Articulatory and Spectrum Features Integration Using Generalized Distillation Framework

K. Markov, University of Aizu
T. Matsui, ISM

Outline

- Generalized Distillation
  - Hinton’s Distillation
  - Vapnik’s Privileged Information
- Application in ASR
  - Articulatory and Spectrum Information Integration
    - Feature based Integration
    - Model based Integration
- Experiments and Results
- Conclusions

Hinton's Distillation

- Training many different models on same training data:
  - Improves the performance, but
  - Makes the whole model big and unsuitable in practice.
- How to train single, small model with simlar performance?
  - Use the output of the big model as "soft" targets for the small model – *Model Compression* (Caruana, 2006).
- When references, i.e. "hard" targets, are available (Hinton, 2015):
  - Combine the "soft" and "hard" targets and control the **softness** of the "soft" targets.
  - The big model is called *teacher* and the small one – *student*.

Hinton's Distillation

- Given the c-class classification task with training data $\{(x_i, y_i)\}_{i=1}^{n} \sim P^n(x, y),\ \ x_i \in \mathbb{R}^d, y_i \in \mathbb{Q}^c$, where $\mathbb{Q}^c$ is a space of c-dimensional probability vectors, teacher training is to find:

$$f_t = \arg\min_{f \in \mathcal{F}_t} \frac{1}{n} \sum_{i=1}^{n} l\left(y_i, \sigma\left(f(x_i)\right)\right) + \Omega(\|f\|)$$

where $\sigma()$ is a softmax, $l()$ is the loss, and $\Omega()$ is a regularizer.

- Then, for the student we have:

$$f_s = \arg\min_{f \in \mathcal{F}_s} \frac{1}{n} \sum_{i=1}^{n} \left[(1-\lambda) l\left(y_i, \sigma\left(f(x_i)\right)\right) + \lambda l\left(s_i, \sigma\left(f(x_i)\right)\right)\right]$$

where $s_i = \sigma(\frac{f_t(x_i)}{T}) \in \mathbb{Q}^c$ and $T > 0$ controls the smoothness.

Vapnik's Privileged Information

- Often during training some additional information is available which is **not** accessible during test time. Given training data

$$\{(x_i, x_i^*, y_i)\}_{i=1}^{n} \sim P^n(x, x_i^*, y)$$

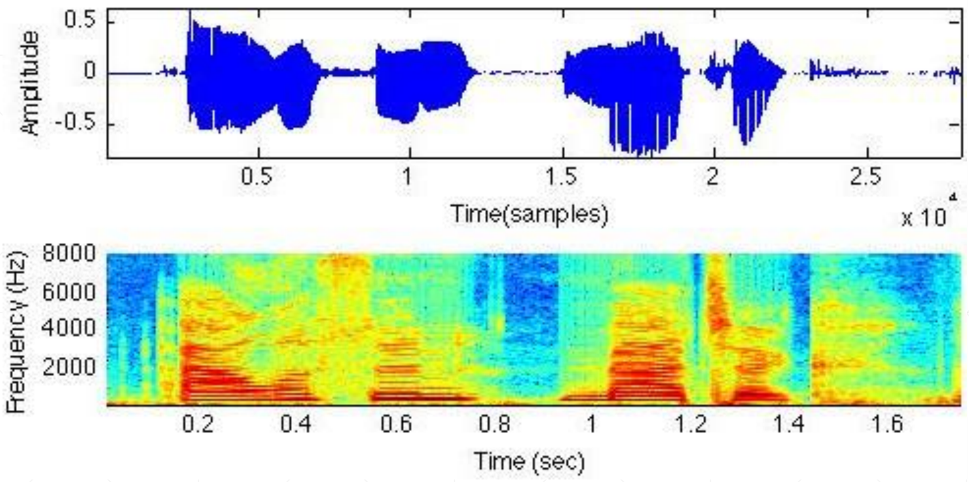
- How to leverage this information to make better model?
  - The naïve way – estimate the mapping $x \xrightarrow{f} x^*$ and generate x^* during testing.
  - Vapnik's way (restricted to SVMs):
    - *Similarity Control* (Vapnik, 2009). Implemented in SVM+ objective.
    - *Knowledge transfer* (Vapnik, 2015). Train f_t on $\{(x_i^*, y_i)\}_{i=1}^{n}$ and use it during the training of f_s on $\{(x_i, y_i)\}_{i=1}^{n}$.

