Structuring Media Data Using Latent Modelling

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Mortaza Doulaty
Oscar Saz, Raymond Ng,
Thomas Hain

University of
Sheffield

29 June 2016

Outline

Natural Speech Technology

- Introduction and Motivation
- (Acoustic) Latent Dirichlet Allocation
- Discovering Latent Domains
- Automatic Identification of Shows and Genres in Media Data
- Acoustic Model Adaptation to Latent Domains
- Conclusion

Introduction and Motivation

Natural Speech Technology

- Media data is usually very diverse
  - e.g. a news show may start with clean, in studio read speech and continue with report outside studio
  - a show exhibits various different characteristics
- Complex characteristics and attributes of media data may not be easily observable or measurable
- Unsupervised latent modelling approaches may better suit the diverse nature of media data

Introduction and Motivation

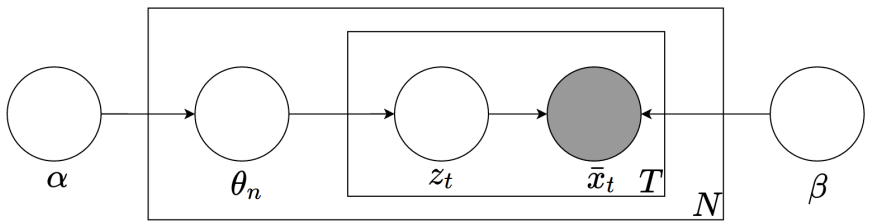
Natural Speech Technology

- Domain adaptation to diverse data such as media data can be challenging
- Conventional adaptation approaches such as MLLR and MAP might not be very suitable for this complex type of data
- New adaptation techniques that use information about the complex structure of media data can further improve ASR performance

Latent Dirichlet Allocation

Natural Speech Technology

- Unsupervised probabilistic generative model for collection of discrete data
  - aims to describe how every item within a collection is generated
  - assuming there are a set of latent variables
  - each item is modelled as a finite mixture over latent variables
- Initially proposed for latent topic modelling of text corpora

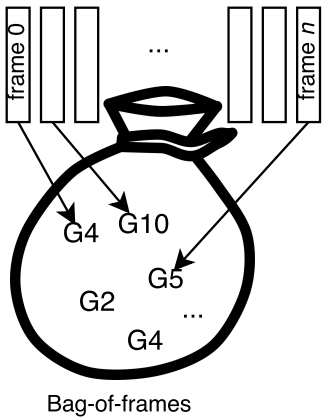


$$p(\theta, \mathbf{z}, \bar{\mathbf{d}}|\alpha, \beta) = p(\theta|\alpha) \prod_{t=1}^{T} p(z_t|\theta) p(x_t|z_t, \beta)$$

