A New Perspective on Combining GMM and DNN Frameworks for Speaker Adaptation

Natalia Tomashenko[1,2,3]
natalia.tomashenko@univ-lemans.fr
Yuri Khokhlov[3]
khokhlov@speechpro.com
Yannick Esteve[1]
yannick.esteve@univ-lemans.fr

[1]University of Le Mans, France
[2]ITMO University, Saint-Petersburg, Russia
[3]STC-innovations Ltd, Saint-Petersburg, Russia

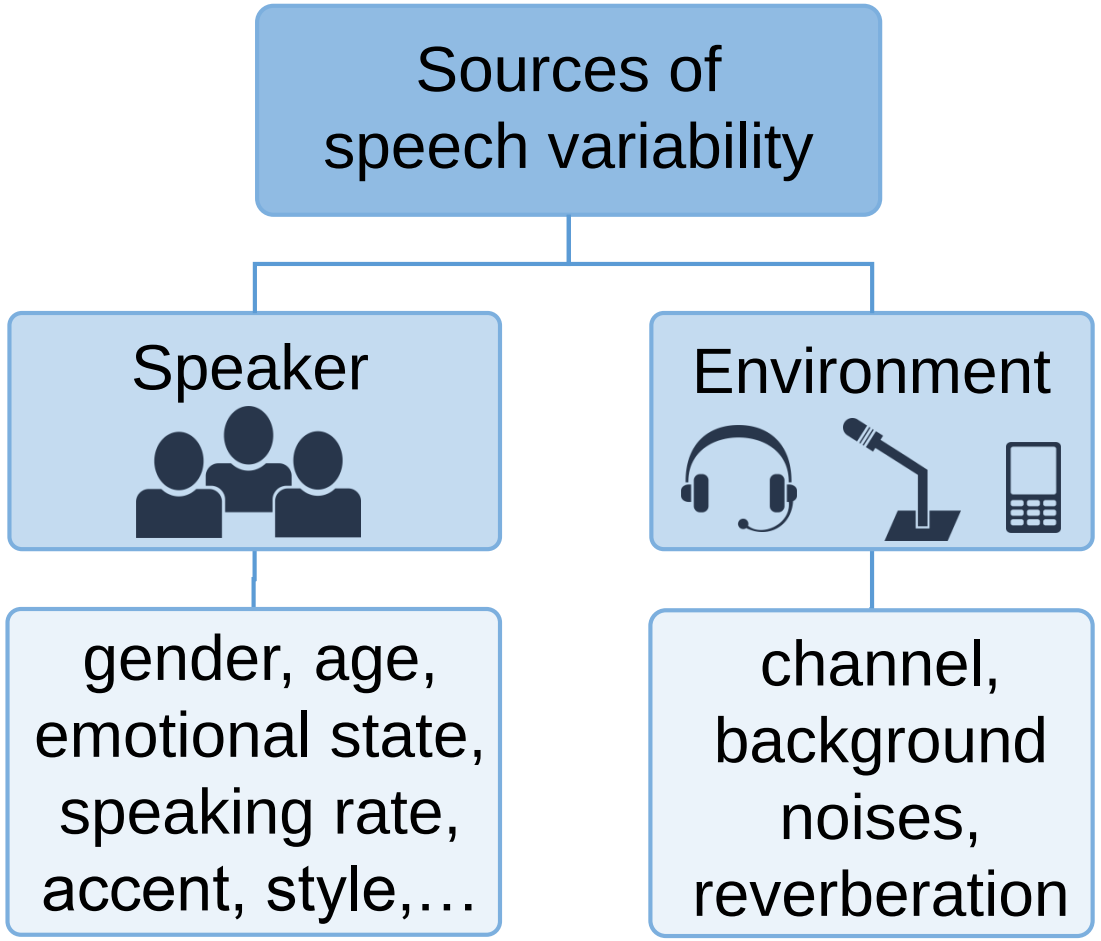
Outline

1. Introduction
   - Speaker adaptation
   - GMM vs DNN acoustic models
   - GMM adaptation
   - DNN adaptation: related work
   - Combining GMM and DNN in speech recognition
2. Proposed approach for speaker adaptation: GMM-derived features
3. System fusion
4. Experiments
5. Conclusions
6. Future work

Outline

1. Introduction
   - Speaker adaptation
   - GMM vs DNN acoustic models
   - GMM adaptation
   - DNN adaptation: related work
   - Combining GMM and DNN in speech recognition
2. Proposed approach for speaker adaptation: GMM-derived features
3. System fusion
4. Experiments
5. Conclusions
6. Future work