Generalized Distillation

- *Combination* of Hinton's distillation and Vapnik's privileged information approaches (Lopez-Pas, 2016).
- Three step process. Given training data $\{(x_i, x_i^*, y_i)\}_{i=1}^n$
  1. Learn teacher $f_t \in \mathcal{F}_t$ using $\{(x_i^*, y_i)\}_{i=1}^n$;
  2. Compute teacher "soft" labels $s_i = \sigma\left(\frac{f_t(x_i^*)}{T}\right)$ for some temperature T;
  3. Learn student $f_s \in \mathcal{F}_s$ using both $\{(x_i, s_i)\}_{i=1}^n$ and $\{(x_i, y_i)\}_{i=1}^n$, distillation objective and imitation parameter $\lambda \in [0,1]$.
- Generalized distillation reduces to:
  - Hinton's distillation when $x_i = x_i^*$.
  - Vapnik's method when x_i^* is privileged description of x_i.

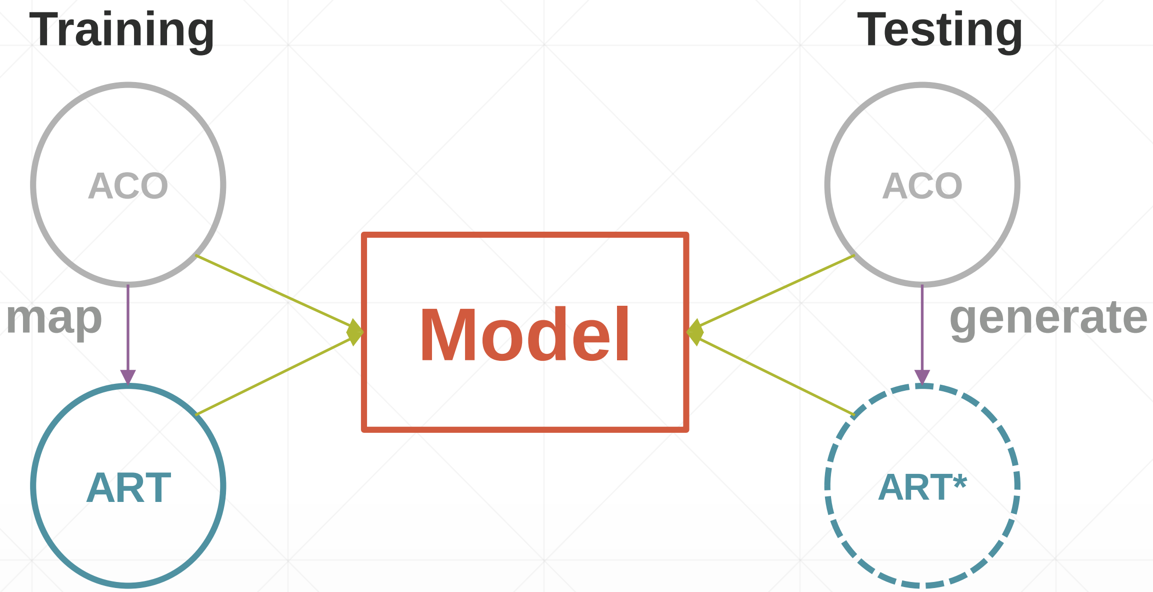
Application in Speech Recognition

- Features for ASR:
  - Spectrum based – MFCC, FBANK, etc.
    - Main features, widely used.
    - Easy to obtain.
    - Highly variable.
    - Affected by noise, etc.
  - Articulatory movements based.
    - Not affected by noise.
    - Less variable.
    - Difficult to obtain – EMA, X-rays, MRI.
    - Impractical for real time ASR.

