
Proceedings Paper:
Milner, R. and Hain, T. orcid.org/0000-0003-0939-3464 (2017) DNN approach to speaker diarisation using speaker channels. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 5-9, 2017, New Orleans, USA. IEEE , pp. 4925-4929. ISBN 9781509041176

https://doi.org/10.1109/ICASSP.2017.7953093

DNN APPROACH TO SPEAKER DIARISATION USING SPEAKER CHANNELS

Rosanna Milner, Thomas Hain

Speech and Hearing Research Group, University of Sheffield, UK

rmmilner2,t.hain@sheffield.ac.uk

ABSTRACT

Speaker diarisation addresses the question of "who speaks when" in audio recordings, and has been studied extensively in the context of tasks such as broadcast news, meetings, etc. Performing diarisation on individual headset microphone (IHM) channels is sometimes assumed to easily give the desired output of speaker labelled segments with timing information. However, it is shown that given imperfect data, such as speaker channels with heavy crosstalk and overlapping speech, this is not the case. Deep neural networks (DNNs) can be trained on features derived from the concatenation of speaker channel features to detect which is the correct channel for each frame. Crosstalk features can be calculated and DNNs trained with or without overlapping speech to combat problematic data. A simple frame decision metric of counting occurrences is investigated as well as adding a bias against selecting nonspeech for a frame. Finally, two different scoring setups are applied to both datasets. The stricter SHEF setup finds diarisation error rates (DER) of 9.2% on TBL and 23.2% on RT07 while the NIST setup achieves 5.7% and 15.1% respectively.

***Index Terms*—** speaker diarisation, multi-channel, crosstalk, deep neural networks, speaker channels

1. INTRODUCTION

The task of speaker diarisation is an important prerequisite task for audio indexing, automatic speech recognition (ASR) and more [1, 2]. The objective is to split the audio into segments which are associated with a single speaker, and to identify among the set of segments those that are spoken by the same speaker. Diarisation systems generally consist of three main stages: speech activity detection (SAD), speaker segmentation and speaker clustering. SAD aims to detect speech segments which are passed to a speaker segmentation stage to split the segments further at speaker change points (speaker boundaries). Speaker clustering aims to group speaker segments together into speaker-homogeneous clusters. The objective is not only to group the speakers correctly, but also to find the correct number of clusters (i.e. speakers). Diarisation has been well studied over the years, and toolkits are available for this task which are designed to perform well for a specific type of data [3, 4, 5].

The challenges to multi-channel diarisation differ by domain. For conversational telephone speech (CTS) only two speakers are present. However, channel echo, speaker overlap, poor quality phone lines and noise cause errors, despite independent channels for each speaker [2]. Broadcast news (BN) data has background noises such as music, but also a large number of speakers who may only occur very briefly [6, 7]. Meeting data has become the focus for diarisation for considerable time [8]. Speech is conversational with significant amounts of speaker overlap, as it is for CTS. However, there are more speakers, and speech may be recorded with distant or far-field microphones. Multi-channel diarisation operates in two different modes, depending on the distance between the microphones and the speakers: using beam-forming to focus on speakers [9]; or detecting automatically which speaker is closer and disregarding other speech [10, 11]. The former case is much harder. It helps beam-forming to know who speaks and when [12], but knowing where the speech is coming from can improve speaker segmentation performance [13, 14], e.g. through the use of inter-channel delay information [9]. Work presented here is related to the latter case: microphones are far apart and assigned to speakers, although not in close proximity to the speakers mouth.

Deep neural networks (DNNs) have been introduced into different stages of a diarisation system. Artificial neural networks (ANN) have been trained to learn a feature transform [15] and DNNs can be trained to detect speech/nonspeech in an SAD stage where adapting the DNN leads to improved performance [16]. A speaker segmentation stage using auto-associative neural networks (AANN) was proposed in which a windowing method is used where an AANN model is trained for the left half of the window and tested on the right to give a confidence score on how likely each part belongs to the same speaker [17]. Finally, DNNs have been applied to the clustering stage by training speaker separation DNNs and adapting these to specific recordings [16, 18].

