Transfer Learning: from Bayesian Adaptation to Teacher-Student Modeling

Chin-Hui Lee
School of ECE, Georgia Tech
chl@ece.gatech.edu

Outline

- Transfer learning: an introduction
  - ➢ Avoiding catastrophic forgetting by adaptation
  - ➢ Maintaining performances in adapted conditions
- Transfer learning of generative models: Bayesian
- Transfer learning of discriminative network models
  - ➢ Direct Bayesian learning via the same neural network
  - ➢ Indirect Bayesian learning via bottleneck features
- Teacher-student learning: three speech examples
  - ➢ Adapting student models with auxiliary teacher networks
  - ➢ Going beyond conventional Bayesian adaptation capability
- Summary

Transfer Learning: An Introduction

- **Transfer learning:** knowledge of previously already learned models for Task A is adapted to models for Task B with some adaptation data from new Task B
- **Issues and challenges** in transfer learning
  - A large number of model parameters to adapt but with only limited amounts of adaptation data
  - Catastrophic forgetting in transfer learning: when adapting to specific new test conditions (Task B), knowledge learnt in the training Task A might be lost
  - Performances of Task B often degraded from Task A
  - Differences with generative (e.g., probability density function) and discriminative (e.g., DNN) models

Robustness Issue: Speech Recognition

Due to the training/test mismatch, *performance of a recognition system in-the-field may not reflect performance measured during system design*

Resource Management Task (1000 words)

RM Task	WER
Native Speakers	3.6 %
Non-Native Speakers Telephone Channel	34.9%

Wall Street Journal Task (5000 words)

WSJ0 5K Task (Nov92)	WER
Native speakers	4.7 %
Non-Native Speakers	29.1%

10-fold increase in WER!

Topic 1: Transfer of Generative Models

- Task A learning is summarized in a likelihood function, $f(X|\theta)$, *with* $\theta = \theta_A$ as parameter learned from training data X for inferencing, and θ_A is often estimated via maximum likelihood (ML)
- Learned knowledge is often characterized in a prior, $g(\theta \mid \phi)$, with ϕ being the *hyperparameter*, *and* $\theta = \theta_A$, representing what was learned in Task A
- Task B transferring involves learning $\theta = \theta_B$ from a set of adaptation data, Y, through the prior density
- θ_B is often estimated via maximum a posteriori (MAP) and a conjugate prior $g(\theta \mid \phi)$ is often chosen

MAP Adaptation: Motivations

- Providing a *mathematically well-founded* and *optimal* way for combining an existing model and new data into a new model in transfer learning
- Offering a natural way for domain adaptation to new speaker, channel, environmental and others
- Achieving asymptotically equivalence to ML as the amount of adaptation data increases
- Solving MAP is similar to ML with conjugate priors
- Speaker adaptation for ASR, TTS and SID/SV

MAP versus ML Estimation

- Given density form $f(.|\theta)$ and a set of training observations X *or adaptation data* X, we want to estimate the parameter vector θ
  - If θ is fixed but unknown

$$\boxed{\theta_{MLE} = \arg\max_{\theta} f(X|\theta)}$$

  - If θ is random, with a given prior density $g(\theta)$

$$\boxed{\begin{aligned}\theta_{MAP} &= \arg\max_{\theta} f(\theta|X) \\ &= \arg\max_{\theta} f(X|\theta)g(\theta)\end{aligned}}$$

A "Good" Prior: Conjugate Density

- $f(X|\theta)$ has sufficient statistic $t(X)$ of finite dimension for θ if it can be factorized: $f(X|\theta) = h(X)\, k(\theta \mid t(X))$
  - $h(X)$ is independent of θ
  - Kernel density $k(\theta|t(X))$ depends on X only through $t(X)$
- $k(X|\theta)$ is called the conjugate family of $f(X|\theta)$
- If the prior density $g(\theta) = k(\theta|\phi)$ is a member of the conjugate family, then posterior $f(\theta|X) = k(\theta \mid \phi')$
- Under this condition, *the ML and MAP optimization problems are similar*: finding the mode of $k(\theta \mid X)$
  - ML: $\theta = \mathrm{argmax}_\theta\, f(X|\theta) = \mathrm{argmax}_\theta\, h(X)\, k(\theta \mid t(X))$
  - MAP: $\theta = \mathrm{argmax}_\theta\, f(\theta|X) = \mathrm{argmax}_\theta\, k(\theta \mid \phi')$