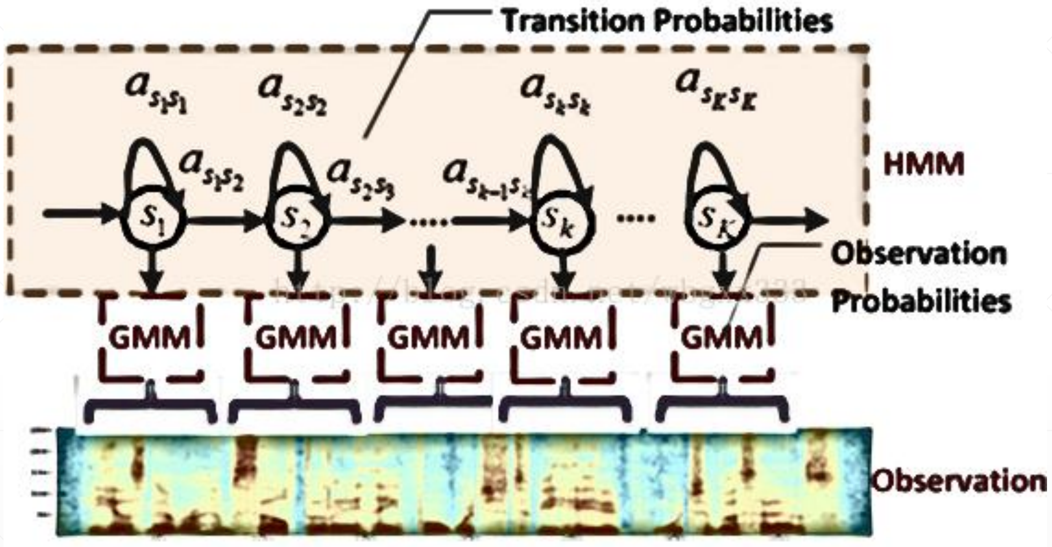


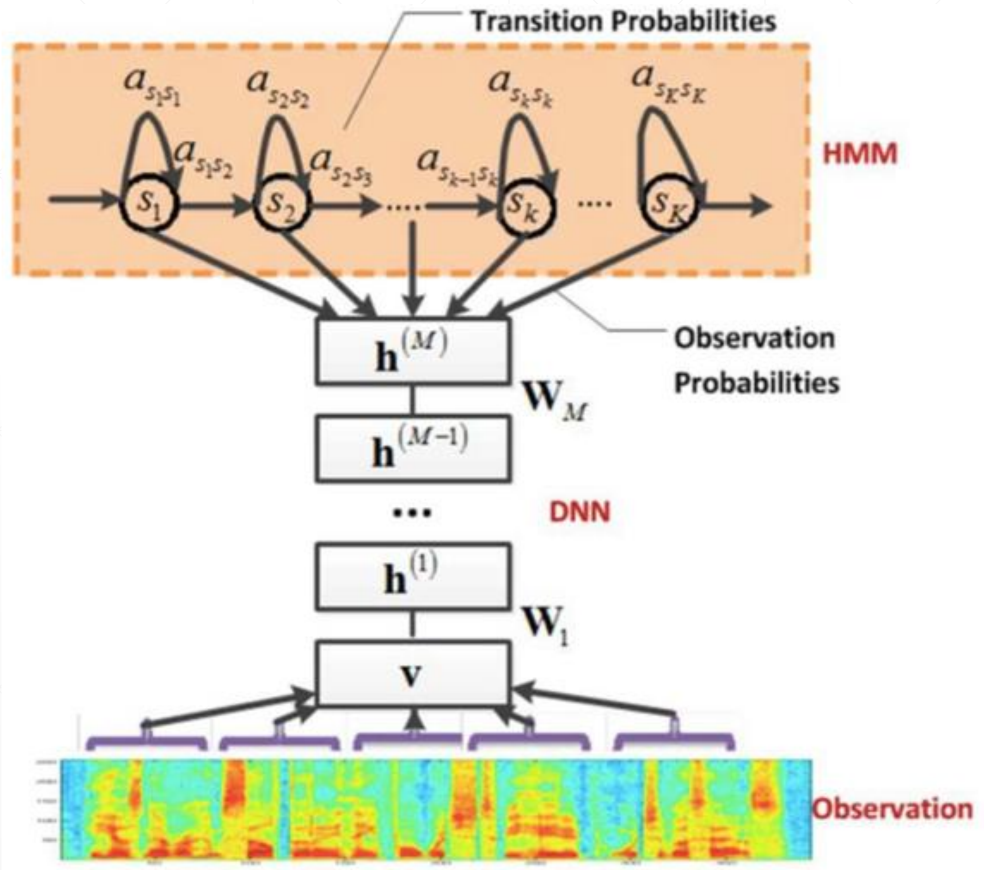
Articulatory and Spectrum Feature Integration

- **Feature based.**
  - Articulatory Inversion.
  - Most popular approach.
  - Mapping with various regression techniques.
- **Model based.**
  - More difficult.
  - HMM/BN (Markov, 2006)
  - Generalized Distillation (this work).