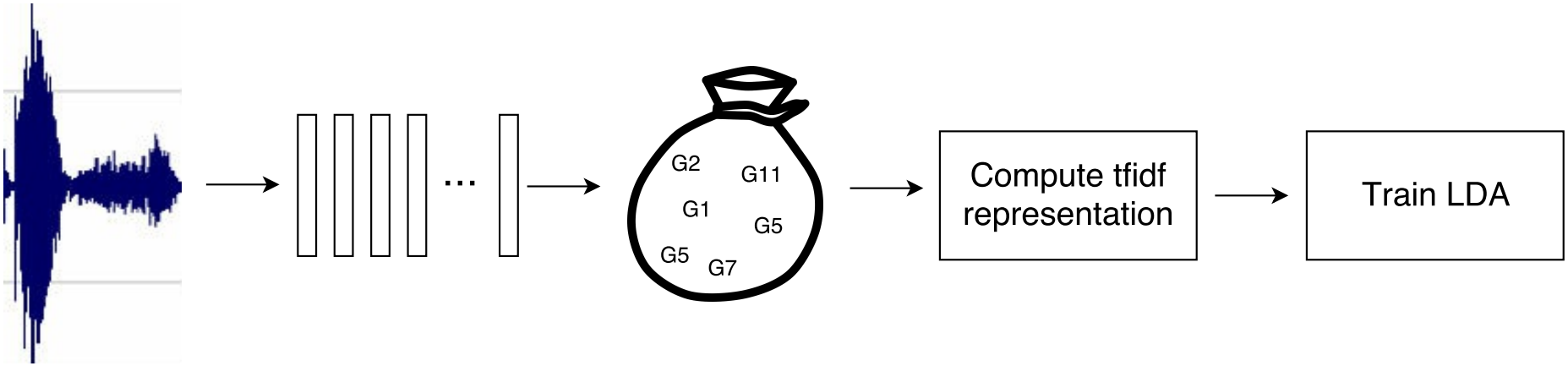
Acoustic LDA

- Latent factors in speech signal are interpreted as “domains”
- To fit into discrete LDA concept, speech signals need to be converted into discrete symbols

- Bag-of-frames representation of audio:

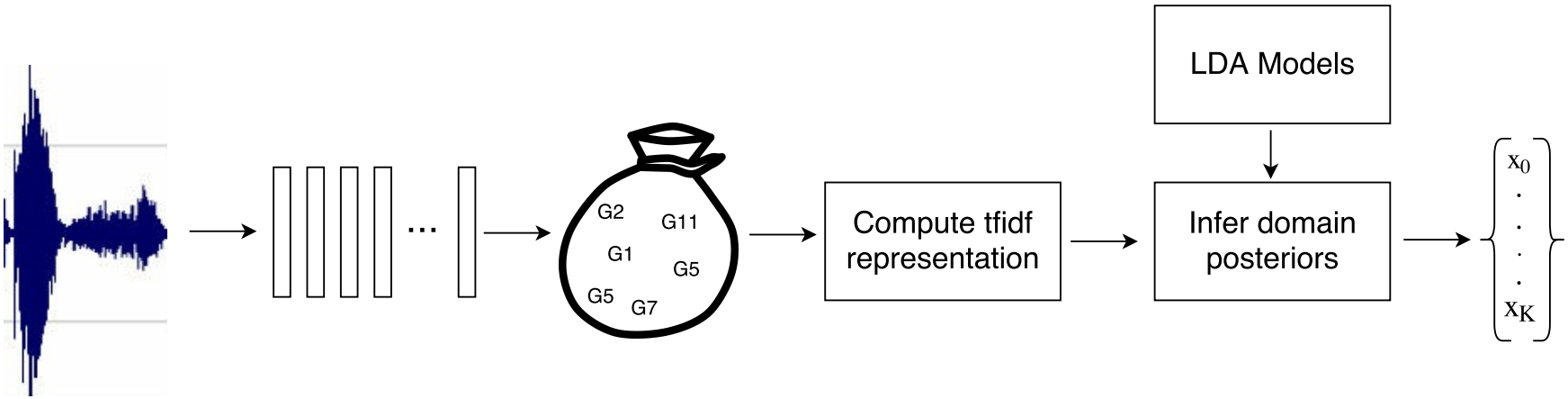


Acoustic LDA Training



- With discrete symbols, LDA models can be trained
- For training LDA models, tf-idf representation of frame symbols are used
- Number of latent domains needs to be decided before training