Adaptation: Motivation

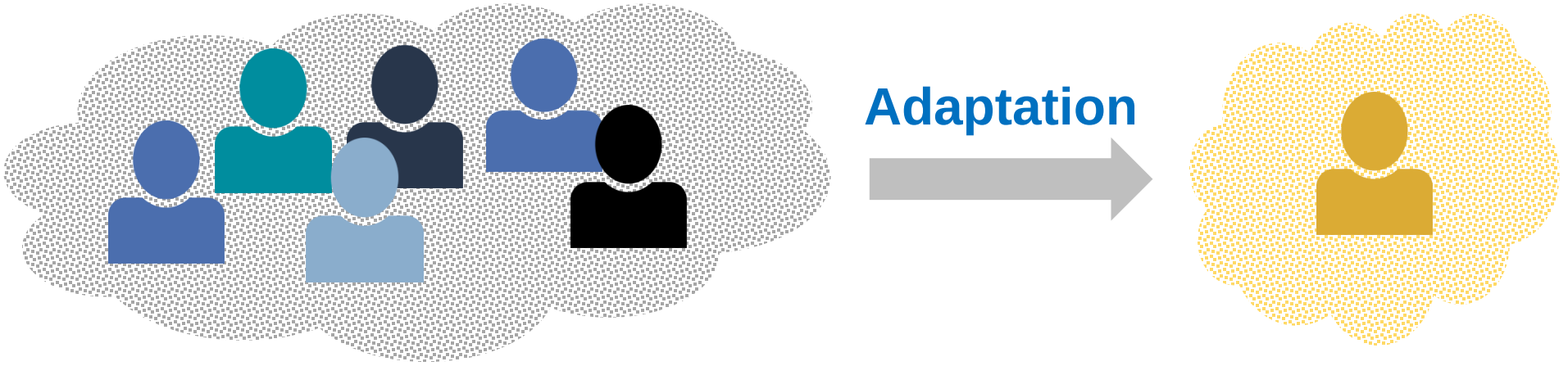


Why do we need adaptation?

- Differences between training and testing conditions may significantly degrade recognition accuracy in speech recognition systems.
- Adaptation is an efficient way to reduce the mismatch between the models and the data from a particular speaker or channel.

Speaker adaptation

The adaptation of pre-existing models towards the optimal recognition of a new target speaker using limited adaptation data from the target speaker

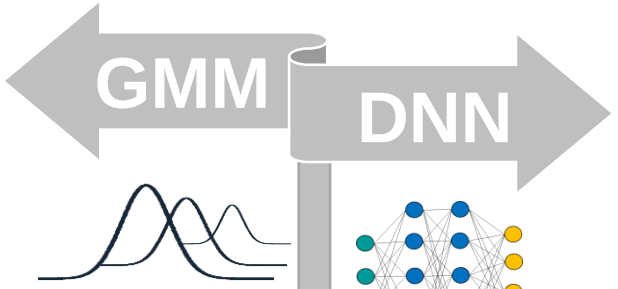


General speaker independent **(SI)** acoustic models trained on a large corpus of acoustic data from different speakers

Speaker adapted acoustic models, obtained from the **SI** model using data of a new speaker

Acoustic Models: GMM vs DNN

Gaussian Mixture Models



- GMM-HMMs have a long history: since 1980s have been used in speech recognition
- Speaker adaptation is a well-studied field of research

Deep Neural Networks

- Big advances in speech recognition over the past 3-5 years
- DNNs show higher performance than GMMs
- Neural networks are state-of-the-art of acoustic modelling
- Speaker adaptation is still a very challenging task

GMM adaptation

Model based: Adapt the parameters of the acoustic models to better match the observed data

- Maximum a posteriori (**MAP**) adaptation of GMM parameters

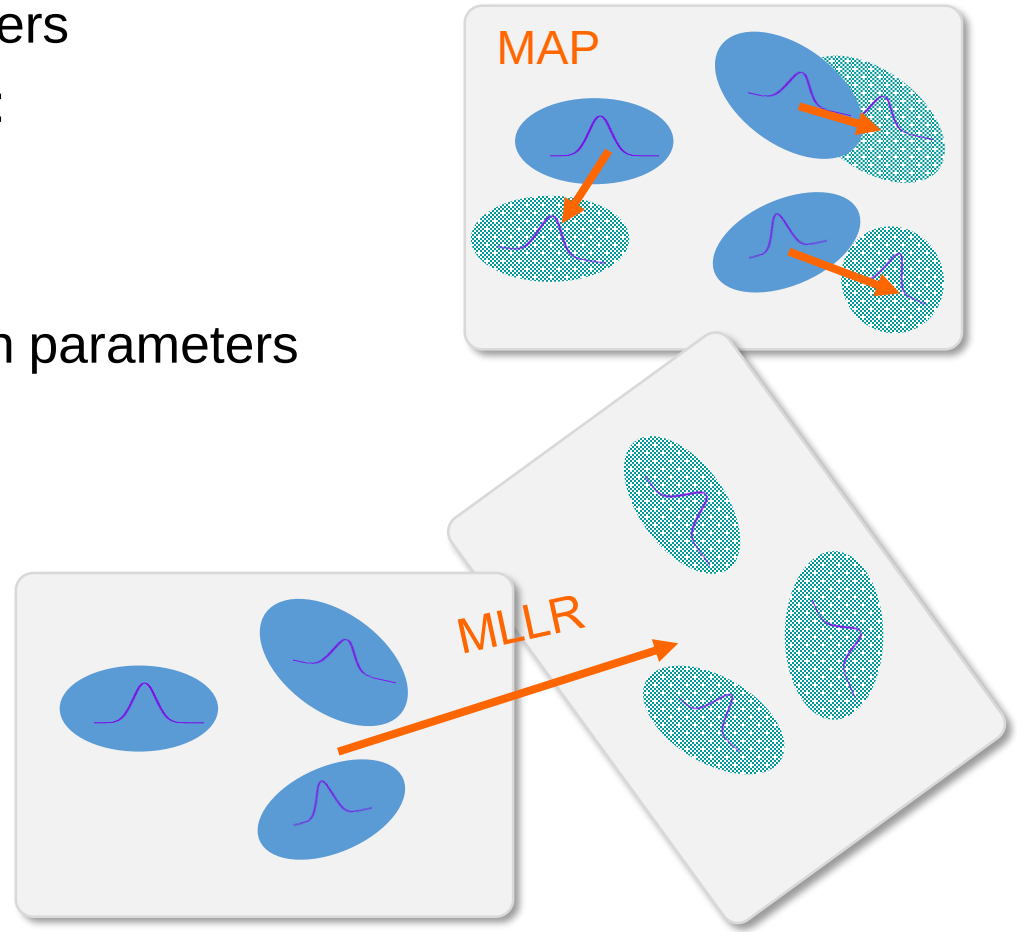
  In **MAP** adaptation each Gaussian is updated individually:

$$\hat{\mu}_m = \frac{\tau \mu_m + \sum_t \gamma_m(t) x_t}{\tau + \sum_t \gamma_m(t)}$$