Typically, speaker diarisation is unsupervised meaning no a priori information or metadata is used to aid a system. The desired output of a system is speaker labelled segments with timing information. Whether diarisation is performed unsupervised (ICSI system [19]), semi or lightly supervised (supplementary data such as imperfect transcripts [20]) or supervised (known speakers [21]), the desired output remains the

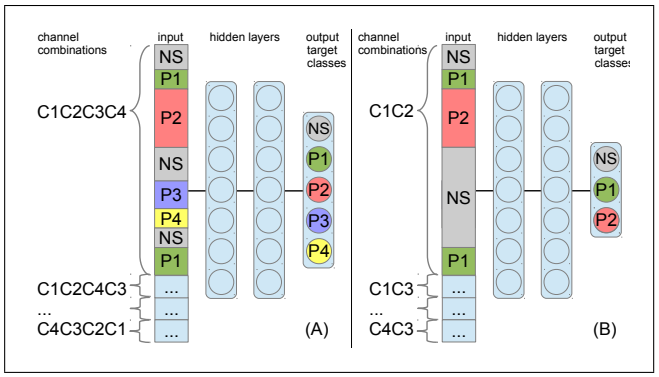

Fig. 1. Feature concatenations and input labels are shown for methods (A) a fixed number of channels for every recording and (B) a mixed number of channels across recordings.

same. It will be shown that the obvious method of performing diarisation on the individual headset microphone (IHM) channels is not satisfactory given imperfect data, such as channels containing heavy crosstalk. Thus, two methods are proposed which train DNNs to detect which channel contains the correct speaker at a given frame. Both methods concatenate speaker channel features in training and testing. The first concatenates all speaker channels from a recording so it requires each recording in a dataset to contain the same number of speakers. As this is not portable to datasets which do not have this trait, a second method is proposed which trains DNNs on pairs of speaker channels. Furthermore, the problems of crosstalk and overlapping speech are considered and as well as simple counting frame decision metric vs. adding a bias against selecting nonspeech.

2. DNN APPROACH USING SPEAKER CHANNELS

Two methods are presented: the first method is channel detection when the specific number of channels is fixed and the second is an extension to the first in which the data consists of a mixed number of channels.

2.1. Fixed number of channels per recording

DNNs are trained on concatenated features from all the speaker channels. It requires every recording to contain the same number of speakers. Every combination of the channels are used for training, as this may help prevent channels being biased in certain positions. Example (A) in Figure 1 depicts the ordering of the concatenated features with their equivalent label file for training. It assumes there are four IHM channels for every recording. The channels are referred to as C1, C2, C3, C4 while each speaker-pure segment is labelled as P1, P2, P3, P4 corresponding to the position of the relevant channel in the feature concatenation. Nonspeech is referred to as NS.

2.2. Mixed number of channels per recording

The fixed method is not portable to datasets which do not contain the same number of speakers in each recording. A different approach is required where pairs of features can be concatenated. Example (B) in Figure 1 displays how the channel pairs are annotated as before, where position labels are necessary to denote which channel contains speech and which is nonspeech. For instances where the speech segment does not belong to either channel, a nonspeech label is given.

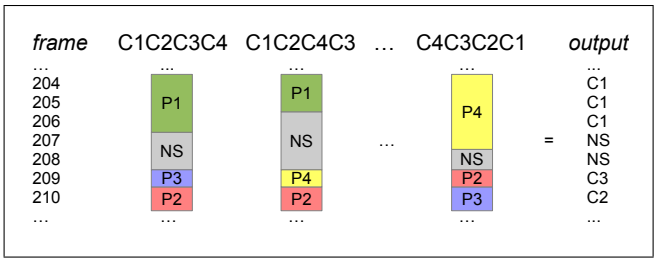

Fig. 2. Frame decisions are made considering the decoded outputs from all combinations of feature concatenations on the testset. The simple counting method gives the output displayed.

As well as being applicable to all datasets, this alternative approach also reduces the amount of data needed for training. For a single recording in the fixed method, the number of possible combinations for training is $x!$, where x in the number of channels. Whereas for this method, the number of possible feature pairs for training becomes $x(x-1)$. For example, if there are 4 channels then the amount of combinations needed for each method is 24 and 12 respectively.

2.3. Frame decision

All the combinations of feature concatenations are used for testing and this gives a channel or nonspeech label to every frame. This results in multiple labels for every frame, across the different decoded feature concatenations, as shown in Figure 2. To make a decision on the correct label, one can simply count the occurrences and select the channel or nonspeech that has been labelled the most. Alternatively, the occurrences can be counted as before with a bias for or against nonspeech applied as a multiplier to increase or reduce the likelihood of selecting nonspeech. A bias for or against specific channels could also be applied, for example if a host in a TV programme is known to talk more than the guests.

3. EXPERIMENTS

3.1. Data

The methods are evaluated with two datasets in different domains. TBL is TV broadcast data which consists of 22 programmes from a talk–show with single distant microphone (SDM) and IHM channels: four speakers as one host and three guests. The recordings have been split into a training set of 12 programmes for DNN training only, and a test set of 10 episodes which has a total of 40 speakers and 8749 segments in 5.3 hours of speech time. The audio was manually transcribed to an accuracy of 0.1s.

