THE 2015 SHEFFIELD SYSTEM FOR LONGITUDINAL DIARISATION OF BROADCAST MEDIA

R. Milner, O. Saz, S. Deena, M. Doulaty, R.W.M. Ng, T. Hain

Speech and Hearing Group, Department of Computer Science, University of Sheffield, UK

{rmmilner2,o.saztorralba,s.deena,mortaza.doulaty,wm.ng,t.hain}@sheffield.ac.uk

MGB Challenge - Task 4

The longitudinal diarisation task for the MGB challenge aimed to perform diarisation across multiple TV series by linking clusters of speech segments, representing speakers, in one programme to the clusters in other programmes.

Data

This challenge provided access to the BBC archives. Data used consisted of various TV series from different genres:

- Documentary, political drama, sci-fi drama, sitcom, sports event

Dataset	Series	Shows	Spkrs	Time (h)
Train	-	2,193	-	1580.5
Development	5	19	348	9.2
Test	2	19	486	10.4

Speech Activity Detection

DNN-based speech activity detection (SAD)

- Two speech/non-speech detectors trained:
  - Input layer of 368, 2 hidden layers of 1000 nodes, output 2 nodes
  - SNS1 - phonemes as speech (759 hours speech, 793 hours non-speech)
  - SNS2 - 60 second chunks of segments which have less than 40% PMER and WMER, and AWD is 0.3-0.7 seconds (116 hours speech, 363 hours non-speech)
  - Tuned number of HMM states for model duration, prior probability for speech model and grammar scale factor
  - Padded with 0.25 seconds

DNN	Tuning			Error		
	States	Prior	Scale	MS	FA	DER
SNS1	1	0.2	6	23.5	0.4	23.9
+Padding	1	0.2	1	8.3	1.2	9.5
SNS2	30	0.05	30	11.0	7.5	18.5
+Padding	1	0.05	12	4.1	8.5	12.6

Resegmentation (refining segments)

- ASR system used to obtain sentence hypothesis for each segment
- Raw word confidence, segment confidence score, length of word (seconds and phonemes), length of segment (seconds) used as input features to a decision tree
- Words above a certain threshold were used to define new segments, words below deemed non-speech and resegmented

SAD adaptation with LM decoding

- Adapting SNS2 DNN with output from ASR resegmentation process on a per show basis (including non-speech)
- Decoding modified to include speech occurrence patterns
  - 2 state output replaced with 6 states based on speech non-speech and duration
  - States represent duration ranges (<4s, 4s-7s, >10s)
  - Bigram language model of speech duration states

Stage	MS	FA	DER
SNS2+Padding	4.1	8.5	12.6
+Resegmentation	6.7	2.7	9.4
+AdaptationLM	4.4	3.8	8.2

Diarisation

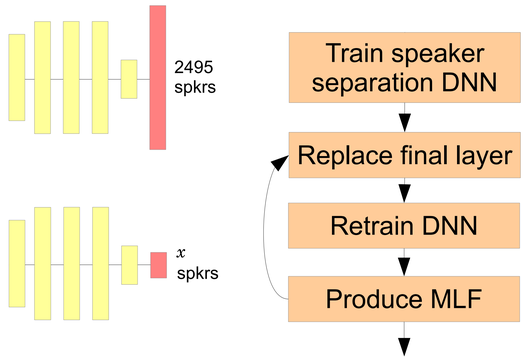
Speaker clustering

- SHoUT toolkit: designed for meetings, BIC segmentation and stopping criterion in an unsupervised model training regime

Speaker clustering adaptation using DNNs

- Trained a speaker separation DNN:
  - Input layer of 368, 3 hidden layers of 1000 modes, 26 node bottleneck, 2495 speakers for final layer
  - No speaker information: aligned segments with subtitle colour information and automatically clustered to give new speaker labels, compared with original subtitle colour information
  - Kept pure speakers with at least 40 seconds of data (33.4 hours)
- DNN adaptation: final layer removed and replaced, modified DNN retrained updating segmentation and clustering, x new speakers in new output used to replace final layer and retrain in iterative fashion

Stage	Spkrs	MS	FA	SE	DER
Clustering	409	3.2	4.2	41.1	48.4
+Adaptation	333	4.6	4.1	37.7	46.4



Longitudinal Diarisation

Speaker linking

- Incremental method: AHC using BIC, allowing for merging within and across episodes
- Short time speakers discarded as these are less likely to be recurrent speakers

Stage	Spkrs	MS	FA	SE	linked DER
SpkrLink	312	4.6	4.1	42.0	50.8

Further details of the different genres, shows how different and difficult some genres are:

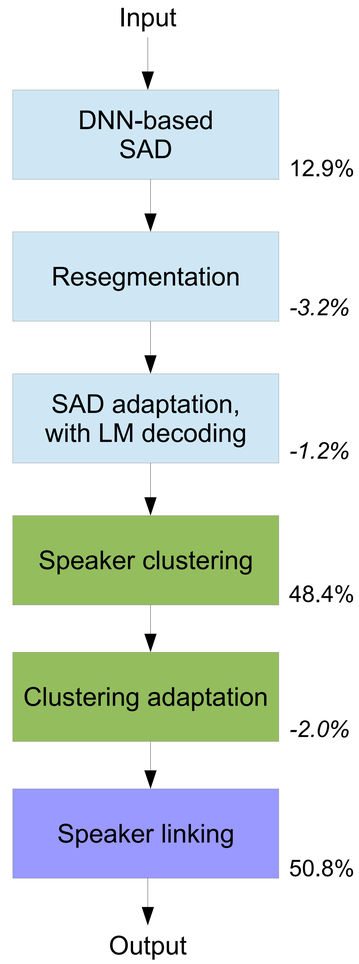
Genre	MS	FA	SE	unlinked DER	linked DER
Documentary (3)	3.9	1.5	15.8	21.1	22.0
Political drama (6)	4.0	2.5	28.6	35.1	36.8
Sci–fi drama (2)	11.0	1.6	59.6	72.1	75.7
Sitcom (6)	4.0	10.4	59.0	73.4	85.6
Sports event (2)	3.2	4.9	52.1	60.2	65.5

Discussion

Achieved best final segmentation result of 8.7% (MS+FA) and fourth best linking result of 50.8%.
Novel methods presented:

- Decoding speech/non-speech using language modelling
- DNN-based adaptation for speaker clustering

System Description



Acknowledgments

Supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology). The authors would also like to thank Jana Eggink and the BBC for supporting this work.