DNN Approach to Speaker Diarisation Using Speaker Channels

Rosanna Milner and Thomas Hain

Machine Intelligence for Natural Interfaces
Speech and Hearing
University of Sheffield

The University Of Sheffield.

SPandH

Outline

- Introduction
- Background
- DNN approach using speaker channels
  - Fixed or mixed number of channels
- Experiments
  - Test Data
  - Setup
  - Evaluation
  - Results
- Conclusion

Introduction

Speaker diarisation - 'who speaks when'

- the 3 main tasks are SAD, speaker segmentation and speaker clustering
- **step-by-step**: performs stages separately
- **integrated**: performs some stages together

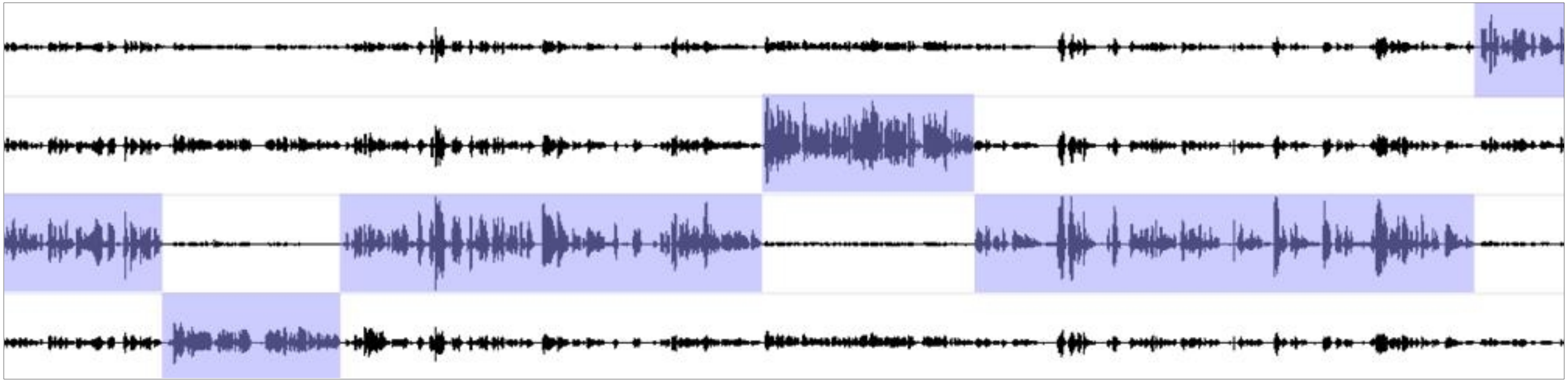
Typically unsupervised

- **unsupervised**: no prior knowledge or information
- **lightly/semi-supervised**: auxiliary information or metadata available
- **supervised**: prior knowledge about test data known

Presenting a semi-supervised integrated method using DNNs trained on concatenated IHM features

- <u>semi-supervised</u>: uses IHM speaker channels (instead of SDM)
- <u>integrated</u>: performs all three tasks together

Multi-channel Diarisation Approaches



- Single channel (SDM)
  - Segmentation, change detection and clustering
- How close are the microphones to the speaker ?
  - Associated speaker channels (IHM++)
  - Distant speaker channels (MDM)

Scoring Diarisation Output

Diarisation Error Rate (DER)

- Frame based metric
- Collar changes reference
- No penalty for data fragmentation
- Alternative scoring (Milner & Hain, ICASSP'16)