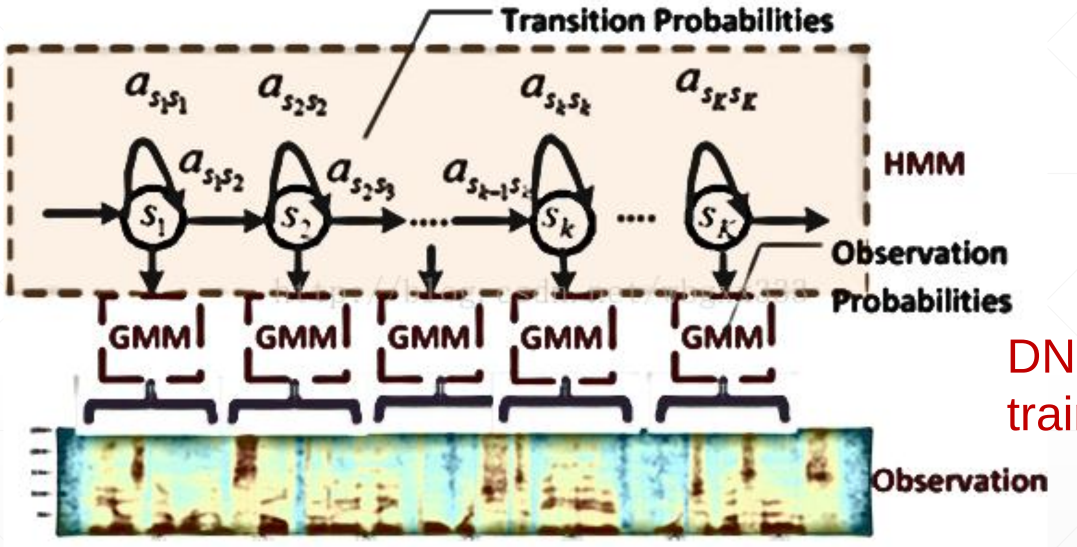


GMM-HMM versus DNN-HMM AMs

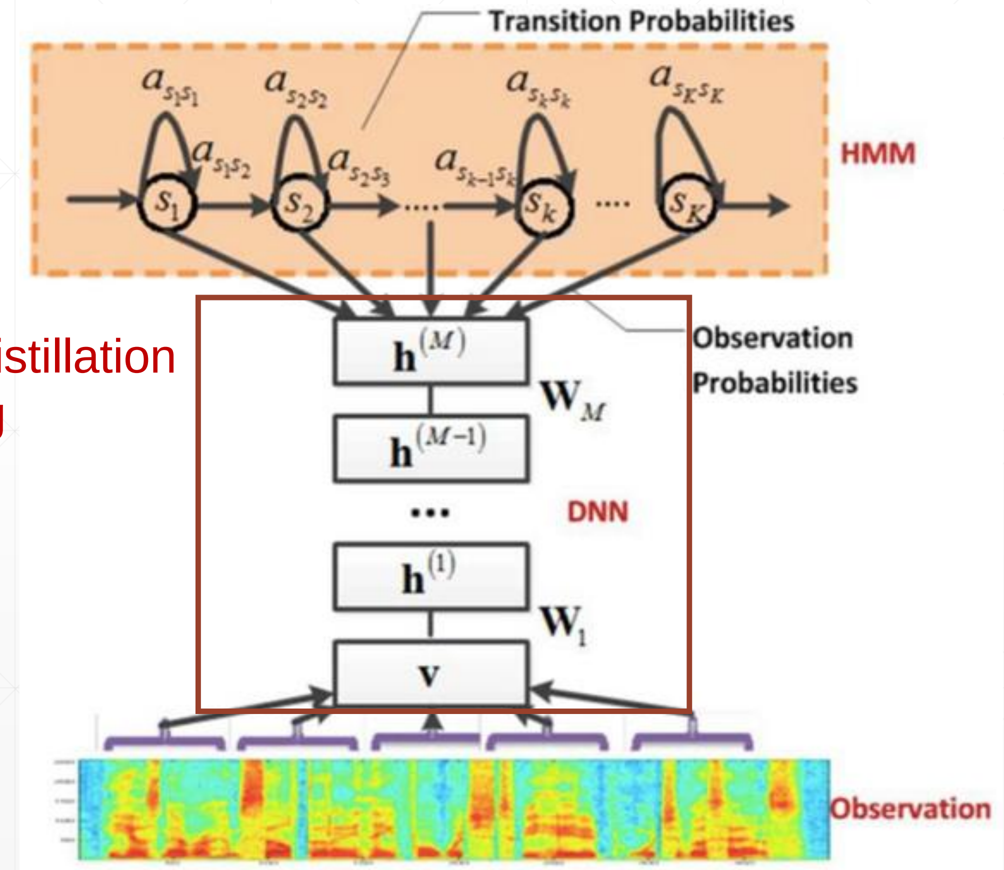




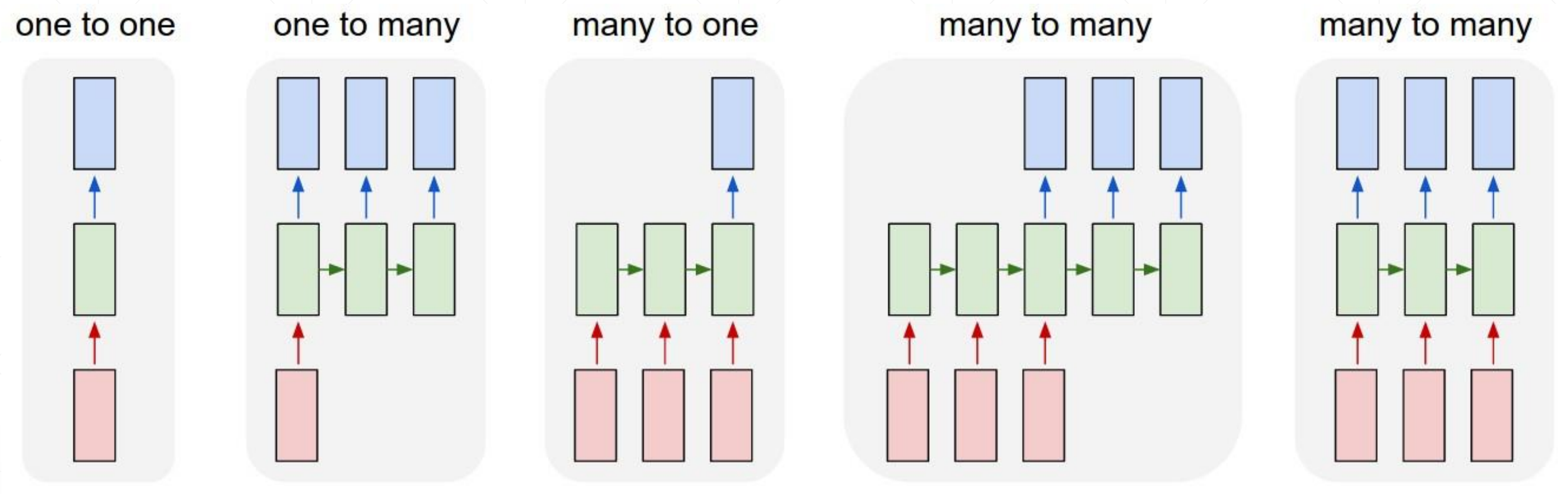
GMM-HMM versus DNN-HMM AMs



DNN distillation
training

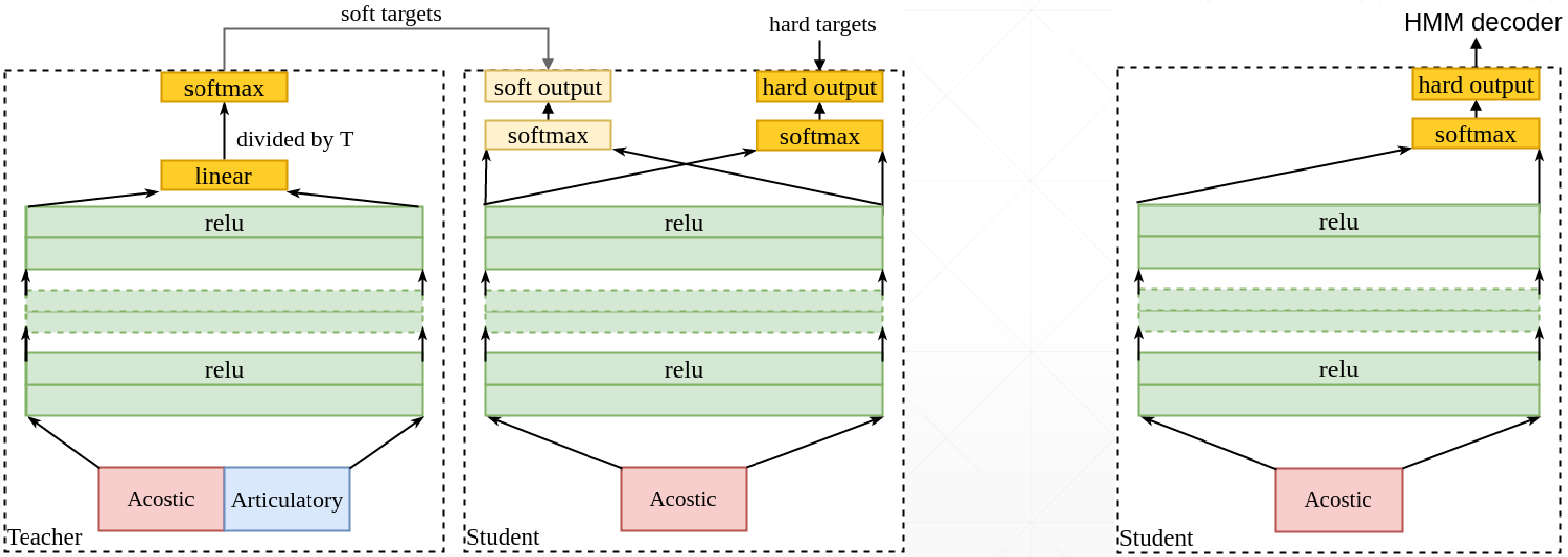


DNN Variants