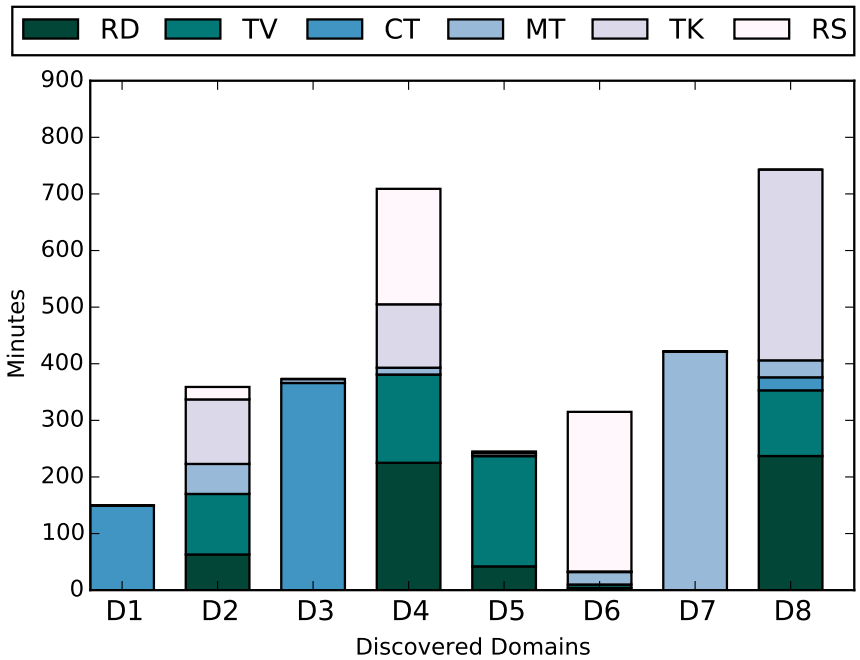
Acoustic LDA Inference



- LDA models can be used to infer the latent domain posteriors given the acoustic data
- Domain posteriors contain information about the latent structure of the data
- They can be used other downstream processes:
  - discovering latent domains and adapting to them
  - show entities and genre identification

Discovering Latent Domains

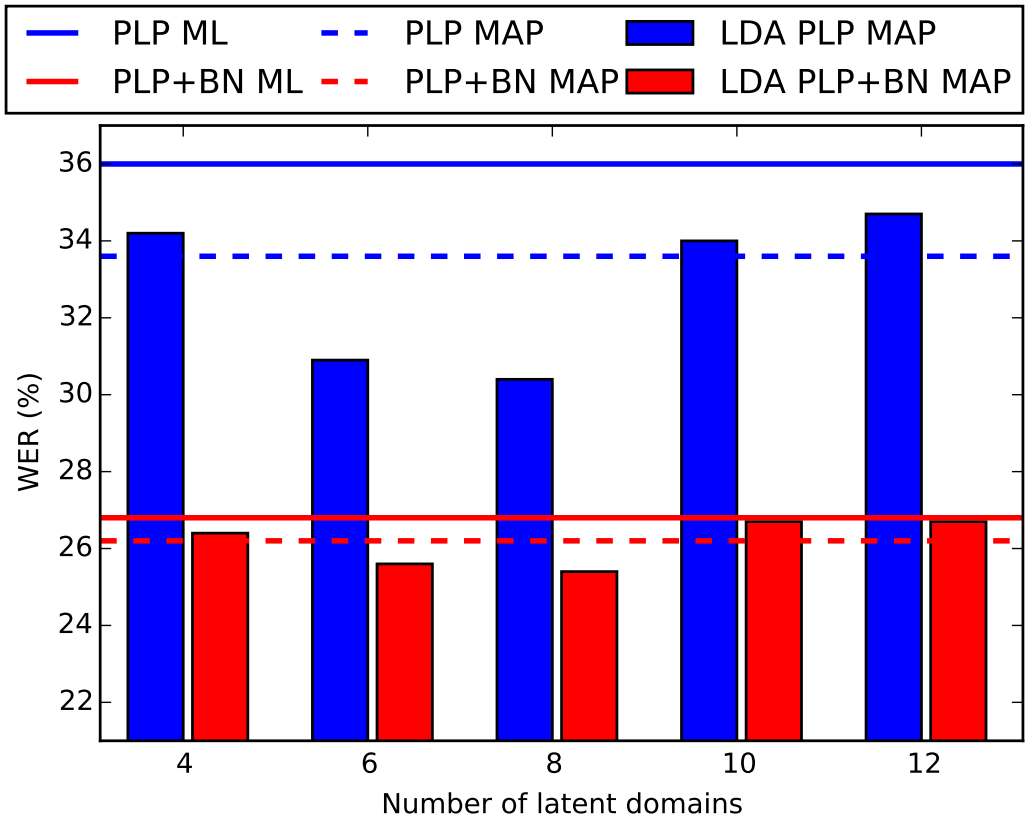
- Discovering latent domains using LDA in a very diverse dataset and adapting to them
- Diverse data set consisting radio (RD), television (TV), telephone speech (CT), meeting (MT), lectures (TK) and read speech (RS)



M. Doulaty, O. Saz, and T. Hain, "Unsupervised Domain Discovery using Latent Dirichlet Allocation for Acoustic Modelling in Speech Recognition", Proc. Interspeech, Germany, 2015.