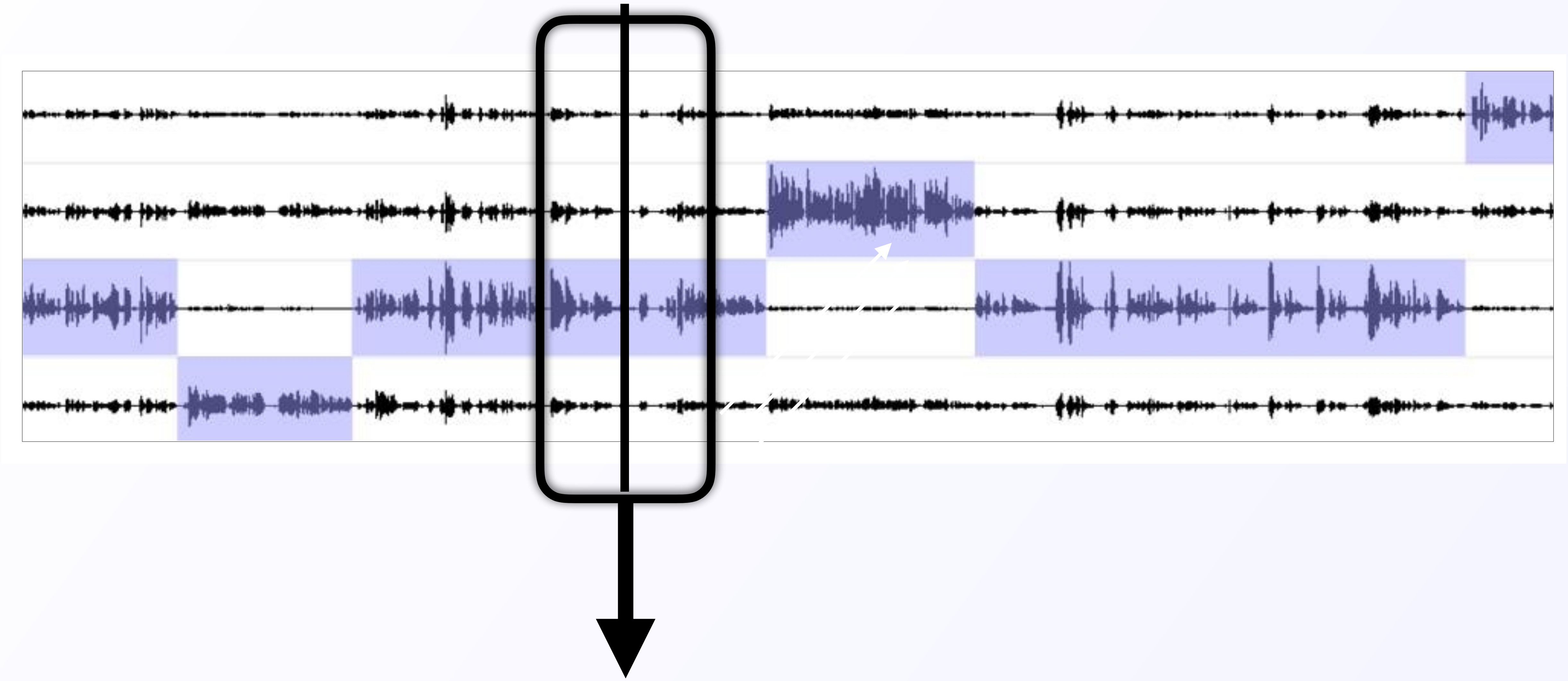
Scoring target

- on individual channels - IHM
  - very low number due to non-speech prior !
- on one 'global channel' - SDM
- based on true activity - MDM

Related Work

- **Multichannel diarisation**
  - Beamforming focuses on speakers (Anguera et al. 2007)
  - Detecting closest speech and disregarding other speech (Dines et al. 2006, Wrigley et al. 2005)
- **DNNs for diarisation**
  - Feature transforms using ANNs (Yella et al. 2014, 2015)
  - DNNs trained for SAD
    (Dines et al., 2006, Milner & Hain, 2015)
  - Windowing segmentation method and clustering using autoassociateive NNs (Jothilakshmi et al. 2009)
  - Clustering by adapting speaker separation DNNs to specific recordings (Milner & Hain, 2016)