- **Feed-Forward**: can learn one-to-one mapping
- **Recurrent**: can learn mapping between two sequences

DNN Distillation Training and Testing

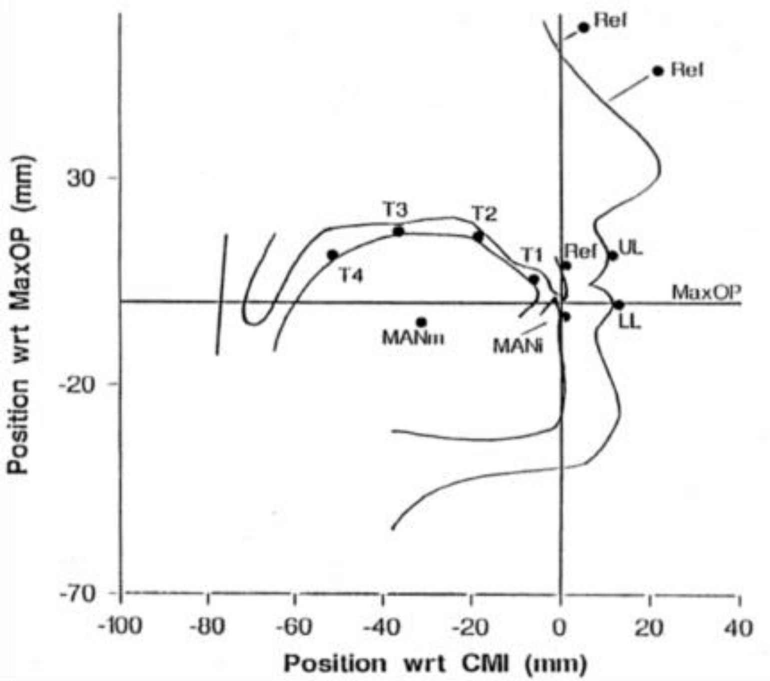


Training

Testing

Experiments

- Database.
  - University of Wisconsin X-ray micro-beam database (XRMB).
  - Consists of **simultaneously** recorded acoustic and articulatory measurements from 47 American English speakers.
- Features
  - Acoustic – MFCC (39 dim.)
  - Articulatory – Displacement of 8 articulatory points (16 dim.)
  - All feature vectors normalized and synchronized.