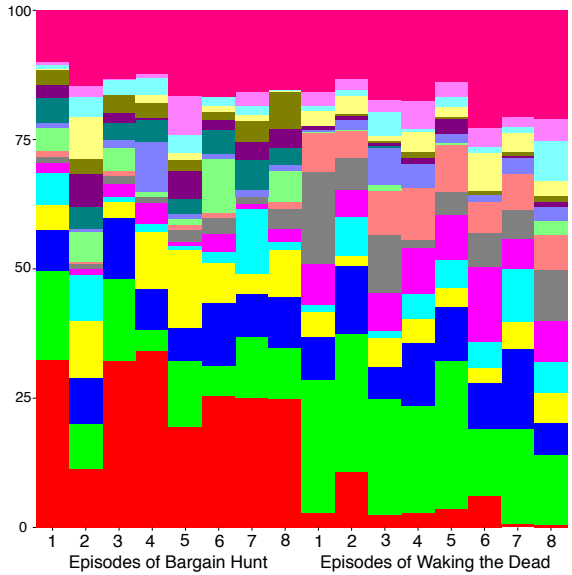
MAP Adaptation to Latent Domains

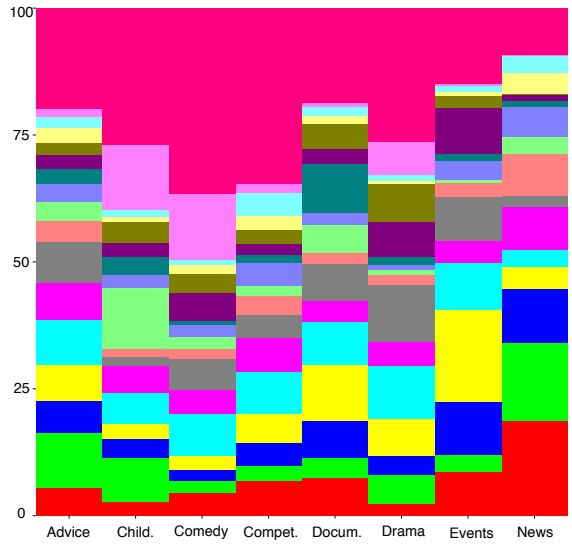
- MAP adaptation to latent domains improves the WER over the adaptation to the named domains



Distribution of Latent Domains

- BBC TV broadcasts used to train LDA models
- Distribution of LDA domain posteriors plotted for few episodes of the two different shows
- Similarities within show and differences across shows are visible in posterior distributions
- The distribution is different across genres as well