Approach in this work



Channel 2 → Rosanna

Approach: Using speaker assigned channels

Fixed number of chanels

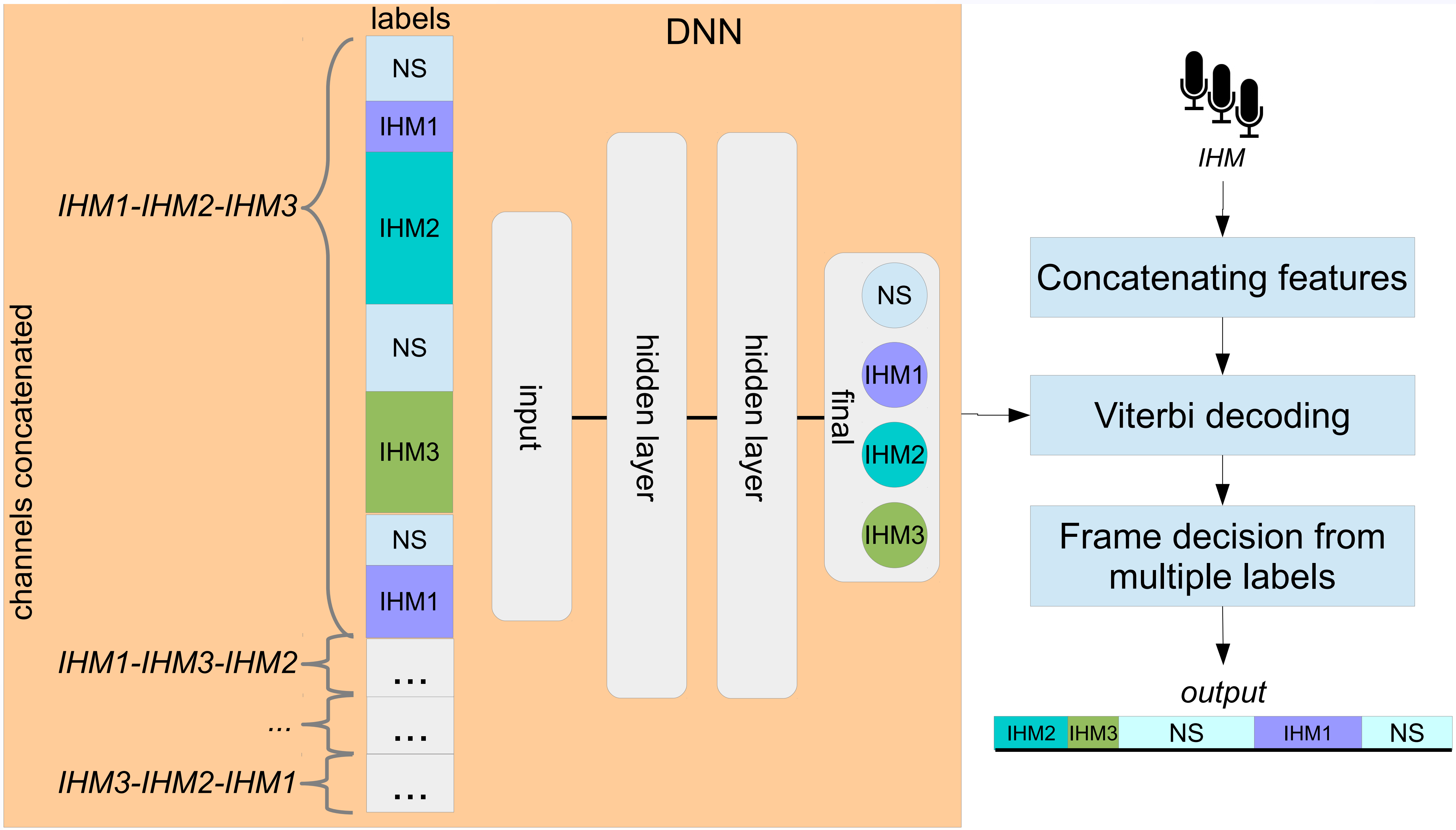
- every recording must contain the same number of speaker channels, x concatenate x channels in all permutations: x! features per recording
- final layer in DNN is x+1, representing x speakers and NONSPEECH

Mixed number of channels

- recordings contain different number of speaker channels
- concatenate all pairs of channels: $x(x - 1)$ features per recording
- final layer in DNN is 3, representing 2 speakers and NONSPEECH
- a speaker in label source but not in channel pair has NONSPEECH label

Both methods require every speaker having their own channel

Approach



labels
DNN
IHM1-IHM2-IHM3
NS
IHM1
IHM2
NS
IHM3
NS
IHM1
IHM1-IHM3-IHM2
...
IHM3-IHM2-IHM1
channels concatenated
input
hidden layer
hidden layer
final
NS
IHM1
IHM2
IHM3
IHM
Concatenating features
Viterbi decoding
Frame decision from multiple labels
output
IHM2
IHM3
NS
IHM1
NS

Frame Decisions

Combinatorial Voting

<table>
<tr><th>FRAME</th><th>IHM1-IHM2-IHM3</th><th>IHM1-IHM3-IHM2</th><th>…</th><th>IHM3-IHM2-IHM1</th><th></th><th>OUTPUT</th></tr>
<tr><td>…</td><td>...</td><td>…</td><td></td><td>…</td><td></td><td>...</td></tr>
<tr><td>204</td><td rowspan="3">IHM1</td><td rowspan="2">IHM1</td><td rowspan="7">…</td><td rowspan="4">IHM1</td><td rowspan="7">=</td><td>IHM1</td></tr>
<tr><td>205</td><td>IHM1</td></tr>
<tr><td>206</td><td rowspan="3">NS</td><td>IHM1</td></tr>
<tr><td>207</td><td rowspan="2">NS</td><td>NS</td></tr>
<tr><td>208</td><td>NS</td><td>NS</td></tr>
<tr><td>209</td><td>IHM3</td><td>IHM3</td><td>IHM2</td><td>IHM3</td></tr>
<tr><td>210</td><td>IHM2</td><td>IHM2</td><td>IHM3</td><td>IHM2</td></tr>
<tr><td>…</td><td>…</td><td>…</td><td></td><td>…</td><td></td><td>...</td></tr>
</table>

