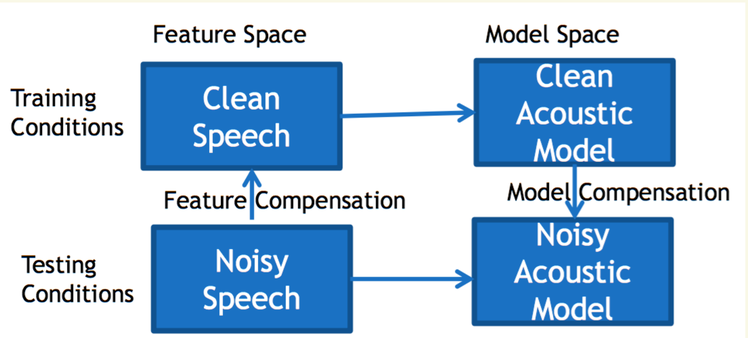
An Investigation of Deep Neural Networks for Noise Robust Speech Recognition

Topic description

- Speech recognition in noisy environment.
- Different noise types: subway, bubble, car, exhibition, restaurant, etc.
- Different SNRs.
- Applications of DNN and effect of dropout.

General solutions overview

- Feature Compensation.
- Model Compensation.



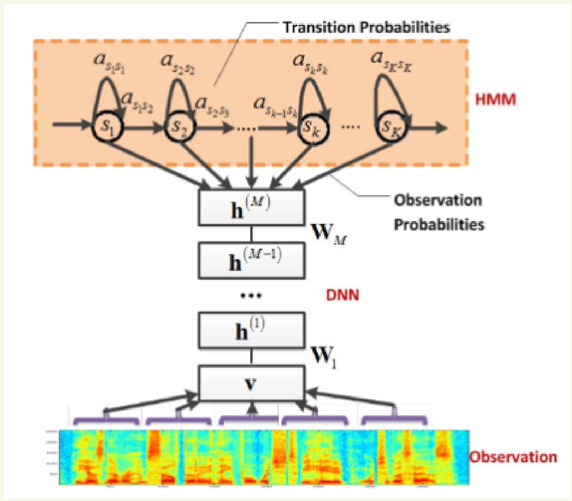
Data

- Tidigit + noise → Aurora 4 (noisy speech).

Set	Clean	Noise types	SNR level
training	✓	subway, bubble, car, exhibition	5, 10,15
test A	⊠	subway, bubble, car, exhibition	-5, 0, 20
test B	✓	restaurant, street, airport	-5 ∼ 20
test C	✓	subway, street	-5 ∼ 20
test D	✓	subway, bubble, car	-5 ∼ 20

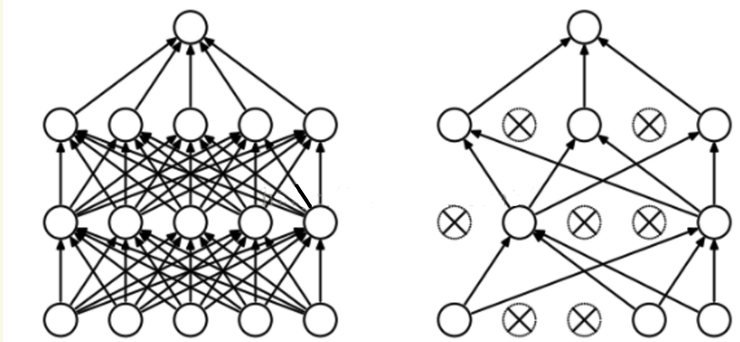
Deep Neural Network acoustic model

- DNN-HMM hybrid with input containing extended context window.



Regularization using dropout

- DNN training with dropout is used to achieve better regularization.
- Avoids overfitting caused by learning the noise.

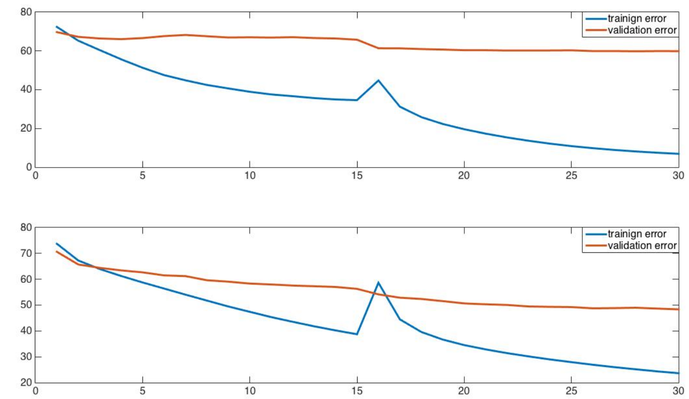


(a) Standard Neural Net (b) After applying dropout.

Experiments

- DNN architecture:
  - 792 input units for FBANK features with first and second order derivatives and 5 frames of context
  - 64 output units for phoneme-states
  - 3 hidden layers with 2048 units each
  - ReLU activation function
- DNN training:
  - 15 epochs with learning rate 0.1
  - 15 epochs of fine tuning with learning rate 0.004
  - SGD with momentum equal 0.9
  - mini-batch size of 512
- Tested variable: 20% of dropout.

Results



Error rate for test sets on frame and phoneme level:

DNN	A	B	C	D	AVG
baseline	63.15	69.39	62.63	62.02	64.30
	54.40	58.92	55.03	54.16	55.63
dropout	50.00	54.03	51.66	52.89	52.15
	42.31	45.93	44.05	45.52	44.45

Mateusz Buda, Masoumeh P. Ghaemmaghami, Lin Qi
buda@kth.se, mpg@kth.se, linqi@kth.se
KTH Royal Institute of Technology, Department of Computer Science