Speaker Adaptive Joint Training of Gaussian Mixture Models and Bottleneck Features

Zoltán Tüske, Pavel Golik, Ralf Schlüter, Hermann Ney

Human Language Technology and Pattern Recognition, Computer Science Department, RWTH Aachen University, Aachen, Germany

RWTH AACHEN UNIVERSITY

Motivation & Goals

Motivation:

- Speaker adaptation is a crucial part of acoustic modeling
- With GMMs: Constrained Maximum Likelihood Linear Regression (CMLLR) [Gales 98]
  - Robust feature space adaptation technique
  - Works with neural network (NN) based tandem bottleneck (BN) features
  - Often combined with speaker adaptive training (SAT)
- Consistent tandem approach has been introduced recently
  - Jointly trained tandem features and GMM [Tüske 15]
  - Speaker independent

Goals:

- Extension of our previous work [Tüske 15]
- Full integration of CMLLR-SAT GMM into the NN framework
- Joint (end-to-end) training of: GMM, NN-feature, and feature space adaptation
  - Specific parts of a single model

State-of-the-art – Overview

Gaussian Mixture Models (GMM)

- Trained on cepstral features (e.g. MFCC, Gammatone) and
- Neural network (NN) based features (tandem approach): e.g. bottleneck (BN)

Hybrid approach

- Estimates state posteriors directly
- Better or similar performance to GMM with deep tandem BN features

Tandem vs. Hybrid?

- NN integration of Gaussian Mixtures via log-linear mixtures [Tüske 15]
- Jointly optimized mixture and NN-features: tandem should NOT perform worse than classical hybrid

Integration of GMM into DNN [Tüske 15]

- GMM is equivalent to a log-linear mixture model (LMM) [Saul 01]
- Assuming globally shared covariance matrix (Σ) results in
  - 1st- and 0th-oder "feature functions"

$$p_\theta(s|x) = \sum_i p_\theta(s,i|x) = \frac{p(s)\sum_i p(i|s)\mathcal{N}(y|\mu_{si},\Sigma)}{\sum_{s'} p(s')\sum_i p(i|s')\mathcal{N}(y|\mu_{s'i},\Sigma)} = \frac{\sum_i \exp(w_{si}^T y + b_{si})}{\sum_{s',i} \exp(w_{s'i}^T y + b_{s'i})}$$

- (Generalized) softmax layer with hidden variables
  - **NN interpretation:** huge softmax layer followed by sum/max pooling
- Conversion of (generative) GMM to (discriminative) LMM:

$$b_{si} = -\frac{1}{2}\mu_{si}^T\Sigma\mu_{si} + \ln\ p(s) + \ln\ p(i|s) \qquad w_{si} = \Sigma^{-1}\mu_{si}$$

- Where:

x	observation, input feature
$y = f(x)$	non-linearly transformed input feature x, e.g. output of a BN layer
$p_\theta(s\|x)$	posterior probability of state s given observation x
i	hidden variable/mixture component
$\mathcal{N}(y\|\cdot,\cdot)$	Normal distribution
μ_{si}, Σ	GMM means and (pooled) covariance
$p(i\|s)$	mixture component weight
$p(s)$	state prior
$\theta = \{w_{si}, b_{si}\}$	log-linear model parameters

- "Smoothing" the posterior distribution of the converted model is important
  - see details in the paper

References

[Saul 01] L. Saul and D. Lee, "Multiplicative updates for classification by mixture models," in NIPS, vol. 2, 2001, pp. 897-904.

[Gales 98] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Computer Speech and Language, vol. 12, pp. 75-98, 1998.

[Tüske 15] Tüske et al., "Integrating Gaussian mixtures into deep neural networks: softmax layer with hidden variables," in ICASSP, 2015, pp. 4285-4289.

Feature space adaptation of Gaussian mixture models

- Assuming single Gaussian target model (no hidden variables)
- Optimal feature space transformation maximizes the likelihood

$$\mathcal{N}(y|\mu_{s,r},\Sigma_{s,r}) = |A_r|\mathcal{N}(A_r y + a_r|\mu_s,\Sigma)$$

- Where:

r	speaker cluster
μ_s, Σ	non-adapted acoustic model parameters
$\mu_{s,r}, \Sigma_{s,r}$	speaker dependent Gaussian parameters of state s
A_r, a_r	affine feature-space transformation for the speaker r

- Transformation can already be robustly estimated on a few minutes of speech

Speaker dependent layer and joint SAT of BN-GMM

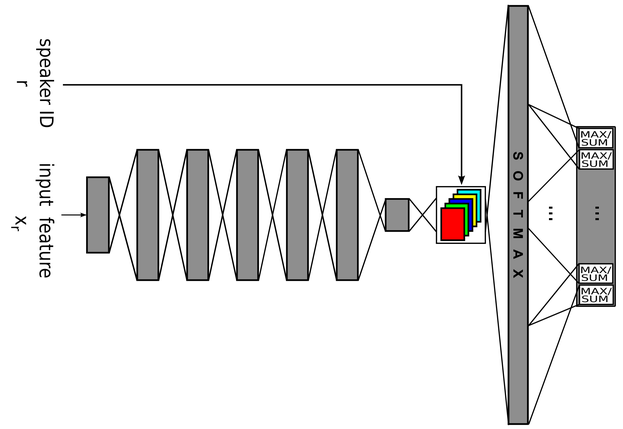
- NN integration of speaker dependent feature transformation
- Denoting:
  $z^{(l)}(x)$ the output of the lth layer
  - Forward:

$$z^{(l)}(x_r) = A_r z^{(l-1)}(x_r) + a_r$$

  - Backpropagation (BP):

$$\frac{\partial E}{\partial z^{(l-1)}} = \sum_j \frac{\partial E}{\partial z_j^{(l)}}\frac{\partial z_j^{(l)}}{\partial z^{(l-1)}} = A_r^{(l)T}\frac{\partial E}{\partial z^{(l)}}$$