- Maximum likelihood linear regression (**MLLR**) of Gaussian parameters

  In **MLLR** adaptation all Gaussians of the same regression class share the same transform:

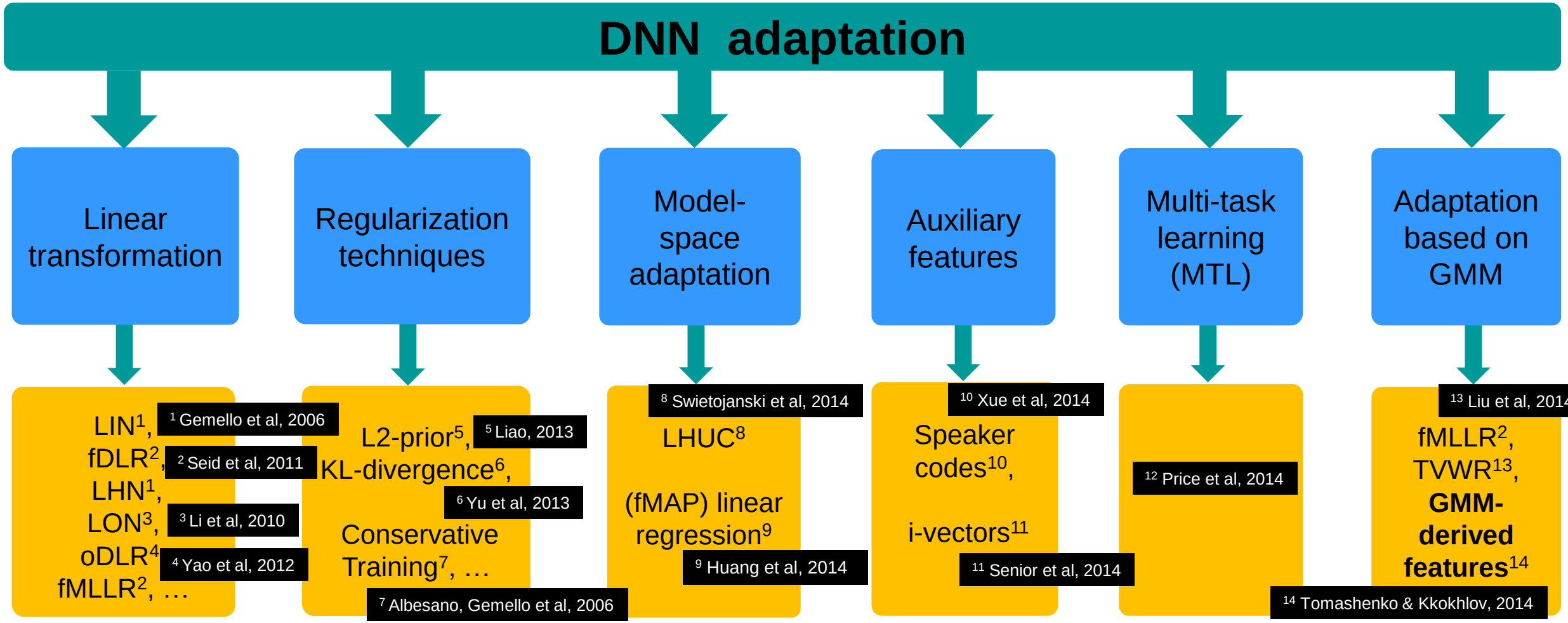
$$\hat{\mu} = A\mu + b$$



Feature space: Transform features

- Feature space maximum likelihood linear regression (**fMLLR**)

DNN adaptation: Related work



Combining GMM and DNN in speech recognition

- Tandem features[17] [17] Hermansky et al, 2000
- Bottleneck features[18] [18] Grézl et al, 2007
- GMM log-likelihoods as features for MLP[19] [19] Pinto & Hermansky, 2008
- Log-likelihoods combination
- ROVER*, lattice-based combination, CNC**, …