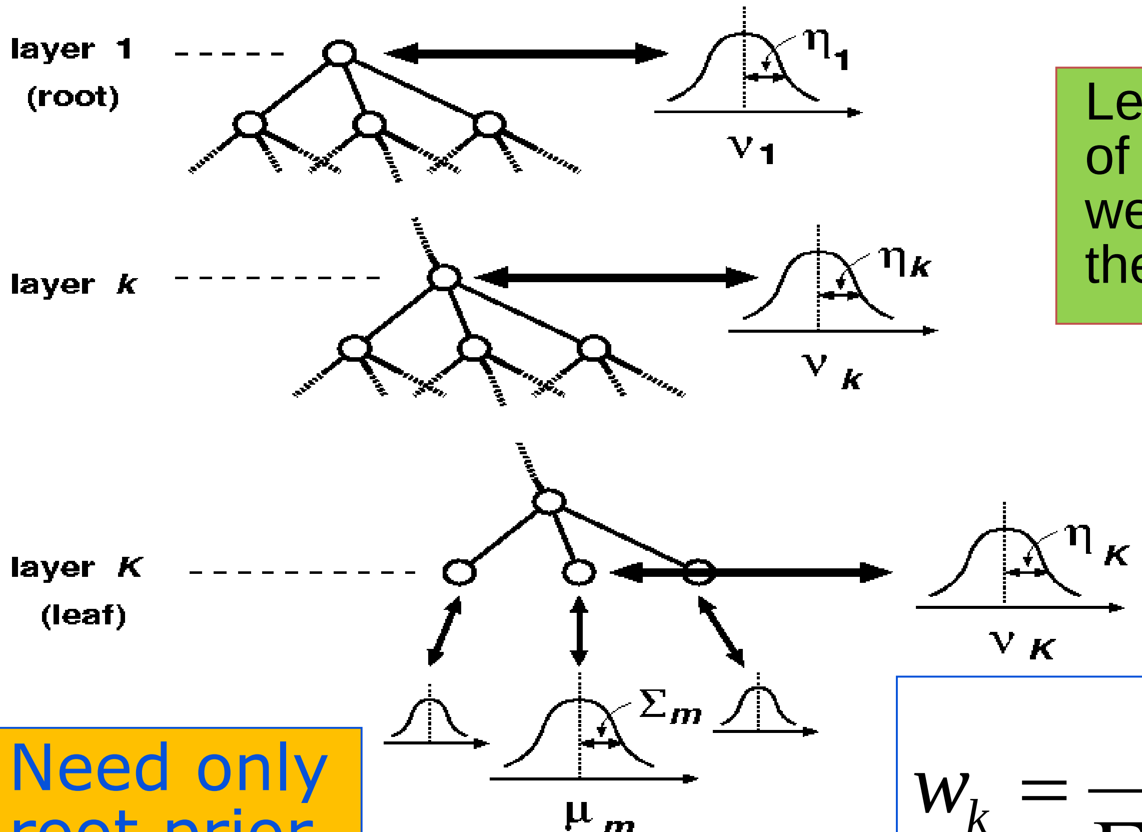
MAP has been developed for many useful pdf, including HMM

Estimation of Hyperparameters

- Hyperparameters are often more than parameters
- The value of these hyperparameters is key to MAP adaptation, i.e., controlling adaptation quality
- One key issue for practical Bayesian deployment
- Potential solutions for estimating hyperparameter ϕ
  - Hyperparameter tying (e.g., structural MAP or SMAP)
  - Ad-hoc settings (e.g., tuning on a development set)
  - Empirical Bayes (e.g., learning from training data)
  - Spatial, temporal and incremental prior evolution (e.g., online adaptation, Huo and Lee, T-SAP 1997)

Key reference: Lee and Huo, *Proceedings IEEE*, August 2000

SAMP: Recursive Estimation with Hierarchical Prior Evolution in a Tree



Leaf mean is a weighted sum of the parent means in which weights are determined by the amount of data and priors

$$\widehat{\nu}_K^{m} = \sum_k w_k \widehat{\nu}_k^{mk}$$

$$w_k = \frac{\Gamma_k}{\Gamma_k + \tau_k} \prod_{i=k+1}^{K} \frac{\tau_i}{\Gamma_i + \tau_i}$$

Need only root prior

Key reference: Shinoda and Lee, *IEEE T-SAP*, 2001

Topic 2: Transfer for Discriminative Models

- Mostly derived for deep neural networks (DNNs)
- Bayesian transfer learning of DNNs (Bayesian DNN)
  - ➢ Speaker adaptation on the same network but adapting only a subset of parameters (with little adaptation data)
  - ➢ Using prior density to avoid catastrophic forgetting
- Transfer learning via teacher-student modeling
  - ➢ Using an auxiliary network (teacher) to adapt the task network (student) which can be with more parameters than the teacher network, but with less data to adapt
  - ➢ GAN can be used to generate more adaptation data

Key reference: Huang, Siniscalchi and Lee, *Neorocomputing*, 2016

General Transfer Learning Settings

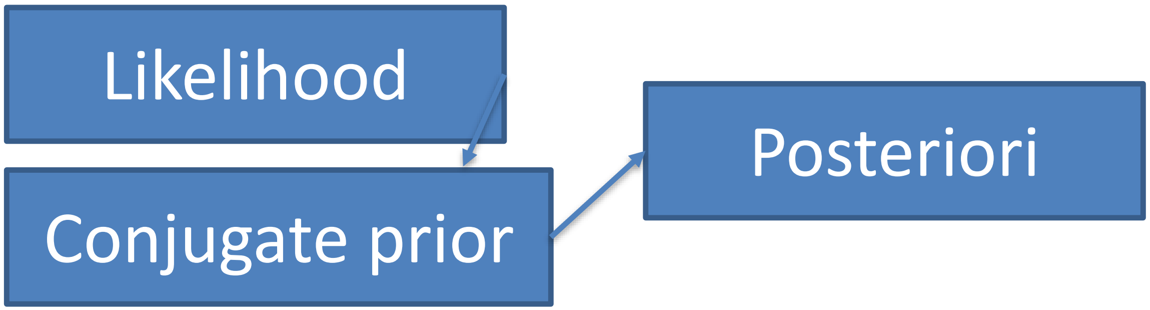
- Existing DNN to be adapted
  - Bayesian learning: SI acoustic model for LVCSR
  - Teacher-student learning (teacher as “prior”)
    - Audio-visual ASR: audio-only AM as teacher
    - CHiME-4: SE DNN as teacher to help ASR student
    - SE DNN: ASR as teacher to help SE DNN student
- Cross-domain transfer learning metrics
  - Bayesian learning: approximate likelihood from new data together with prior density from existing DNN model
  - Teacher-student learning
    - Audio-visual ASR: KL divergence between outputs of teacher and student DNNs
    - CHiME-4: teacher output to evaluate improved speech presence probability dynamically to serve as learning target for student
    - SE DNN: two ASR teachers generate KL to update SE model