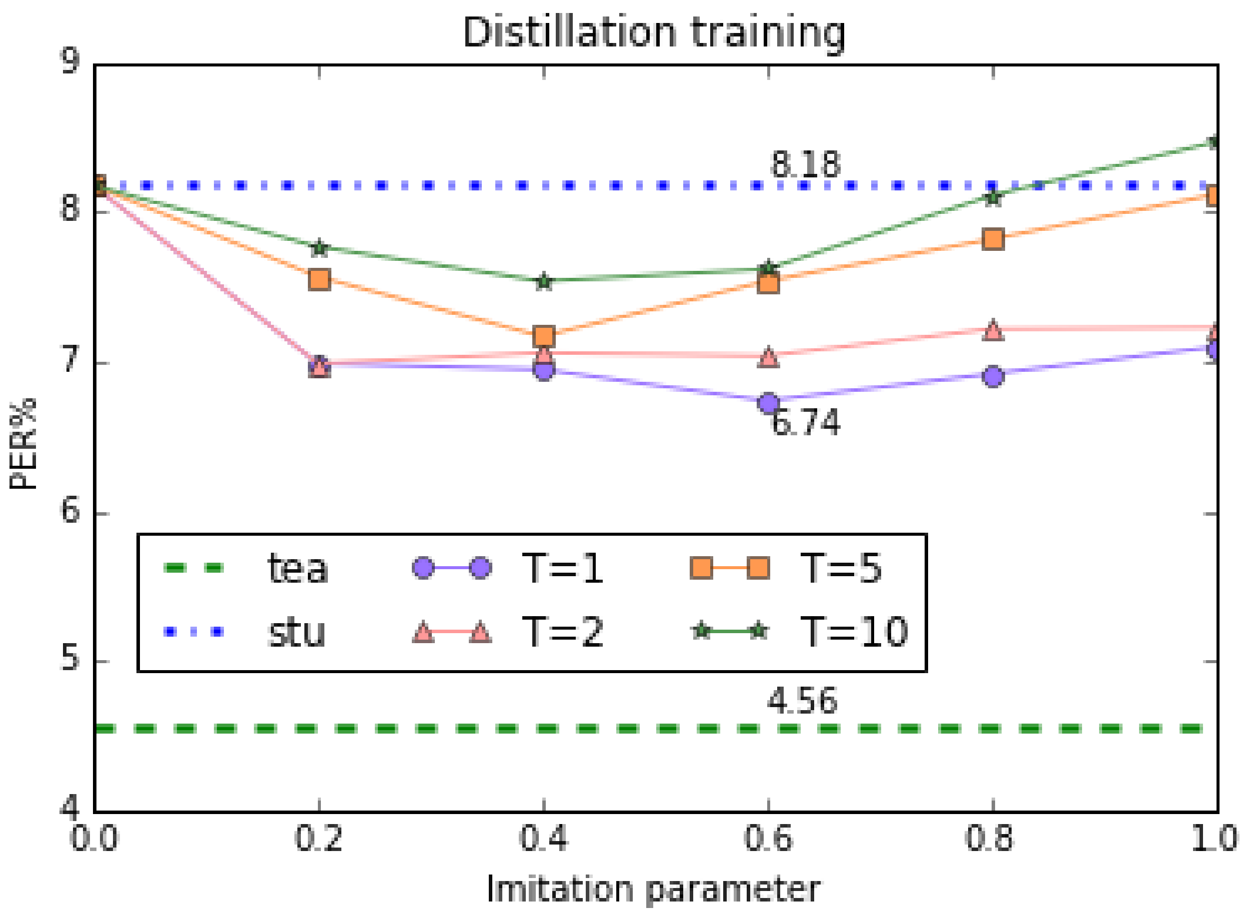


Experiments

- Training procedure
  1. Train conventional GMM-HMM model using both acoustic and articulatory features.
  2. Perform forced alignment to obtain DNN “hard” targets.
  3. Train teacher DNN using both acoustic and articulatory features.
  4. Train student DNN using acoustic features only and guided by the “teacher”.
- Testing procedure
  1. Use student DNN with acoustic features only to obtain HMM state probabilities.
  2. Use standard HMM decoding (Viterbi) to obtain recognition result.
- Evaluation metric – Phoneme Error Rate (PER)

Results

- DNN parameters:
  - Feed Forward.
  - Input widow – 17 frames.
  - Activation – ReLU.
  - Dropout – 40%.
  - Teacher DNN
    - 5 layers
    - 3073 nodes.
  - Student DNN
    - 4 layers
    - 2048 nodes.



Results

By DNN type:

Feed Forward

vs.