*ROVER – Recognizer Output Voting Error Reduction

**CNC – Confusion Network Combination

Outline

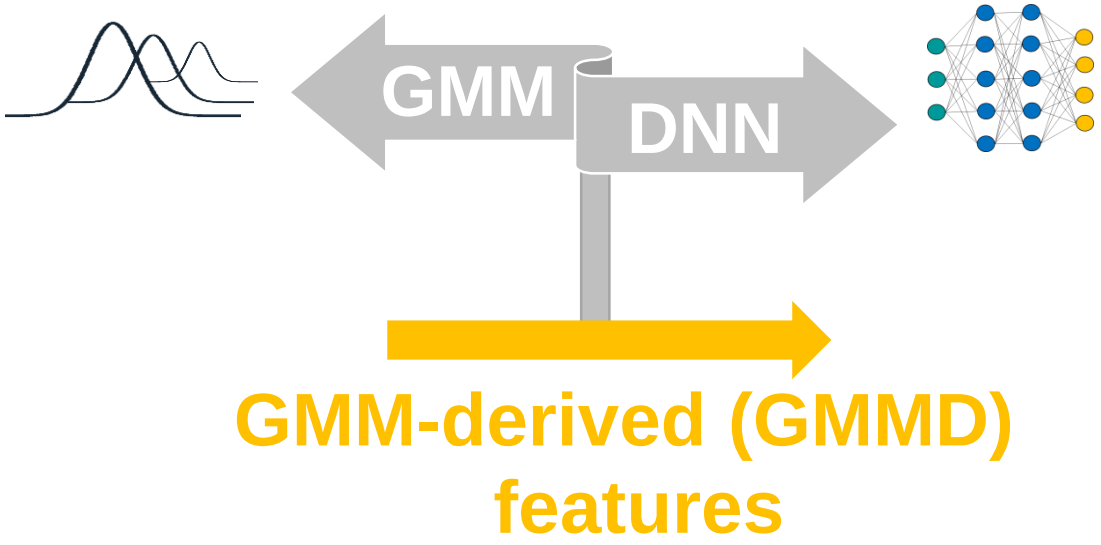
1. Introduction
   - Speaker adaptation
   - GMM vs DNN acoustic models
   - GMM adaptation
   - DNN adaptation: related work
   - Combining GMM and DNN in speech recognition
2. Proposed approach for speaker adaptation: GMM-derived features
3. System fusion
4. Experiments
5. Conclusions
6. Future work

Proposed approach: Motivation

- It has been shown that speaker adaptation is more effective for **GMM** acoustic models than for **DNN** acoustic models .
- Many adaptation algorithms that work well for **GMM** systems cannot be easily applied to **DNN**s.
- Neural networks and **GMM**s may be complementary and benefit from their combination.

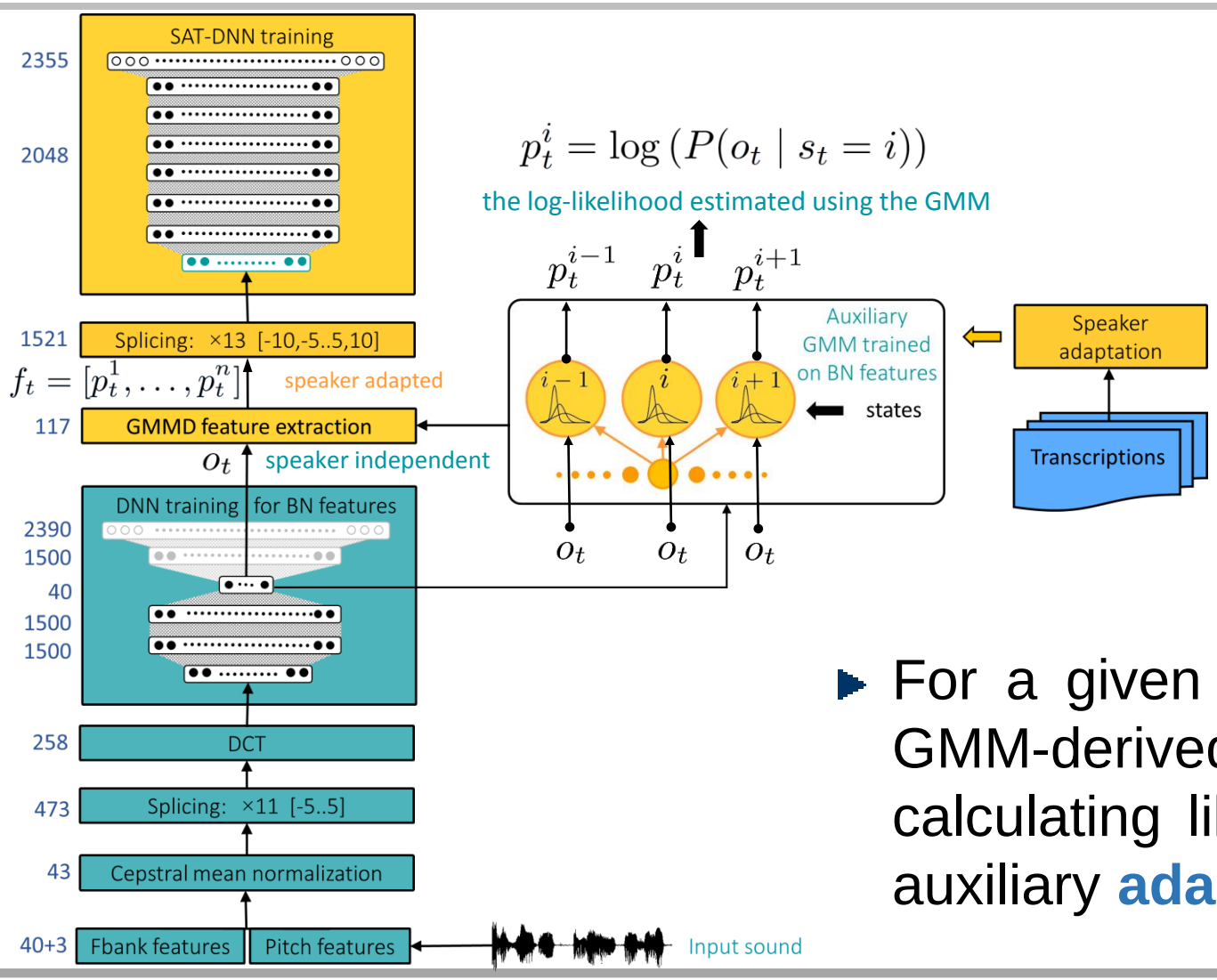
- To take advantage of existing adaptation methods developed for **GMM**s and apply them to **DNN**s.

Proposed approach: GMM-derived features for DNN



- Extract features using **GMM** models and feed these **GMM-derived** features to **DNN**. Train **DNN** model on **GMM-derived** features.
- Using **GMM** adaptation algorithms adapt **GMM-derived** features.