Bayesian DNN Adaptive Learning

- Direct adaptation of discriminative DNNs
  - Adapting a small parameter subset while keeping the other parameters frozen (avoid catastrophic forgetting)
  - MAP/SMAP adaptation for DNNs using Gaussian priors
- Indirect adaption on converted generative DNNs
  - Utilizing bottleneck feature (BN) derived from DNN
  - MAP/SMAP adaptation for GMMs with BN features
- Bayesian system combination (not here) of the two adaptive models with a few thousand sets of weights
  - Leveraging upon complementarity of the discriminative and generative models adapted with totally different methods
  - Using same adaptation set to train all adaptation weights

Bayesian DNN Adaptation: Framework

- MAP/SMAP adaptation for GMM based system
  - GMM as generative pdf: straightforward



- MAP/SMAP adaptation for DNN based system
  - DNN as a discriminative function: generative output posterior form not directly available

How to perform Bayesian adaptation for DNN?

Direct Bayesian Adaptation of DNNs

- First step:
  - Look at the DNN as an approximation of a pdf
  - Explain the DNN objective function in a probabilistic way (likelihood): $L = \log p(\boldsymbol{o}_t|\boldsymbol{W})$

- Second step
  - Estimate posterior not likelihood parameters

$$L_{MAP} = \log p(\boldsymbol{W}|\boldsymbol{o}_t) = \log \frac{p(\boldsymbol{o}_t|\boldsymbol{W})p(\boldsymbol{W})}{p(\boldsymbol{o}_t)} \propto \log p(\boldsymbol{o}_t|\boldsymbol{W}) + \log p(\boldsymbol{W})$$

Likelihood + Conjugate prior = Posteriori

prior

Objective Functions and Prior Forms

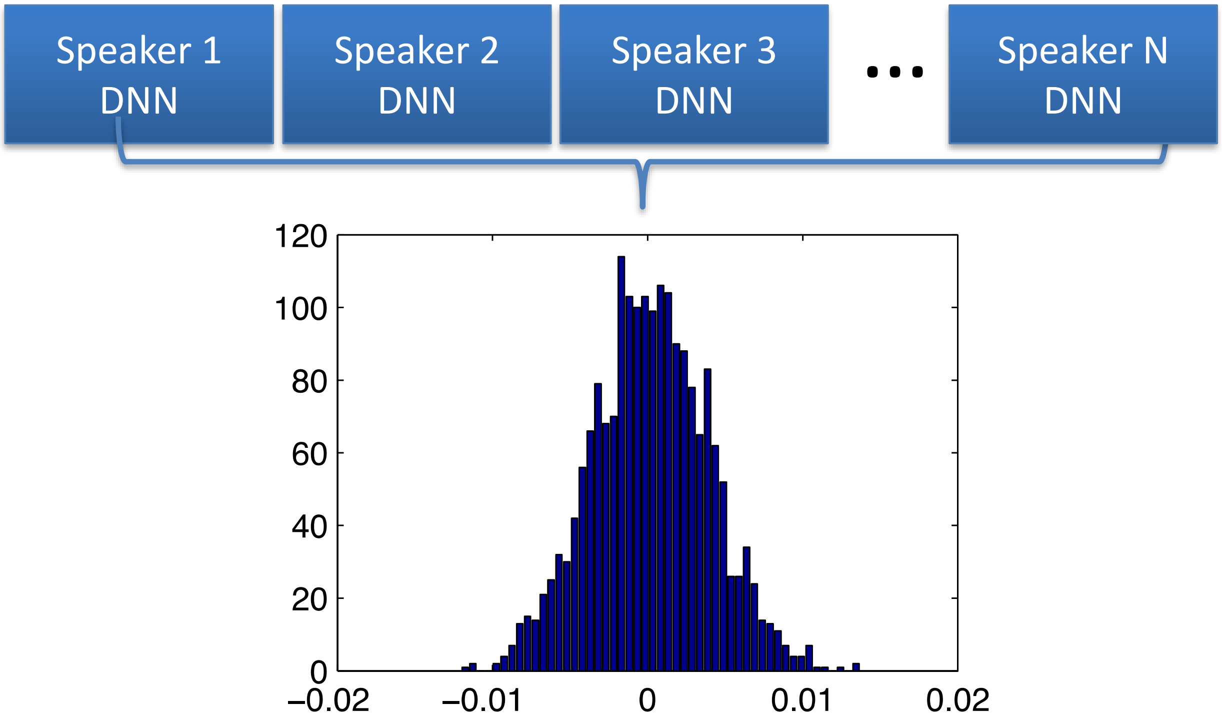
$$L_{MAP} = L + \log p(\boldsymbol{W})$$

?