- All combinations of feature concatenations used for testing
- results in multiple labels for every frame
- simply count occurrences and choose label which occurs most often
- additionally: apply a prior for NONSPEECH

Data - Meetings

- NIST RT'07 - meeting data
  - NIST reference and manually transcribed reference (0.1 sec precision)
  - 11144 segments, 35 speakers
  - 8 meetings
  - 6 meetings: 4 speakers, 1 meeting: 5 speakers, 1 meeting: 6 speakers

Improved manual reference on
http://mini.dcs.shef.ac.uk/resources/dia-improvedrt07reference/

Data: Talk Show Radio 4

- The Bottom Line - BBC Radio4
- Topics in Business and Economy
- 3 participants
- 1 interviewer (Evan Davis)

- manually transcribed reference
- 8749 segments, 40 speakers
- 12 train and 10 test programmes

Evaluation

Diarisation error rate

- DER=MS+FA+SE
- does not consider the segmentation quality so all tables show the number of detected segments

Two scoring settings

- NIST
  - collar 0.25s
  - score specified times only (UEM)
  - NIST provided reference (where possible)
- SHEF
  - collar 0.05s
  - score complete recordings
  - manually transcribed references

Baseline results

- LIUM SpkrDiarization (Rouvier et al., 2013)
  - tailored for TV and radio broadcasts
  - BIC segmentation with CLR and integer linear programming and i-vector clustering

Channel	#Segs	#Spkrs	NIST DER%	SHEF DER%
Data: TBL				
SDM	2030	82	16.6	27.8
IHM	8478	40	393.9	335.9
Data: RT07				
SDM	2648	72	40.1	66.4
IHM	13070	35	308.1	371.0

Crosstalk on channels which results in high false alarm

Channel	#Segs	#Spkrs	NIST DER%	SHEF DER%
ICSI - SDM	3082	54	21.7	66.2

Features and configuration

- Features
  - Log filterbank (23 coefs, 32 frames, compressed)
  - Cross talk features (Wrigley et al, 2006) - normalised energy, kurtosis, mean/max cross correlation and differentials, 7 per channel

- DNN configurations
  - 2 hidden layers (1000 hidden units)
  - With cross talk features (31 frames)
  - trained with or without overlapping speech (OV) - unqiue labels - TBL only - overlap 7.5%

Fixed Channel Experiments - TBL

- Only possible on TBL data
  - with or without overlap in training
  - with or without cross talk features

DNN Train	OV	CT	#Segs	MS%	FA%	SE%	SHEF DER%
TBL	x		6732	4.3	2.4	1.2	8.0
TBL	x	x	7136	4.3	2.4	1.7	8.4
TBL			7269	4.3	2.5	1.5	8.3
TBL		x	2964	4.6	3.7	1.4	9.7