Bottleneck-based GMM-derived features for DNNs



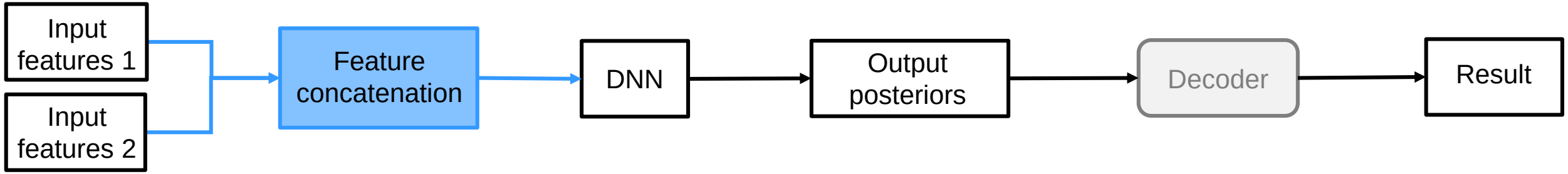
- For a given acoustic BN-feature vector O_t a new GMM-derived feature vector f_t is obtained by calculating likelihoods across all the states of the auxiliary **adapted GMM** on the given vector.

Outline

1. Introduction
   - Speaker adaptation
   - GMM vs DNN acoustic models
   - GMM adaptation
   - DNN adaptation: related work
   - Combining GMM and DNN in speech recognition
2. Proposed approach for speaker adaptation: GMM-derived features
3. System fusion
4. Experiments
5. Conclusions
6. Future work

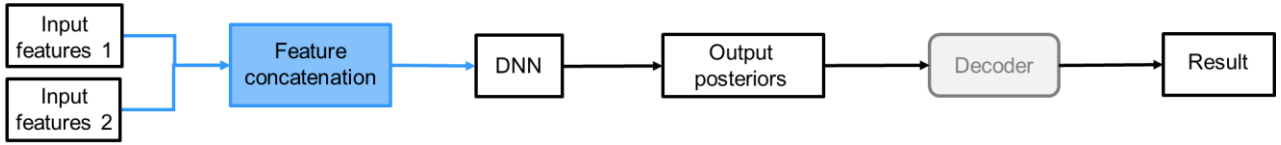
System Fusion

- **Feature** level: fusion for training and decoding stages

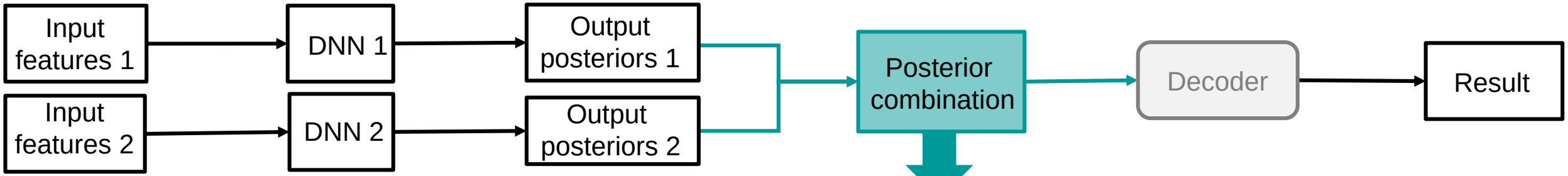


System Fusion

- **Feature** level: fusion for training and decoding stages



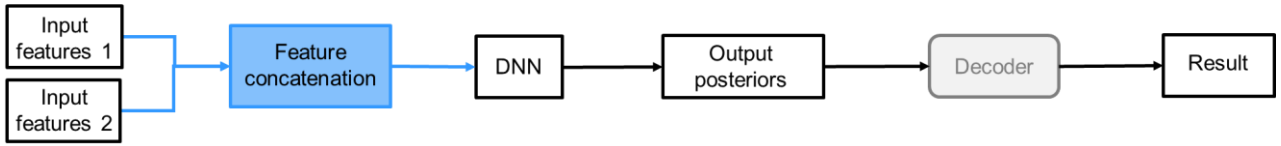
- **Posterior** combination



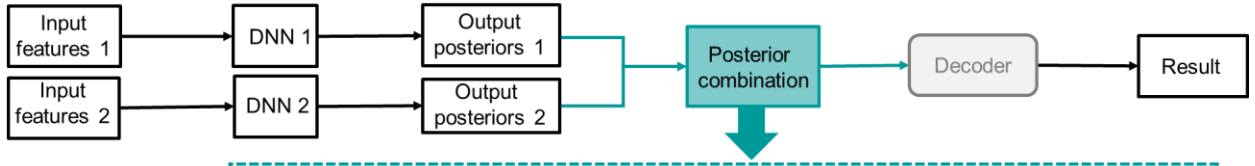
$$\log\left(p(o_t \mid s_i)\right) = \alpha \log\left(p_{DNN_1}(o_t \mid s_i)\right) + (1-\alpha)\log\left(p_{DNN_2}(o_t \mid s_i)\right)$$