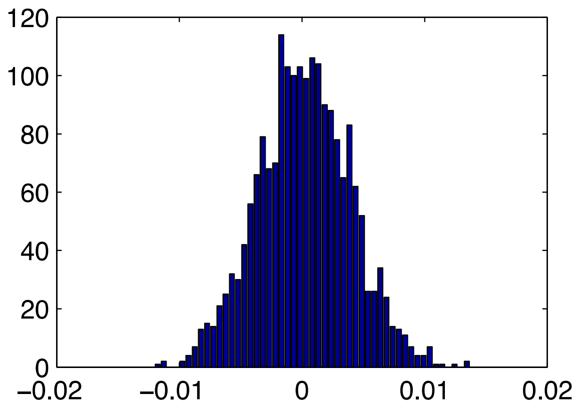
Cross-Entropy (CE)
Maximum Mutual Information (MMI)
Minimum Phone Error (MPE)
Minimum Classification Error (MCE)
…

Prior Estimation: Empirical Bayes

- Performing adaptation on all training speakers, and then analyze the parameter distribution across them
- Treating each adapted DNN as observed samples



Prior Estimation Cont'd



vectorize

$$p(\boldsymbol{W}) = \lambda \exp(-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{u})^* \boldsymbol{\Sigma}^{-1}(\boldsymbol{w} - \boldsymbol{u}))$$

$\boldsymbol{u}$**=0** $\boldsymbol{\Sigma}$**= I**

$$p(\boldsymbol{W}) = \lambda \exp\{-\frac{1}{2} tr(\boldsymbol{W} - \boldsymbol{M})^* \boldsymbol{\Sigma}^{-1}(\boldsymbol{W} - \boldsymbol{M})\boldsymbol{\phi}^{-1}\}$$