Genre and Show Identification

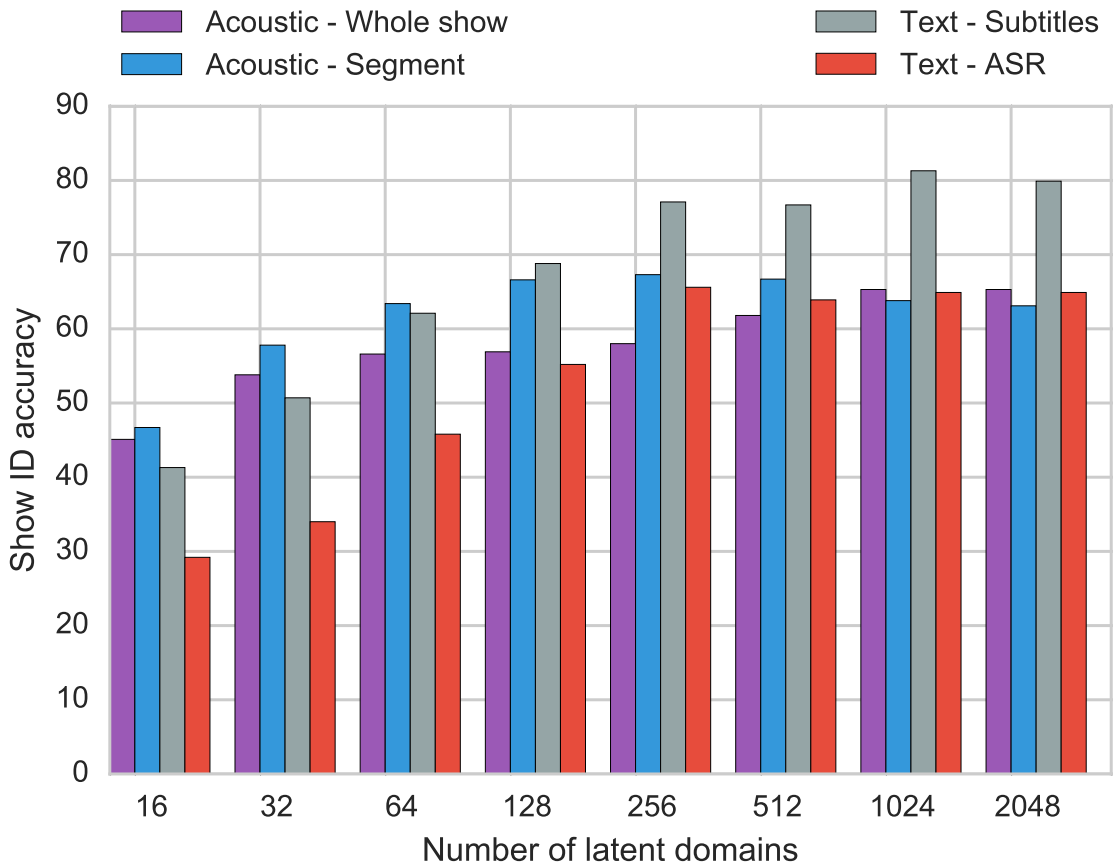
Natural Speech Technology

- LDA domain posteriors has unique distribution across different shows and genres
- These posterior vectors can be used to classify shows and genre entities
- 1,000h of BBC TV broadcasts, consisted of over 1,500 shows (133 unique shows and 8 genres) used for training and 200h, consisted of 288 shows used for testing
- Baseline models were GMMs and classification accuracy was 61.5% and 70.1% for genre ID and show ID
- LDA posteriors used a features and classified using SVMs

M. Doulaty, O. Saz, R. W. M. Ng, and T. Hain, Automatic Genre and Show Identification of Broadcast Media, Proc. Interspeech, USA, 2016.