The second is based on the established testset from the NIST Rich Transcription evaluation in 2007 [8]. The com-

plete files were also manually transcribed to an accuracy of 0.1s[1], which produces a different reference to the original testset. This updated reference contains 8 conference meetings with both SDM and IHM channel data and contains 35 speakers and 11144 segments over 8.9 hours of speech time. Six meetings contain 4 participants, one has 5 and another 6.

3.2. Experimental setup

DNNs require training on concatenated IHM channels and log-Mel filterbanks of 23 dimension are used as opposed to Mel frequency cepstral coefficients (MFCCs) as they are found to yield better performance with DNNs [22]. Crosstalk features (denoted CT), of 7 dimensions, may help reduce errors caused by speech on the wrong channel [10]. The energies are normalised across all N channels by

$$E_i^{norm}(n) = \frac{E_i(n)}{\sum_{k=1}^{N} E_k(n)} \quad (1)$$

where $E_i(n)$ is the current channel i energy at frame n. Further features are calculated such as kurtosis [23] and mean cross-correlation and maximum normalised cross-correlation. DNNs for the fixed method are trained on TBL, whereas DNNs for the mixed method are trained on TBL and the AMI corpus [24]. The number of input neurons depends on the number of concatenated channels. For 4 channels, there are 1472 input neurons, increasing to 1920 with CT, two hidden layers of 1000 hidden units and 5 output neurons, which represent the 4 channels and nonspeech. For 2 channels, there are 736 neurons, increasing to 960 with CT, two hidden layers of 1000 hidden units and 3 output neurons, representing the 2 channels and nonspeech. Training on overlapping speech may cause DNNs to learn errors and affect the performance thus DNNs are trained with or without overlapping speech (denoted OV).

3.3. Diarisation evaluation

Diarisation error rate (DER) is the standard metric for speaker diarisation and is the sum of three error values: miss (MS), false alarm (FA) and speaker error (SE) [25]. The DER does not consider the segmentation quality in its evaluation of a system, so all tables depict the number of detected segments [**?**]. Two scoring methods are investigated. The standard evaluation method for RT07 data is to use a collar of 0.25s and score specific portions of time only, not complete recordings, with the NIST reference [8]. This will be referred to as the NIST setup. In terms of the TBL dataset for the NIST setup, the collar of 0.25s will be employed however the complete recordings will be evaluated with the manually transcribed reference. The second scoring setup will be referred to as SHEF. As both datasets have been manually transcribed to an accuracy of 0.1s, a stricter collar of 0.05s is used, and scoring occurs on the complete files with this reference.

[1] mini.dcs.shef.ac.uk/resources/dia-improvedrt07reference/

Scoring	Channel	#Segs	#Spkrs	DER%
Data: TBL				
NIST	SDM	2030	82	16.6
	IHM	8478	40	393.9
SHEF	SDM	2030	82	27.8
	IHM	8478	40	335.9
Data: RT07				
NIST	SDM	2648	72	37.9
	IHM	13070	35	308.1
SHEF	SDM	2648	72	66.4
	IHM	13070	35	371.0

Table 1. Baseline performance for both datasets on the SDM and IHM channels evaluated using both NIST and SHEF scoring setups, where #Segs represents the number of hypothesis segments and #Spkrs represents the number of speakers.

3.4. Baseline experiments

The public domain toolkit, LIUM_SpkrDiarization [4], is tailored for TV and radio broadcasts and consists of Bayesian information criterion (BIC) segmentation with cross-likelihood ratio and integer linear programming and i-vector clustering. Table 1 displays results for both datasets and a distinction is made between the two scoring setups as previously described: NIST and SHEF. Scoring also occurs on both SDM and IHM channels. For the SDM results for the TBL dataset, changing the collar has a dramatic effect on the DER, from 16.6% to 27.8% with the stricter collar. For RT07 SDM, the NIST scoring gives 37.9% against the SHEF result 66.4%, again, a large improvement in DER performance is seen. For the IHM results, the imperfect data has large amounts of crosstalk which negatively affects the performance and causes large false alarms from incorrectly detected speech for both datasets, seen in both scoring setups.

The SHEF setup is a stricter scoring method however arguably more reliable to show the true performance given the more accurate references. The rest of the paper will use this scoring method. The NIST setup can be seen as more lenient scoring as 0.25s collar around every boundary is a large portion of time to ignore from evaluation. However, for the results to be comparable to other papers, the best result will be scored in the NIST setup at the end.