Multivariate Gaussian

Used in final formulation

Matrix variate Gaussian

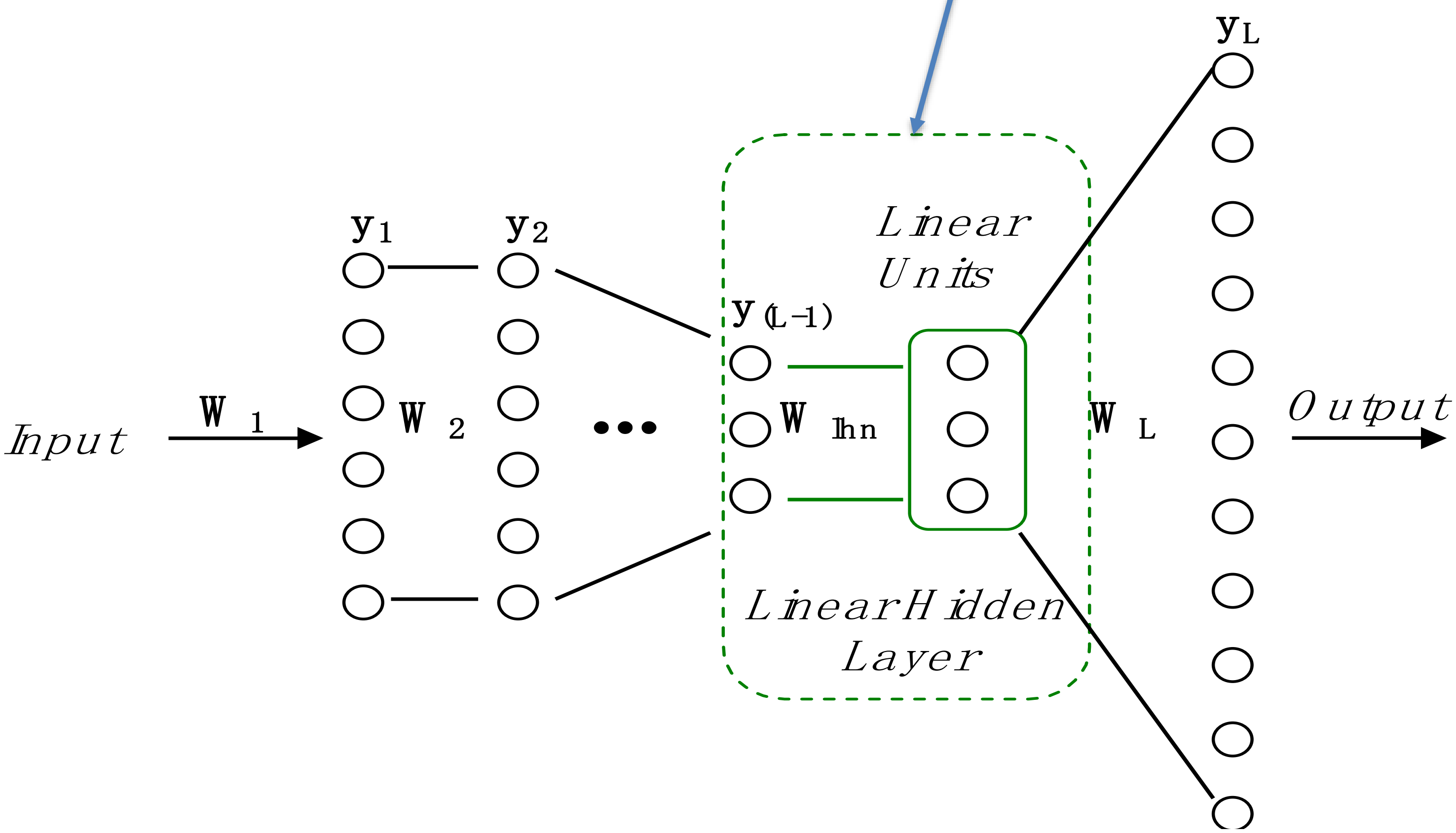
Not used in final formulation

Multivariate Gaussian prior easily reduced to L2 regularization

$$L^{MAP} = \log p(\boldsymbol{W}) + L^{CE} = -\frac{\lambda}{2}\boldsymbol{w}^*\boldsymbol{w} + L^{CE}$$

Controlling Number of Parameters

Adapt only weights in an inserted linear hidden layer (LHN): other parameters remain frozen (no catastrophic forgetting), same for LIN and LON, but LHN Is better



Direct DNN Adaptation: A Remark

- **MAP DNN adaptation represents the first Bayesian effort on deep model training/adaptation in literature**
- A paper titled "Overcoming catastrophic forgetting in neural networks" was published in **March 28, 2017**, in "**Proceedings of the National Academy of Sciences**" by Google's **DeepMind** group (Google owns Google Scholar)
- It used almost <u>**exactly the same idea and even the same core equations**</u> as our work published 3 years ago (2014)
- This 2017 paper <u>**didn't cite**</u> our work (we published 2 conferences and 2 journal papers on this topic)
- The first author <u>**admitted his mistake**</u> and said he was unaware of our work before we notified him, but still refused to cite our four papers even after he knew them