System Fusion

- **Feature** level: fusion for training and decoding stages

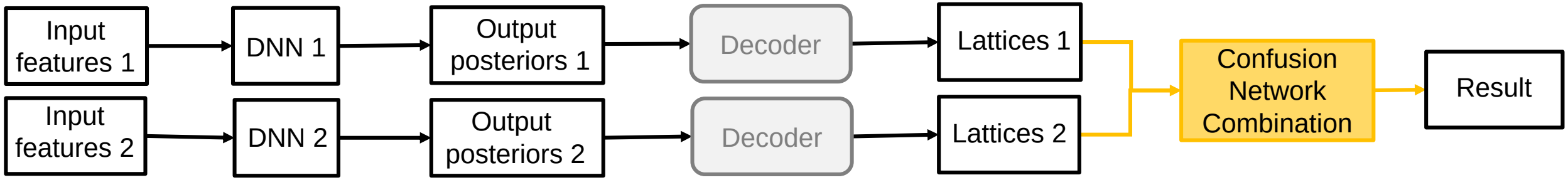


- **Posterior** combination



$$\log\left(p(o_t \mid s_i)\right) = \alpha \log\left(p_{DNN_1}(o_t \mid s_i)\right) + (1-\alpha)\log\left(p_{DNN_2}(o_t \mid s_i)\right)$$

- **Lattice** combination



Outline

1. Introduction
   - Speaker adaptation
   - GMM vs DNN acoustic models
   - GMM adaptation
   - DNN adaptation: related work
   - Combining GMM and DNN in speech recognition
2. Proposed approach for speaker adaptation: GMM-derived features
3. System fusion
4. Experiments
5. Conclusions
6. Future work

Experiments: Data

TED-LIUM corpus:* 1495 TED talks, 207 hours: 141 hours of male, 66 hours of female speech data, 1242 speakers, 16kHz

Data set	Duration, hours	Number of Speakers	Mean duration per speaker, minutes
Training	172	1029	10
Development	3.5	14	15
$Test_1$	3.5	14	15
$Test_2$	4.9	14	21

LM:** 150K word vocabulary and publicly available trigram LM

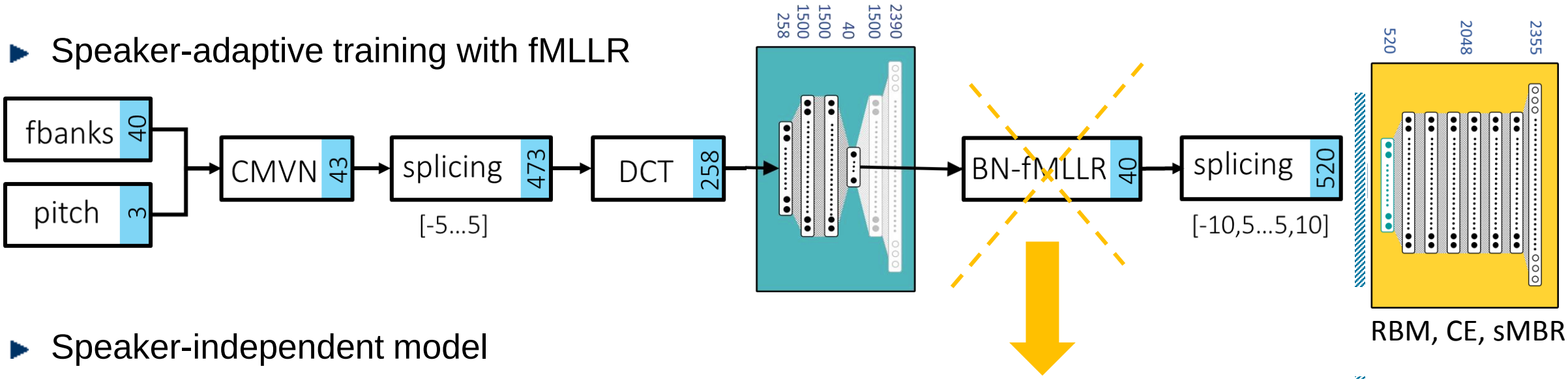
*A. Rousseau, P. Deleglise, and Y. Esteve, “Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks“ 2014

**cantab-TEDLIUMpruned.lm31

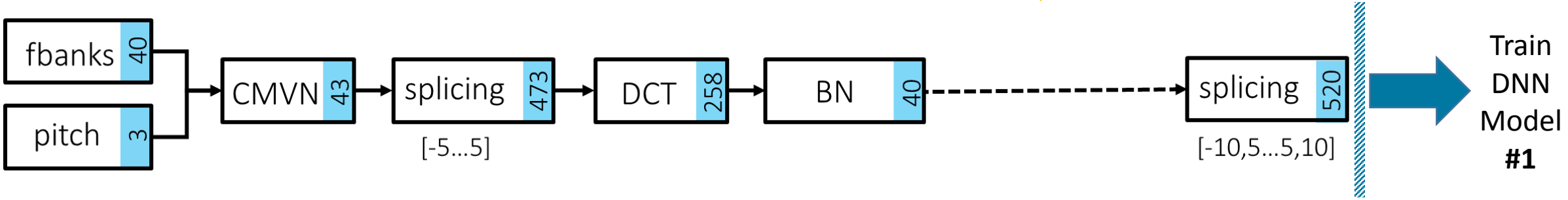
Experiments: Baseline systems

We follow Kaldi TED-LIUM recipe for training baselines models:

- Speaker-adaptive training with fMLLR



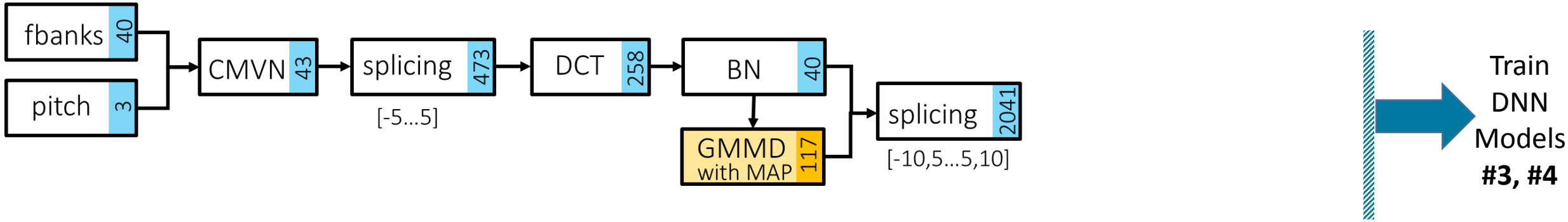
- Speaker-independent model



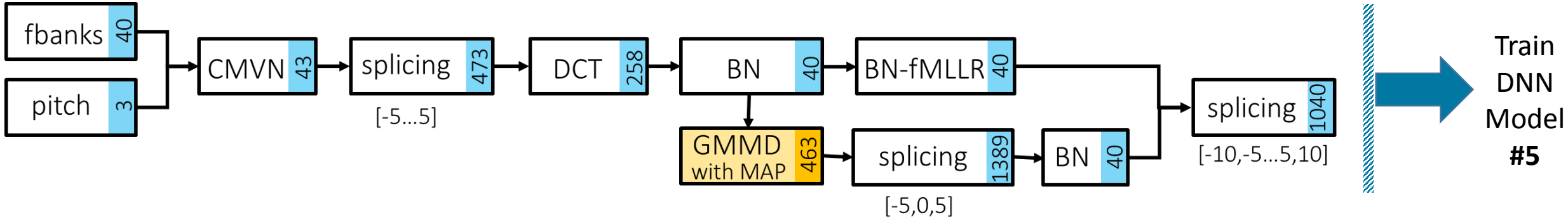
Experiments: Training models with GMMD features

2 types of integration of GMMD features into the baseline recipe:

- 1. Adapted features AF_1 (with monophone auxiliary GMM)



- 2. Adapted features AF_2 (with triphone auxiliary GMM)



Results: Adaptation performance for DNNs

GMMD baseline

#	Adaptation	Features	τ	WER, %		
				Dev	$Test_1$	$Test_2$
1	No	BN		12.14	10.77	13.75
2	fMLLR	BN		10.64	9.52	12.78
3	MAP	AF_1	2	10.27	9.59	12.94
4	MAP	AF_1 + align. #2	5	10.26	9.40	12.52
5	MAP+fMLLR	AF_2 + align. #2	5	10.42	9.74	13.29

better than speaker-adapted baseline

τ parameter in MAP adaptation

Results: Adaptation and Fusion

α is a weight of the baseline model in the fusion

	#	Adaptation	Features	α	WER, % Dev		WER, % $Test_1$		WER, % $Test_2$	
GMMD baseline	1	No	BN		12.14*		10.77*		13.75*	
	2	fMLLR	BN		10.57		9.46		12.67	
	4	MAP	AF_1 + align. #2		10.23		9.31		10.46	
	5	MAP+fMLLR	AF_2 + align. #2		10.37		9.69		13.23	
fusion	6	Posterior fusion: #2 + #4		0.45	9.91	↓ 6.2	9.06	↓ 4.3	12.04	↓ 5.0
	7	Posterior fusion: #2 + #5		0.55	9.91	↓ 6.2	9.10	↓ 3.8	12.23	↓ 3.5
	8	Lattice fusion: #2 + #4		0.44	10.06	↓ 4.8	9.09	↓ 4.0	12.12	↓ 4.4
	9	Lattice fusion: #2 + #5		0.50	10.01	↓ 5.3	9.17	↓ 3.1	12.25	↓ 3.3

* WER in #1 was calculated from lattices, in other lines – from consensus hypothesis

↓ Relative WER reduction in comparison with adapted baseline #2

Best improvement

- Two types of fusion: **posterior** level and **lattice** level provide additional comparable improvement,
- In most cases posterior level fusion provides slightly better results than the lattice level fusion.

Outline

1. Introduction
   - Speaker adaptation
   - GMM vs DNN acoustic models
   - GMM adaptation
   - DNN adaptation: related work
   - Combining GMM and DNN in speech recognition
2. Proposed approach for speaker adaptation: GMM-derived features
3. System fusion
4. Experiments
5. Conclusions
6. Future work

Conclusions

- We investigate a new way of **combining GMM and DNN** frameworks for speaker adaptation of acoustic models

- The main advantage of GMM-derived features is the possibility of performing the adaptation of a DNN-HMM model through the adaptation of the auxiliary GMM. **Other methods for the adaptation** of the auxiliary GMM can be used instead of MAP or fMLLR adaptation. Thus, this approach provides a **general framework for transferring adaptation algorithms** developed for GMMs to DNN adaptation

- Experiments demonstrate that in an unsupervised adaptation mode, the proposed adaptation and fusion techniques can provide, approximately,
  - **11–18%** relative Δ WER (in comparison with speaker independent model)
  - **3–6%** relative ΔWER (in comparison with strong fMLLR adapted baseline)

