$

$$\left[\begin{aligned} \mathbf{c}_t^{l+1} &= \underline{\mathbf{d}_t^{(l+1)} \odot \mathbf{c}_t^l} + \mathbf{f}_t^{(l+1)} \odot \mathbf{c}_{t-1}^{(l+1)} \\ &+ \mathbf{i}_t^{(l+1)} \odot \tanh(\mathbf{W}_{xc}^{(l+1)}\mathbf{x}_t^{(l+1)} + \mathbf{W}_{hc}^{(l+1)}\mathbf{m}_{t-1}^{(l+1)} + \mathbf{b}_c), \\ \mathbf{d}_t^{(l+1)} &= \sigma(\mathbf{b}_d^{(l+1)} + \mathbf{W}_{xd}^{l+1}\mathbf{x}_t^{(l+1)} + \mathbf{w}_{cd}^{l+1} \odot \mathbf{c}_{t-1}^{(l+1)} + \underline{\mathbf{w}_{ld}^{(l+1)} \odot \mathbf{c}_t^l}), \end{aligned}\right.$$

Paper: Training Very Deep Networks/Recurrent Highway Networks
Highway long short-term memory RNNS for distant speech recognition

Residual

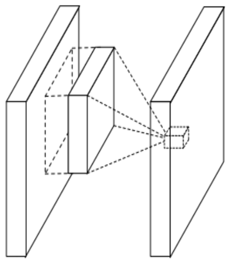




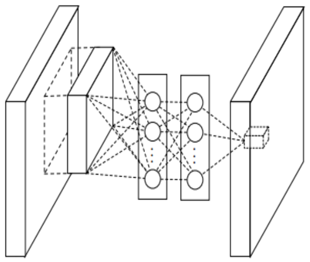
1. along the **spatial and temporal** dimension
2. a row **convolution** layer on the top

Paper: Multidimensional Residual Learning Based on Recurrent Neural Networks for Acoustic Modeling

Network in Network



(a) Linear convolution layer



(b) Mlpconv layer

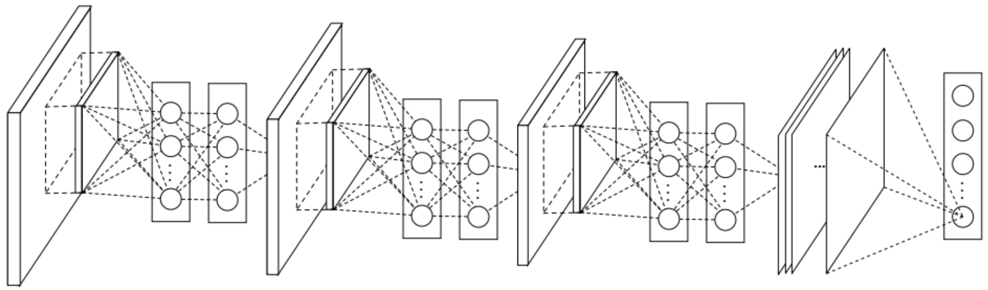
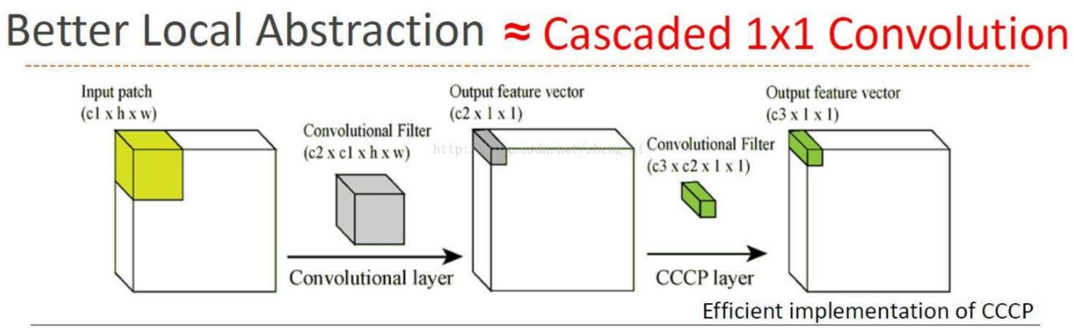


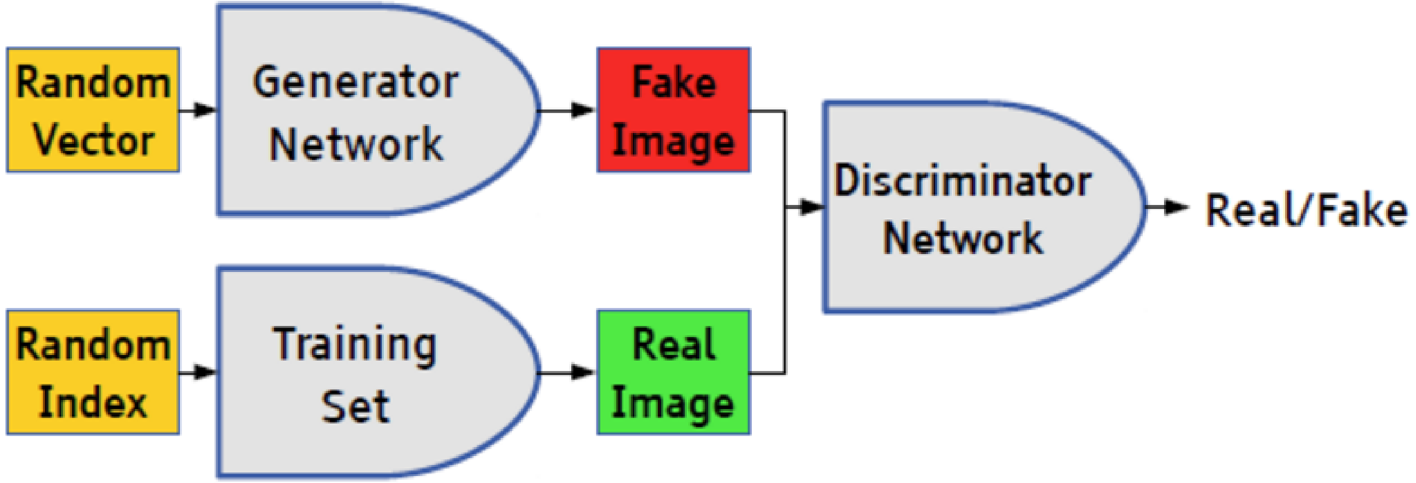
Figure 2: The overall structure of Network In Network. In this paper the NINs include the stacking of three mlpconv layers and one global average pooling layer.



cascaded cross channel parametric pooling

Paper: Network In Network

Adversarial Network



Paper: Generative Adversarial Nets

Others

Paper: Lower Frame Rate Neural Network Acoustic Models

1. lower frame rate: typical 10ms log-mel frontend, subsample frames (keep every n-th one);
2. convolution + LSTM + dnn;
3. reduced latency and graceful degradation (data dependency) on smaller datasets, gains with convolution, compared to CTC-trained model.

Paper: How Neural Network Depth Compensates for HMM Conditional Independence Assumptions in DNN-HMM Acoustic Models

1. synthetic data, a bootstrap resampling framework that allows us to control the amount of data dependence;
2. only when data become more dependent that depth improves ASR performance

Paper: Purely sequence-trained neural networks for ASR based on lattice-free MMI

chain in Kaldi: MMI from scratch like CTC, 3-fold reduced frame, 4-gram phone level language model

Paper: Advances in Very Deep Convolutional Neural Networks for LVCSR