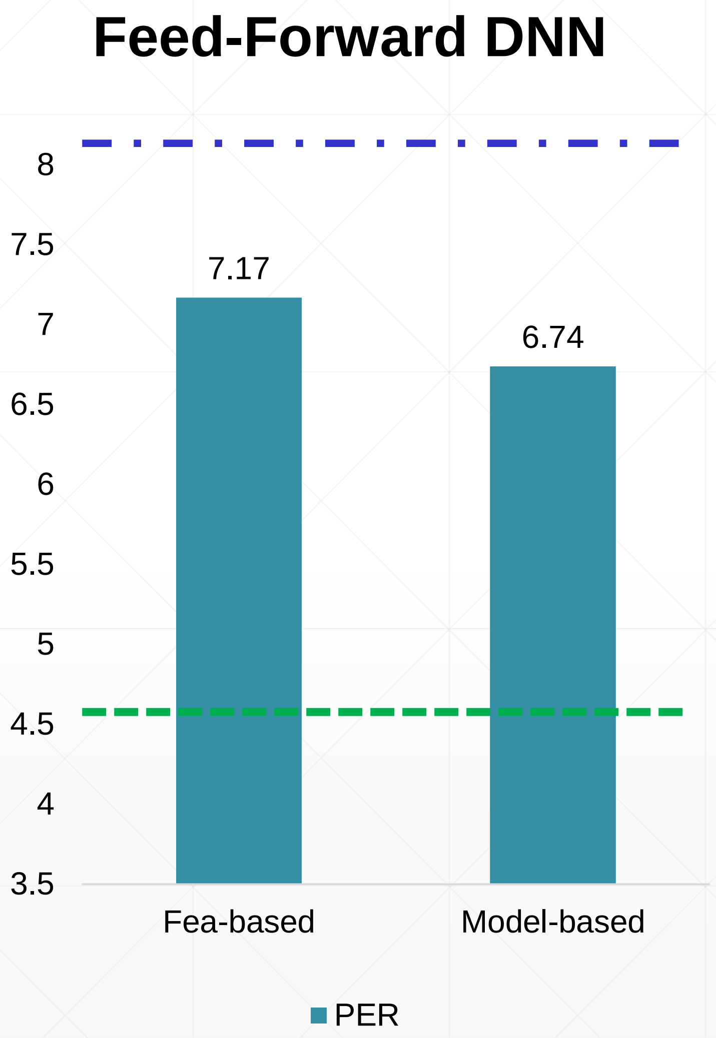
Recurrent

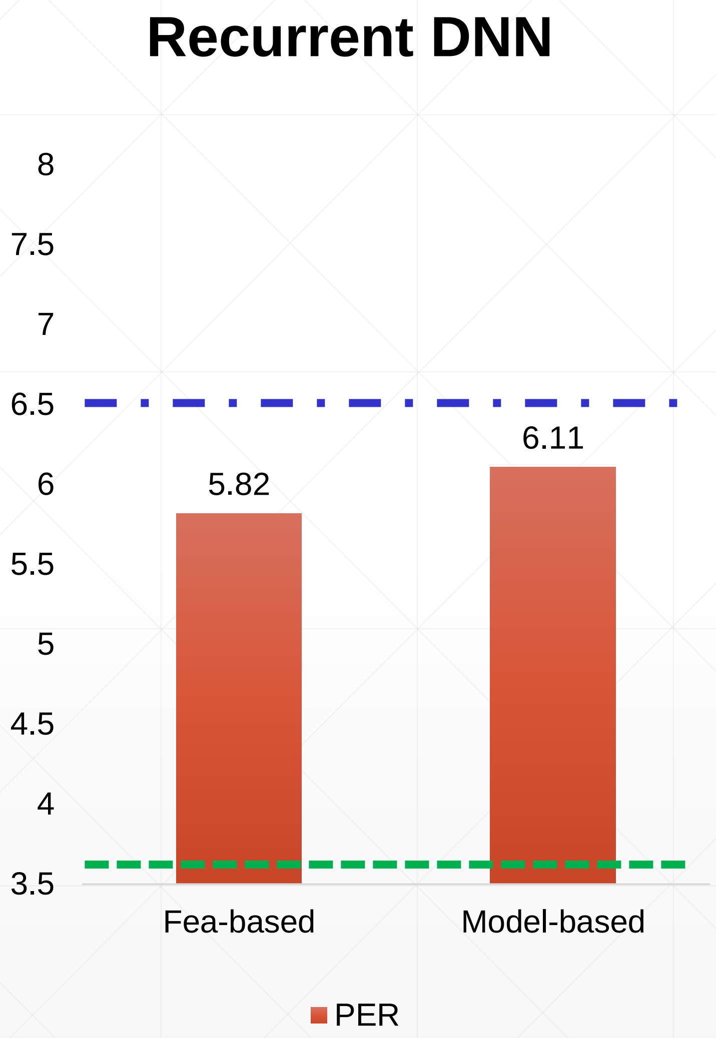
By integration type:

Feature based

vs.

Model based





Conclusions

- **Generalized distillation:**
  - Is an effective method for model based integration of information unavailable at testing time.
  - Allows smaller student models (4 layers / 2048 nodes) to reach performance close to bigger teacher models (5 layers / 3072 nodes).
- **DNN structure:**
  - Recurrent DNNs outperform Feed-Forward DNNs in the ASR task since they better model long-term temporal dependencies.
  - Time complexity of Recurrent DNNs is higher than Feed-Forward DNNs.
- **Integration approach:**
  - Model based and Feature based integration achieve comparable results.
  - Feature based integration requires higher computational power.