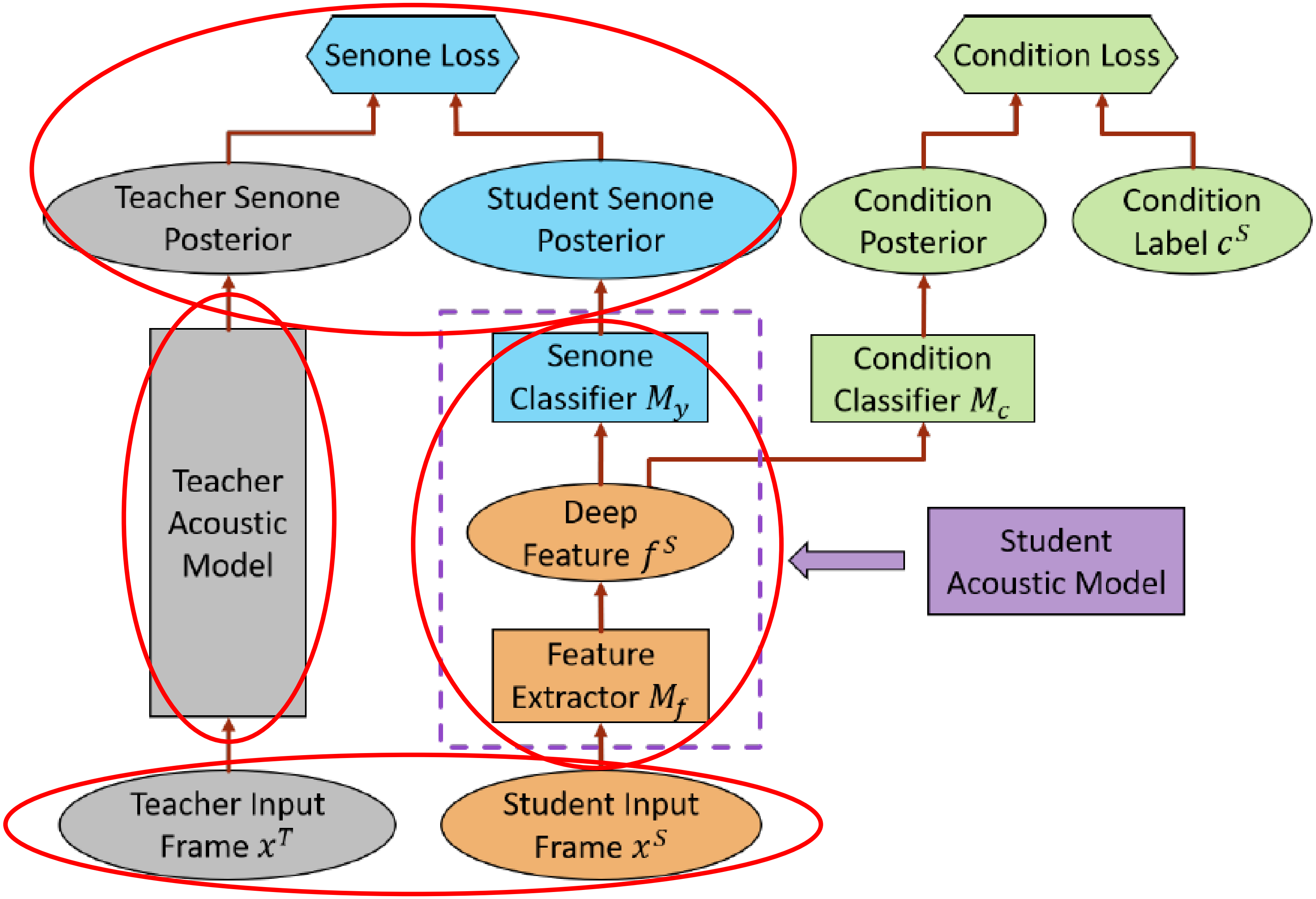
Indirect Bayesian Adaptation of DNNs

- BN features are discriminatively trained data-driven, utilizing DNN's strength in serving as bridge function
- BN features are used to train GMMs at DNN outputs
- To obtain DNN-based features, we can:
  - Train a DNN with a bottleneck layer
  - Train a DNN without BN and do SVD to get BN features
- Rest is straightforward with conventional Bayesian
- Indirect and direct adaptations gave similar results
  - Bayesian combination produced even better results

Topic 3: Transfer via Teacher-Student Models

- Teacher-student learning
  - ➢ Emerged as a new transfer learning framework: typically using an auxiliary network (teacher) to adapt the task network (student) which can be with more parameters than the teacher, but with less data to adapt
  - ➢ Mostly for domain adversarial training or domain adaptive training (DAT)
  - ➢ GAN has been used to generate more adaptation data
- Very flexible and applicable to practical settings
  - ➢ Many new examples have been recently proposed
  - ➢ But a theory is desperately needed

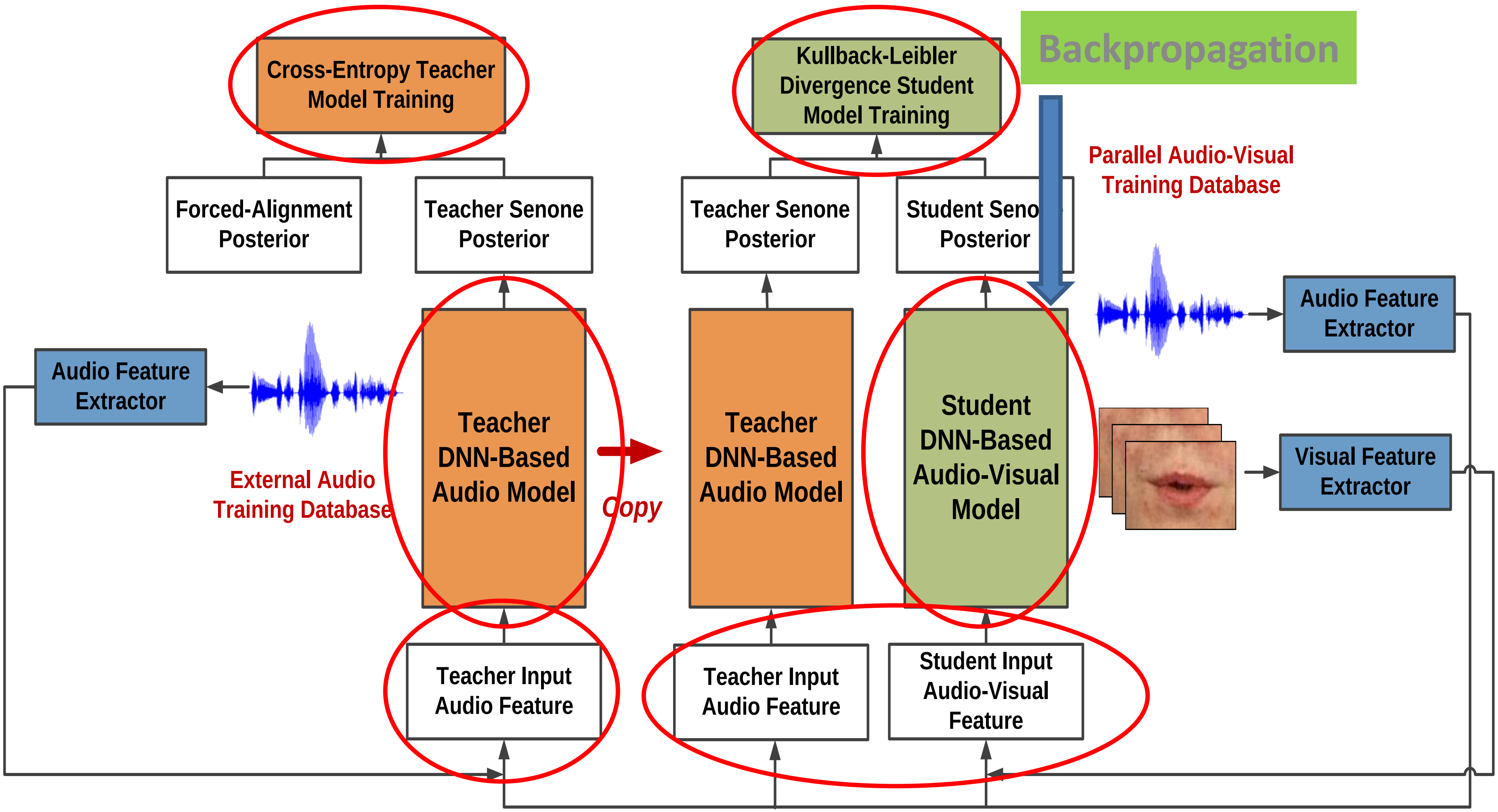
Typical T-S Learning (Meng, *et al*)