3.5. Results

Results for the fixed method can be seen in Table 2 for the TBL dataset, in which there are 4 channels per recording. The DERs are relatively similar apart from the DNN trained on TBL+CT where the number of segments detected is dramatically less than the other three. The DNN trained on TBL+OV achieves the lowest DER of 8.0% with the lowest SE of 1.2%. Training DNNs with crosstalk features degrades the result commpared to DNNs without.

Table 3 displays the performance when the frame decision metric involves a bias against the nonspeech (NS) occur-

DNN			#Segs	MS%	FA%	SE%	DER%
TRN	OV	CT					
Data: TBL							
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

Table 2. Results for the DNNs trained with 4 fixed channels across recordings, with the counting frame decision metric.

NS bias	#Segs	MS%	FA%	SE%	DER%
Data: TBL, DNN: TBL+OV					
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

Table 3. Results when a bias against nonspeech is introduced for the frame decision metric for 4 channels concatenated, specifically for DNN TBL+OV.

rences, multiplier is specified in the table. Errors in the miss rate are reduced but these seem to be moved to the false alarm and speaker error, thus increasing the DERs by 0.2-0.3%.

Table 4 displays results for the mixed method and two additional DNNs are trained on AMI data. Comparing the TBL results to the previous fixed method, more segments are found here although the performance is worse overall. Training DNNs with OV does not help performance as it does in the fixed method. The baseline of 27.8% DER is beaten in all but two of the trained DNNs. A dramatically higher miss rate than the false alarm and speaker error is seen across the trained DNNs. This could imply the counting metric is too simple as nonspeech is selected over the channels. The best DNN is trained on TBL+CT and achieves a DER of 10.9%, the only DNN which improves with CT. The DNNs trained on AMI more than double the error. For RT07, again a large amount of miss across the DNNs is seen, implying a nonspeech bias could help. The DERs are high and range from 58.2% to 80.1% which does not seem promising. The DNNs trained on AMI do not outperform the TBL trained DNNs. The lowest DER is found with the DNN trained on TBL only.

Based on the miss rate reported in Table 4, it can be found that nonspeech is detected often. Table 5 shows the performance when a bias against nonspeech is introduced. As the bias decreases, the likelihood of selecting nonspeech is decreased and the amount of missed speech detected is reduced. For TBL, this is a small gain from 10.9% to 9.2% with a bias of 0.25. However, a large gain is seen for the RT07 dataset which jumps from 58.2% to 23.2% DER with the same bias. These lowest results with the NIST setup would change to 5.7% for TBL and 15.1% for RT07.

4. CONCLUSION

Two methods for training DNNs to detect the correct speaker channel for the purposes for speaker diarisation are presented.

DNN			#Segs	MS%	FA%	SE%	DER%
TRN	OV	CT					
Data: 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
Data: 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

Table 4. Results for the DNNs trained with mixed channels across recordings, with the counting frame decision metric.

NS bias	#Segs	MS%	FA%	SE%	DER%
Data: TBL, DNN: TBL+CT					
0.75	7950	7.3	1.9	1.3	10.6
0.5	7420	5.4	2.4	1.5	9.4
0.25	7468	4.9	2.6	1.7	9.2
Data: RT07, DNN: TBL					
0.75	9940	39.5	1.5	0.6	41.5
0.5	11983	20.3	3.2	0.9	24.4
0.25	13898	14.0	7.4	1.8	23.2

Table 5. Results when a bias against nonspeech is introduced for the frame decision metric for pairs of channels concatenated, specifically for DNN TBL+CT for the TBL dataset and DNN TBL for the RT07 datset.

The first requires a fixed number of speaker channels across recordings and concatenates speaker channel features for training and testing. The second does not require a fixed number of speaker channels and concatenates pairs of features. These were evaluated using two datasets with the former finding the best DER for the TBL dataset, however, it is not applicable to datasets with varying numbers of speaker channels and requires more training data. The mixed method performs well for both TBL and RT07 datasets and achieves best results when a bias against nonspeech is applied, giving 9.2% and 23.2% respectively for the stricter scoring setup. For the NIST setup, this reduces to 5.7% and 15.1% DER.

5. ACKNOWLEDGEMENTS

The authors would like to thank Jana Eggink and the BBC for supporting this work and providing the data. This work was also supported by the EPSRC Programme Grant EP/I031022/1 Natural Speech Technology. Results are found here: https://dx.doi.org/10.6084/m9.figshare.4312469.v1

6. REFERENCES

[1] X. M. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker Diarization: A Review of Recent Research," *IEEE Trans. Audio Speech and Language Processing*, vol. 20, no. 2, pp. 356–370, Feb. 2012.