DNN TBL+OV gives lowest SHEF DER, crosstalk features do not help

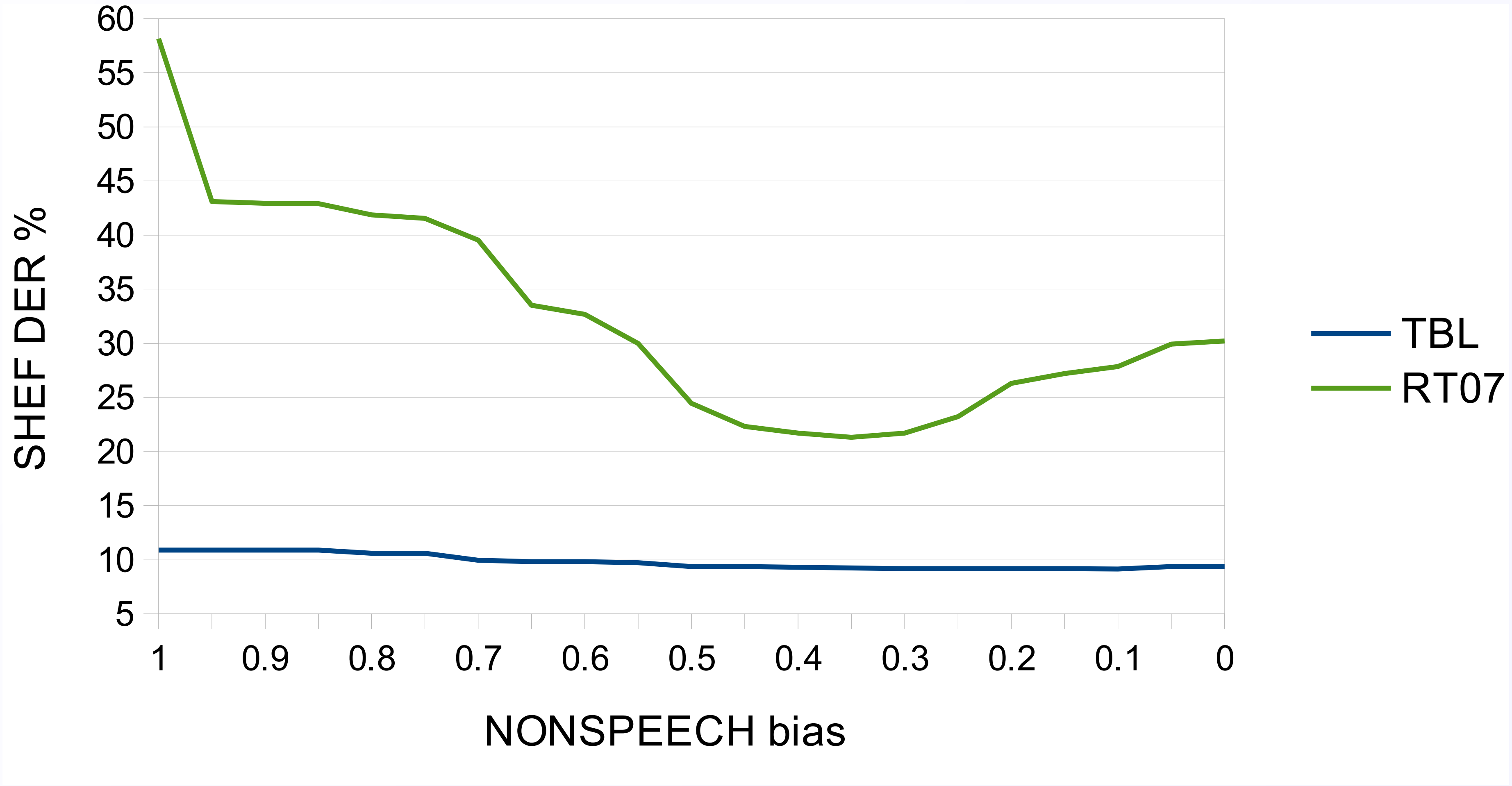
Weight	#Segs	MS%	FA%	SE%	SHEF DER%
0.75	6594	4.3	2.6	1.3	8.2
0.5	6571	4.2	2.7	1.3	8.2
0.25	6569	4.2	2.8	1.4	8.3

Mixed Channel Numbers - 1

Data	DNN Train	DNN OV	DNN CT	#Segs	MS%	FA%	SE%	SHEF DER%
TBL	TBL	x		8295	20.3	1.1	0.9	22.4
	TBL	x	x	10551	34.8	0.7	1.1	36.5
	TBL			8263	17.0	1.4	1.0	19.4
	TBL		x	7932	7.7	0.9	1.2	10.9
	AMI			10354	16.6	1.0	4.9	22.5
	AMI		x	7683	22.9	0.9	5.0	28.8
RT07	TBL	x		7979	60.9	0.8	0.4	62.1
	TBL	x	x	4169	79.6	0.4	0.1	80.1
	TBL			8430	56.5	1.2	0.4	58.2
	TBL		x	5993	59.7	1.3	0.2	61.2
	AMI			8791	58.9	0.5	0.1	59.5
	AMI		x	6873	62.4	0.5	0.1	63.0

- crosstalk features only improve for DNN for TBL
- including overlap in DNN training gives worse performance
- DNNs trained on AMI data do not perform as well as DNNs trained on TBL data without overlap

Mixed Channel Numbers - II



- applying a weight helps both datasets
- RT07 benefits the most with a large performance increase from 58.2%% to 23% SHEF DER

Best results - Mixed channels

Data	SHEF DER%	NIST DER%
TBL	9.2	5.7
RT'07	23.2	15.1

Conclusions

- Presented two approaches for speaker diarisation using only IHM channels
- Evaluated on two datasets: RT07 (meeting) and TBL (broadcast media)
- Compared two scoring settings NIST and SHEF
- Applying nonspeech bias reduces error in mixed method
- Training on OV benefits fixed method but not mixed
- CT only benefit DNN trained on TBL and tested on TBL
- Best result between best reported results on SDM and MDM with single stage processing.

The End

Thank you.