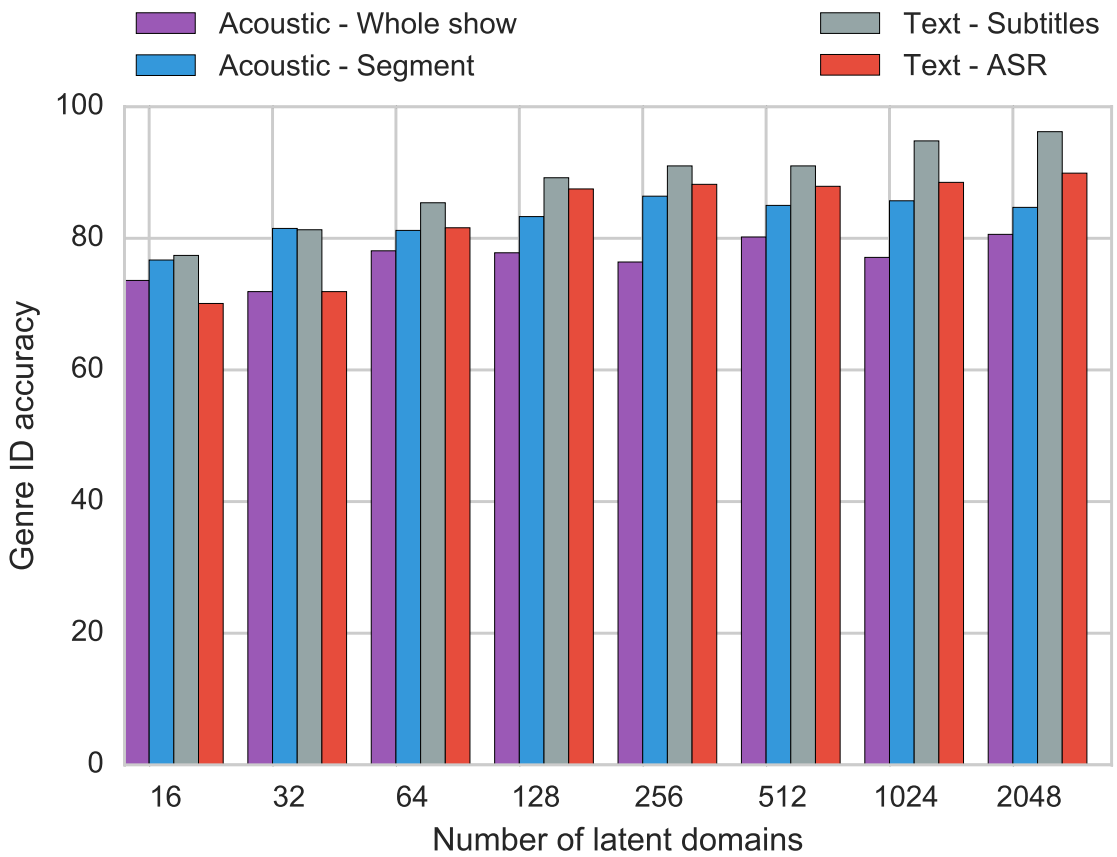
Show Identification Experiments

- Show ID task with 133 classes using acoustic and text LDA is studied
- More latent domains leads to better classification accuracy



Genre Identification Experiments

- Genre ID task with 8 classes using acoustic and text LDA is studied
- Similar to the previous task, more latent domains leads to better classification accuracy



Using Meta-data

Natural Speech Technology

- Meta-data can be used to improve classification accuracy, meta-data includes date and time of broadcast
- Using meta-data improves the classification accuracy for both genre ID and show ID tasks

Meta-Data	Genre ID	Show ID
Acoustic LDA (256)	86.4	67.3
Only Channel & Time	46.7	22.0
Acoustic + Channel	89.6	72.8
Acoustic + Time	89.9	77.7
Acoustic + Channel & Time	92.3	82.6

System Combination

Natural Speech Technology

- What is the best performance we can get for these tasks?
- Acoustic and text based LDA plus meta-data are combined together

Method	Genre ID	Show ID
Baseline (acoustic 256)	86.4	67.3
Baseline (text 2048)	96.2	79.9
Acoustic & Text	97.2	85.0
Acoustic + Meta-data & Text	98.6	85.7

DNN Adaptation to Latent Domains

Natural Speech Technology

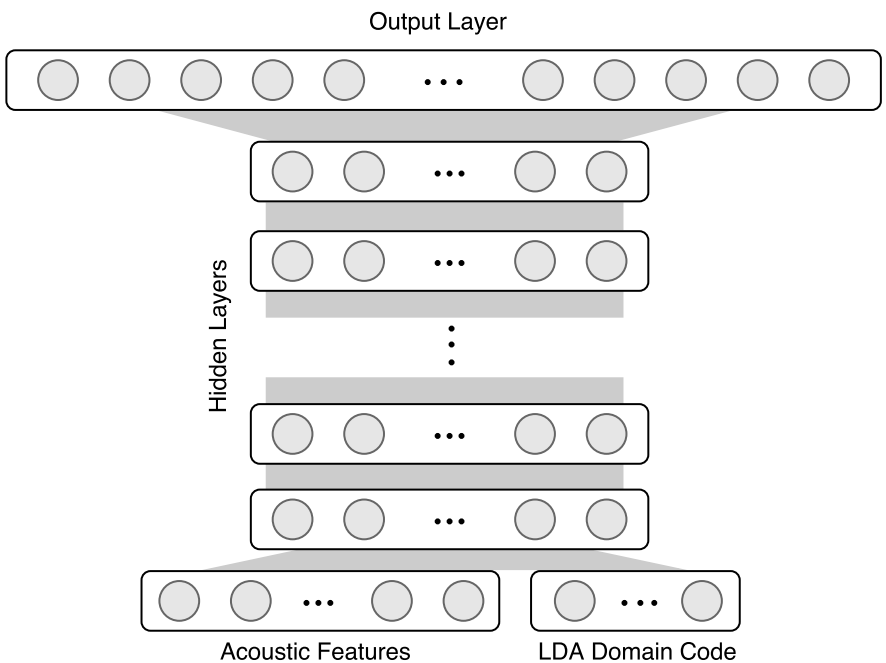
- More latent domains helped in the previous task
- Shortcomings of previous approach for MAP adaptation can be addressed by using a different adaptation technique
- Acoustic models can be adapted to the latent domains using domain codes, similar to iVector type adaptation of DNNs
- Domain information derived from the LDA model with K domains is encoded with a K-dimensional one-hot vector

M. Doulaty, O. Saz, R. W. M. Ng, and T. Hain, Latent Dirichlet Allocation Based Organisation of Broadcast Media Archives for Deep Neural Network Adaptation, Proc. of ASRU, USA, 2015.