[2] S. E. Tranter, "Who Really Spoke When? Finding Speaker Turns and Identities in Broadcast News Audio," *ICASSP*, vol. 1, pp. 1013–1016, 2006.

[3] M. A. H. Huijbregts, "Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled," Ph.D. dissertation, 2008.

[4] M. Rouvier, G. Dupuy, P. Gay, E. el Khoury, T. Merlin, and S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," in *INTERSPEECH, Lyon, France*, 2013, pp. 1477–1481.

[5] D. Vijayasenan and F. Valente, "DiarTk: An Open Source Toolkit for Research in Multistream Speaker Diarization and its Application to Meetings Recordings." *INTERSPEECH*, pp. 5–8, 2012.

[6] T. Hain and P. C. Woodland, "Segmentation and classification of broadcast news audio," in *ICSLP'98*, 1998.

[7] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, "Step-by-Step and Integrated Approaches in Broadcast News Speaker Diarization," *Comput. Speech Lang.*, vol. 20, no. 2-3, pp. 303–330, Apr. 2006.

[8] "NIST Rich Transcription Evaluations," www.itl.nist.gov/iad/mig/tests/rt/, accessed: 12-09-2016.

[9] X. Anguera, C. Wooters, and J. Hernando, "Acoustic Beamforming for Speaker Diarization of Meetings," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 7, pp. 2011–2022, Sep. 2007.

[10] J. Dines, J. Vepa, and T. Hain, "The Segmentation of Multi-Channel Meeting Recordings for Automatic Speech Recognition," *INTERSPEECH*, 2006.

[11] S. N. Wrigley, G. J. Brown, V. Wan, and S. Renals, "Speech and Crosstalk Detection in Multi-Channel Audio," *IEEE Transaction Speech and Audio Processing*, vol. 13, pp. 84–91, 2005.

[12] D. P. W. Ellis and J. Liu, "Speaker Turn Segmentation Based on Between-Channel Differences," *ICASSP*, p. 12, 2004.

[13] U. Anliker, J. F. Randall, and G. Troster, "Speaker Separation and Tracking System," *EURASIP Journal of Applied Signal Processing*, vol. 2006, pp. 1–14, 2006.

[14] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped Speech Detection for Improved Speaker Diarization in Multiparty meetings," *ICASSP*, pp. 4353–4356, 2008.

[15] S. H. Yella, A. Stolcke, and M. Slaney, "Artificial neural network features for speaker diarization," in *SLT, South Lake Tahoe, NV, USA, December 7-10*, 2014.

[16] R. Milner, O. Saz, S. Deena, M. Doulaty, R. W. M. Ng, and T. Hain, "The 2015 sheffield system for longitudinal diarisation of broadcast media," in *ASRU, Scottsdale, AZ, USA*, 2015, pp. 632–638.

[17] S. Jothilakshmi, V. Ramalingam, and S. Palanivel, "Speaker diarization using autoassociative neural networks," *Eng. Appl. of AI*, pp. 667–675, 2009.

[18] R. Milner and T. Hain, "DNN-based speaker clustering for speaker diarisation," in *INTERSPEECH, San Francisco, CA, USA*, 2016, pp. 2185–2189.

[19] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," *Multimodal Technol. Percept. Humans*, vol. 4625, pp. 509–519, 2008.

[20] A. Stan, Y. Mamiya, J. Yamagishi, P. Bell, O. Watts, R. Clark, and S. King, "ALISA: An automatic lightly supervised speech segmentation and alignment tool," *Computer Speech and Language*, vol. 35, pp. 116–133, 2016.

[21] D. Moraru, L. Besacier, and E. Castelli, "Using a priori information for speaker diarization," in *ODYSSEY 2004 - The Speaker and Language Recognition Workshop, Toledo, Spain, May 31 - June 3, 2004*, 2004, pp. 355–362.

[22] H. Hermansky and S. Sharma, "Traps – Classifiers Of Temporal Patterns," in *ICSLP*, 1998, pp. 1003–1006.

[23] J. P. LeBlanc and P. L. D. Leon, "Speech separation by kurtosis maximization," in *ICASSP, Seattle, Washington, USA, May 12-15, 1998*, 1998, pp. 1029–1032.

[24] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI Meeting Corpus: A Pre-announcement," vol. 3869, pp. 28–39, 2006.

[25] "Diarisation error rate scoring code, NIST," www.itl.nist.gov/iad/mig/tests/rt/2006-spring/code/md-eval-v21.pl, accessed: 12-09-2016.