Audio-Visual ASR (Li, *et al*, *Icassp2019*)

- ASR degraded drastically in low SNR conditions
- Visual features are fused with audio features to improve over audio-only ASR
- AV data are hard and expensive to collect, limiting the learning capability of DNN-based AV classifiers
- A huge amount of speech data is available to train a good speech-only teacher model
- AV-based student model can be better trained via an already well-trained audio-only teacher model
- Achieving 17% phone error rate reduction, even better with GAN-based data augmentation

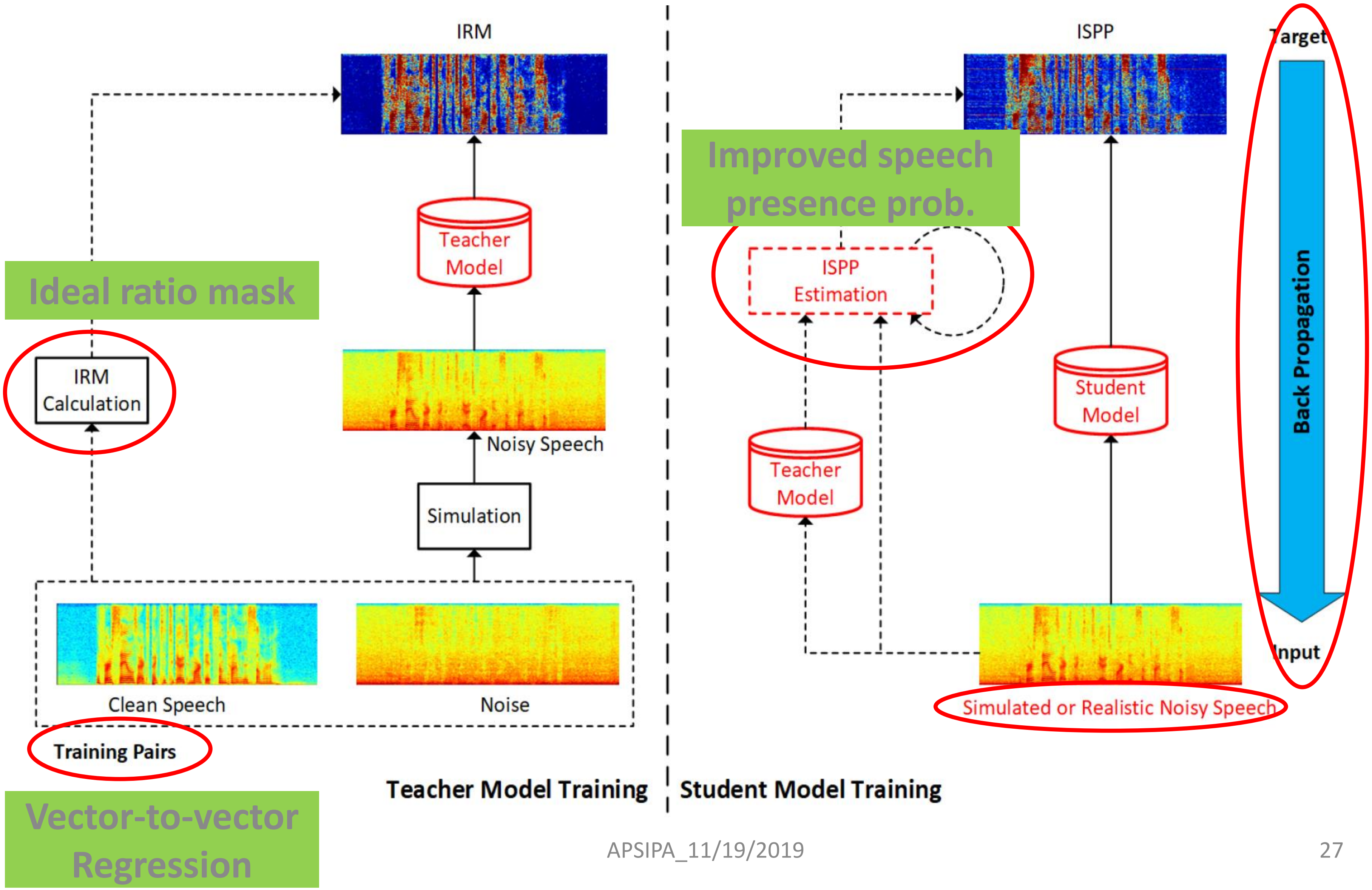
Teacher-Student Learning: Example 1



SE-guided CHiME-4 ASR (Tu, *et al*, *T-ASLP*)

- ASR degraded drastically in unseen noise conditions even with well-trained speech enhancement DNNs
- Clean speech and noises needed in training teacher regression DNN with clean LPS and IRM as targets
- IRM can be used together with ICRMA to estimate improved speech presence probability (ISPP) frame-by-frame (for non-stationary noise) which serves as a new mask target for training student model with only noisy speech collected in adverse conditions
- Speech distortion reduced and continuity maintained
- 8% word error rate reduction from our best CHiME-4