Input Augmentation with Domain Code

- Augmenting DNN input with LDA domain code
- Equivalent to learning a new bias for each latent domain

$$\mathbf{v}^1_{LDaT} = f\left(\left[\mathbf{W}^1_{\mathbf{v}} \mathbf{W}^1_{\mathbf{d}}\right] \begin{bmatrix} \mathbf{v}^0 \\ \mathbf{d} \end{bmatrix} + \mathbf{b}^1_{LDaT} \right)$$
$$= f\left(\mathbf{W}^1_v \mathbf{v}^0 + \underbrace{\mathbf{W}^1_d \mathbf{d} + \mathbf{b}^1_{LDaT}}_{\text{domain specific bias}} \right)$$



Dataset Definition

Natural Speech Technology

- BBC TV broadcasts used for the experiments
- 560h for training and 28h for testing
- Dataset is similar to the MGB 2015 challenge's training and dev set

	Train		Development	
Genre	Shows	Time	Shows	Time
Advice	264	193.1h	4	3.0h
Children's	415	168.6h	8	3.0h
Comedy	148	74.0h	6	3.2h
Competition	270	186.3h	6	3.3h
Documentary	285	214.2h	9	6.8h
Drama	145	107.9h	4	2.7h
Events	179	282.0h	5	4.3h
News	487	354.4h	5	2.0h

Baseline Models

Natural Speech Technology

- Baseline acoustic models were hybrid feedforward DNNs with 6 hidden layers and an output layer of size 6k
- Subtitles from shows with a total of 650 million words were used to train 4-gram language models
- Speaker adapted DNNs with SAT style training were also used, where speaker CMLLR transformations are applied to the input features

Model		WER (%)
GMM	SAT BMMI	41.0
DNN	Baseline	33.3
	Speaker Adapted	31.4

Acoustic LDA Domain Adaptation

Natural Speech Technology

- Segment based acoustic LDA models were trained and used to infer the latent domain posteriors
- LDA vectors of size 64 was used
- Combining speaker adaptation and latent domain adaptation further improves the WER

	WER (%)								
	Adv.	Chld.	Cmdy.	Compt.	Docum.	Drama	Even.	News	**Overall**
–	27.6	29.1	47.8	28.2	31.3	52.0	38.1	17.9	**33.3**
SAT	26.2	27.5	46.1	25.9	29.8	49.3	35.8	15.9	**31.4**
LDA	25.8	27.8	45.1	25.7	28.9	47.7	33.5	15.7	**30.6**
LDA+SAT	24.2	26.5	43.8	23.6	27.3	45.0	31.6	14.3	**28.9**

Conclusion

Natural Speech Technology

- Acoustic Latent Dirichlet Allocation can be used to learn the latent structure of diverse and media data
- Latent domain posteriors also used for genre and show identification of broadcast media
- Latent domain posteriors used for adaptation of GMM/HMM and DNN/HMM systems
- 8% relative WER reduction was achieved in the MGB dataset (comapred with the speaker adapted DNNs)
- Domain adaptation and speaker adaptation are complementary

References

- M. Doulaty, O. Saz, and T. Hain, Unsupervised Domain Discovery using Latent Dirichlet Allocation for Acoustic Modelling in Speech Recognition, Proc. Interspeech, Germany, 2015.
- M. Doulaty, O. Saz, R. W. M. Ng, and T. Hain, Latent Dirichlet Allocation Based Organisation of Broadcast Media Archives for Deep Neural Network Adaptation, Proc. ASRU, USA, 2015.
- M. Doulaty, O. Saz, R. W. M. Ng, and T. Hain, Automatic Genre and Show Identification of Broadcast Media, Proc. Interspeech, USA, 2016.

Thanks!