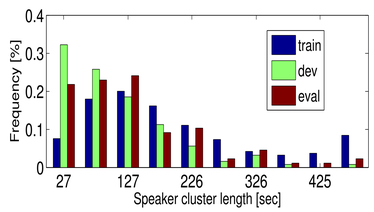
- Joint SAT of BN-GMM:



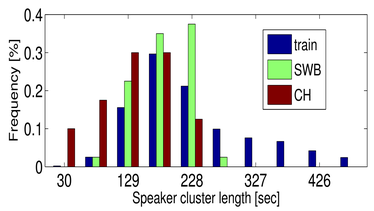
- Joint SAT results in a second set of BN features for the second recognition pass
  - Can be extracted in advance, no speaker dependent parameters
- Backpropagation through speaker dependent layer without updating it
  - NN is made **"CMLLR aware"** → robust adaptation to a new speaker

Experimental setups

- Two different genres:
  - QUAERO: broadcast news and conversations, 50 hours of training data
  - Switchboard: telephony speech, 309 hours of training data
- Models are trained on a fixed Viterbi alignment
- Neural networks with rectified linear units, up to 12 hidden layers
- Input features: 50-dimensional Gammatone instead of MFCC
  - CMLLR transform of input features did not help (see comparisons in the paper)
- Per speaker 2-3 minutes speech are available on average:
  - QUAERO:



  - Switchboard:



Experimental results

Broadcast news - QUAERO 50h

- **Speaker independent** (SI) results obtained with optimal mixture and BN size:
  - Best hybrid: no BN
  - Best tandem: maximum likelihood (ML) BN-GMM after 6 splits, and BN size 64
  - CE joint BN-GMM: after split-5, and BN size 256

AM	Joint training	Training criterion	WER [%] dev	WER [%] eval
Hybrid		CE	**15.4**	20.6
BN-GMM		ML	16.5	21.9
	×	CE	15.7	**20.5**

  - Larger BN is beneficial ***after*** CE based joint BN-GMM training [Tüske 15]
  - Hybrid *vs* BN-GMM: about the same performance
- **Joint speaker adaptive BN-GMM training:**
  - Best ML results: after split-9, BN size 64

BN size	split	BN-GMM/LMM Joint training	BN-GMM/LMM Training criterion	WER [%] dev	WER [%] eval
256	0	×	CE	16.0	20.9
128	5	×	CE	15.5	20.4
		×	CE	15.1	**19.5**
64		×	CE	**14.9**	19.7
	9		ML	15.8	20.7

  - Mixture output performs better than classical softmax (split-0)
  - For joint SAT of BN-GMM, 64-dimensional BN showed the best results:
    20.5% → **19.5%** WER on evaluation set

Telephone conversation - Switchboard 300h

- Best hybrid: BN size 512
- For BN-GMM the BN size was set to 128 (not optimized)
- ML BN-GMM results obtained by split-8 model
- CE BN-GMM results obtained by split-4 model

AM	Joint training	CMLLR/ SAT	Training criterion	WER [%] SWB	WER [%] CH	WER [%] Total
Hybrid			CE	13.7	26.9	20.3
BN-GMM			ML	14.2	27.8	21.0
		×	ML	13.3	26.4	19.9
	×		CE	13.6	27.0	20.3
		×	CE	13.7	26.0	19.9
	×	×	CE	**12.9**	**24.9**	**18.9**

- SI joint BN-GMM as good as hybrid
- SAT joint BN-GMM is better than SI hybrid
- SAT joint BN-GMM (**18.9%**) outperforms SAT hybrid (**19.4%**)
- Joint training (i.e. backpropagation through CMLLR layer) is important
- Improvement related to joint SAT BN-GMM:
  20.3% → **18.9%** WER on Hub5e'00
- Effect of MPE, re-alignment, and external LM data
  - Additional text resources: BN/BC (QUAERO), Gigaword, TED

AM	MPE	MPE re-align.	ext. LM	WER [%] Hub5e'00 SWB	Hub5e'00 CH	Hub5e'01 SWB	Hub5e'01 SWB2-p3	Hub5e'01 SWB-Cell	RT03s SWB	RT03s FSH
Hybrid	×			12.6	24.9	13.4	17.4	23.8	26.4	16.7
joint SAT BN-GMM	×			11.5	23.0	12.5	15.8	21.6	24.0	14.8
	×	×		11.2	21.7	11.9	15.5	20.3	23.0	14.2
	×	×	×	**10.8**	**21.0**	**11.6**	**15.1**	**19.9**	**22.7**	**13.8**

Conclusions

- Joint SAT BN-GMM significantly outperformed our classical hybrid:
  7–11% relative WER improvement
- NN trained "CMLLR aware" resulted in robust adaptation to unseen speakers
- Future work: fine-tuning of the speaker dependent layer

Acknowledgement: The study was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract number W911NF-12-C-0012.

Mail: tuske@cs.rwth-aachen.de