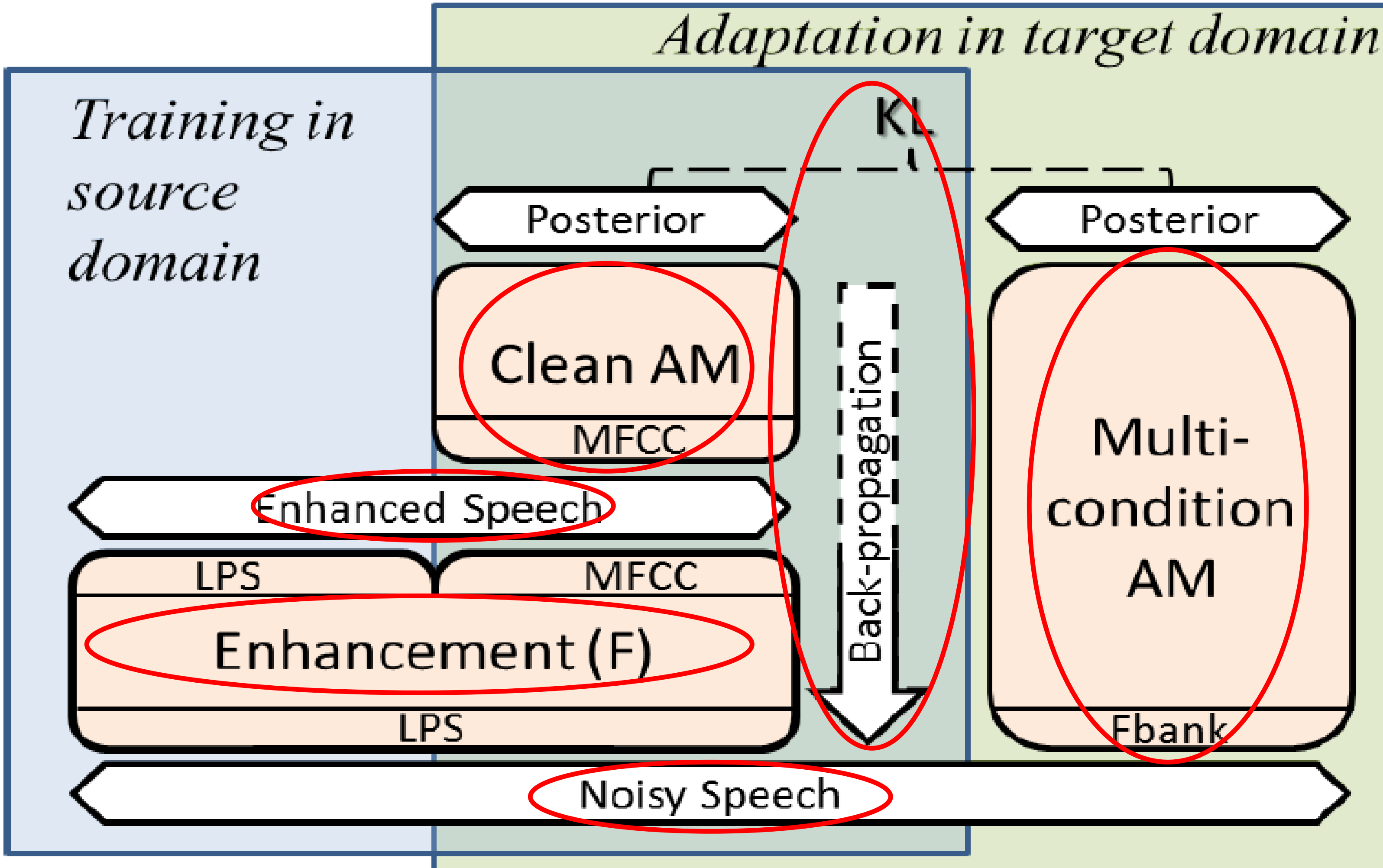
Speech Enhancement helps ASR: Example 2



ASR-Guided SE (Wang, *et al*, ICASSP2020)

- Speech enhancement (SE) performances with deep regression sometimes degrade in unseen noises
- Noisy speech can be fed into a multi-condition-trained acoustic model to generate a set of senone posteriors
- It can also be enhanced by a trained SE DNN and then passed through a clean-trained acoustic model to produce another set of senone posteriors
- KL divergence can be evaluated between the two sets of posteriors and back-propagated via the trained SE DNN in order to update parameters for unseen noises
- Better SE performances achieved with T-S learning

ASR helps Speech Enhancement: Example 3 (submitted to ICASSP2020)



Summary

- Transfer Learning: some more theory needed
  - ➢ Avoiding catastrophic forgetting
  - ➢ Maintaining performances in new conditions
- Using the same network: Bayesian learning
- Via auxiliary networks: teacher-student learning
  - ➢ Example 1: audio-only to help audio-visual ASR
  - ➢ Example 2: speech enhancement to help ASR
  - ➢ Example 3: ASR to help speech enhancement
- ▪ Plenty of new adaptation scenarios & opportunities

Thank You !!!