Outline

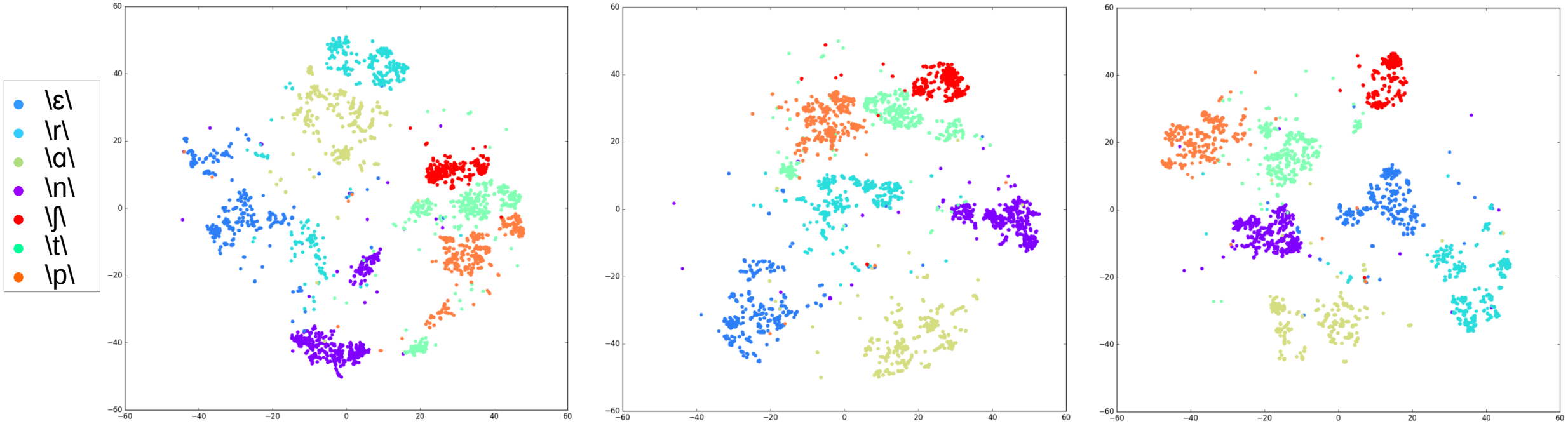
1. Introduction
   - Speaker adaptation
   - GMM vs DNN acoustic models
   - GMM adaptation
   - DNN adaptation: related work
   - Combining GMM and DNN in speech recognition
2. Proposed approach for speaker adaptation: GMM-derived features
3. System fusion
4. Experiments
5. Conclusions
6. Future work

Future work

- Investigate the performance of the proposed method for different types of Neural Networks (Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM),….)
- Other tasks…
- Better understanding and analysis of GMMD features – how we can improve the performance?

Visualization of output vectors using t-SNE*

Visualization of the softmax output vectors of the DNNs (5 speakers, 7 phonems):



1. Baseline speaker-independent DNN, trained on BN features

2. Baseline speaker-adapted DNN, trained on fMLLR adapted BN features

3. DNN, trained using GMMD features with MAP adaptation

* t-Distributed Stochastic Neighbor Embedding: Maaten, L. V. D., & Hinton, G. Visualizing data using t-SNE. 2008.

Key References (1)

Adaptation of DNN acoustic models:

1. R. Gemello, F. Mana, S. Scanzio, P. Laface, & R. De Mori, Adaptation of hybrid ANN/HMM models using linear hidden transformations and conservative Training. 2006.
2. F. Seide, G. Li, X. Chen, & D. Yu, Feature engineering in context-dependent deep neural networks for conversational speech transcription. 2011.
3. B. Li & K. C. Sim, Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems. 2010.
4. K. Yao, D. Yu, F. Seide, H. Su, L. Deng, & Y. Gong, Adaptation of context-dependent deep neural networks for automatic speech recognition. 2012.
5. H. Liao, Speaker adaptation of context dependent deep neural networks. 2013.
6. D. Yu, K. Yao, H. Su, G. Li, & F. Seide, KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. 2013.
7. D. Albesano, R. Gemello, P. Laface, F. Mana, & S. Scanzio, Adaptation of artificial neural networks avoiding catastrophic forgetting. 2006.
8. P. Swietojanski & S. Renals, Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models. 2014.
9. Z. Huang, J. Li, S. M. Siniscalchi, I.-F. Chen, C. Weng, & C.-H. Lee, Feature space maximum a posteriori linear regression for adaptation of deep neural. Networks. 2014.
10. S. Xue, O. Abdel-Hamid, H. Jiang, L. Dai, & Q. Liu, Fast adaptation of deep neural network based on discriminant codes for speech recognition. 2014.
11. A. Senior & I. Lopez-Moreno, Improving DNN speaker independence with i-vector inputs. 2014.
12. Price, R., Iso, K. I., & Shinoda, K. Speaker adaptation of deep neural networks using a hierarchy of output layers. 2014.
13. S. Liu & K. C. Sim, On combining DNN and GMM with unsupervised speaker adaptation for robust automatic speech recognition. 2014.

Key References (2)

Proposed approach for adaptation:

14. N. Tomashenko & Y. Khokhlov. Speaker adaptation of context dependent deep neural networks based on map-adaptation and GMM-derived feature processing. 2014.
15. N. Tomashenko & Y. Khokhlov. GMM-derived features for effective unsupervised adaptation of deep neural network acoustic models. 2015.
16. Kundu, S., Sim, K. C., & Gales, M. Incorporating a Generative Front-End Layer to Deep Neural Network for Noise Robust Automatic Speech Recognition. 2016.

Combining GMM and DNN:

17. Hermansky, H., Ellis, D. P., & Sharma, S. Tandem connectionist feature extraction for conventional HMM systems. 2000.
18. Grézl, F., Karafiát, M., Kontár, S., & Cernocky, J. Probabilistic and bottle-neck features for LVCSR of meetings. 2007.
19. J. P. Pinto & H. Hermansky, Combining evidence from a generative and a discriminative model in phoneme recognition. 2008.

ITMO UNIVERSITY

Speech
Technology
Center

http://www-lium.univ-lemans.fr

http://en.ifmo.ru

http://speechpro.com

Thank you!

Questions?