THE 2015 SHEFFIELD SYSTEM FOR LONGITUDINAL DIARISATION OF BROADCAST MEDIA

Rosanna Milner, Oscar Saz, Salil Deena, Mortaza Doulaty, Raymond W. M. Ng, Thomas Hain

Speech and Hearing Research group, Department of Computer Science, University of Sheffield, UK

ABSTRACT

Speaker diarisation is the task of answering "who spoke when" within a multi-speaker audio recording. Diarisation of broadcast media typically operates on individual television shows, and is a particularly difficult task, due to a high number of speakers and challenging background conditions. Using prior knowledge, such as that from previous shows in a series, can improve performance. Longitudinal diarisation allows to use knowledge from previous audio files to improve performance, but requires finding matching speakers across consecutive files. This paper describes the University of Sheffield system for participation in the 2015 Multi–Genre Broadcast (MGB) challenge. The challenge required longitudinal diarisation of data from BBC archives, under very constrained resource settings. Our system consists of three main stages: speech activity detection using DNNs with novel adaptation and decoding methods; speaker segmentation and clustering, with adaptation of the DNN-based clustering models; and finally speaker linking to match speakers across shows. The final result on the development set of 19 shows from five different television series provided a Diarisation Error Rate of 50.77% in the diarisation and linking task.

Index Terms— speaker diarisation, linking, neural networks, adaptation

1. INTRODUCTION

Speaker diarisation is the task of "who spoke when?" within an audio recording [1, 2]. This is typically performed in three stages: speech activity detection (SAD), speaker segmentation and speaker clustering. Diarisation is traditionally unsupervised and clustering is most commonly performed using agglomerative hierarchical clustering (AHC) with the Bayesian information criterion (BIC) [3] as the similarity measure and stopping criterion. Longitudinal diarisation (speaker linking [4] or partitioning [5]) is diarisation across a collection of connected audio recordings. For example, these could be meetings held by a single group recorded over a few months, or, in this case, a TV series. Speaker linking aims to cluster across recordings to find the speakers which occur in more than one recording. The common method proposed involves agglomerative clustering without a model retraining step and pairs clusters by using the closest segment pairing distance as the score for the cluster pair [4, 6]. Alternatively, complete-linkage clustering works by taking the furthest distance in terms of segment pairings as the score for each cluster pair [7]. Early work was carried out on two-speaker telephone conversations only but since has been extended to meetings [8]. Speaker linking is also referred to cross-show diarisation in the context of broadcast media [9, 10, 11].

While transcription is the most common task in the evaluation of broadcast media systems, speaker diarisation has also been tackled as a task in several challenges. The ESTER [12] and REPERE [13] evaluation campaigns have used French broadcasts to develop diarisation systems, and Albayzin [14] has used Spanish broadcast news data for the same data. The Multi–Genre Broadcast (MGB) challenge, as part of its goal of improving spoken language technology for general broadcast media, has proposed the task of longitudinal speaker diarisation as one of its main components.

The system consists of sveral stages: speech activity detection (SAD), speaker segmentation and clustering, and speaker linking. The SAD is performed using deep neural networks (DNNs) trained to distinguish speech and non–speech. Adaptation is then performed using an improved DNN output. The output is further improved by decoding using a novel duration based language model (LM) approach for the speech and non–speech states. Speaker segmentation and clustering is performed using a standard toolkit, which is unsupervised. Thus, it was suitable to use within this challenge. The second part is again adaptation using a pre-trained DNN to classify or separate speakers, based on a novel approach of speaker clustering. Finally, the speaker linking stage uses BIC to test whether speakers with the largest amount of speaking time should be merged across shows.

The paper is organised as follows: section §3 describes the SAD stage which includes SAD adaptation and decoding using a language model setup, section §4 describes the diarisation with the DNN cluster adaptation and, finally, section §5 describes the speaker linking stage.

This work was supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

2. THE MGB CHALLENGE - TASK 4

The Multi-genre broadcast (MGB) challenge consisted of four different tasks covering the topics of multi–genre broadcast show transcription, lightly supervised alignment, longitudinal broadcast transcription and longitudinal speaker diarisation. The focus of this work was on Task 4: longitudinal diarisation of broadcast television. A full description of this and the other tasks in the challenge can be found in [15], but a brief description of the task is given here.

The task proposed the automatic diarisation of a set of shows broadcast by the British Broadcasting Corporation (BBC). The training data was fixed and limited to more than 2,000 shows, broadcast by the BBC during 6 weeks in April and May of 2008. The development data for the task was 19 shows covering 5 series broadcast by the BBC during June and July of 2008. The amount of shows and broadcast time for training and development data is shown in Table 1. For the training data no speaker labels were provided, and the time of speech segments was semi-automatically derived, in a lightly supervised training setting [15].

Table 1. *Amount of training and development data.*

Train		Development			
Shows	Time	Series	Shows	Spkrs	Time
2,193	1580.5h.	5	19	464	9.3h.

The five series in the development set consisted of: 3 episodes of a nature documentary show, 6 episodes of a political drama series, 2 episodes of a science fiction drama, 2 episodes of a sporting event and 6 episodes of a situation comedy series. These series had a large range of speakers, with both re-occurring speakers and speakers confined to one programme.

The date and time of broadcast for each show was provided as well as the series name. The diarisation of speakers across different episodes of the same series was restricted by allowing only episodes broadcast in previous dates to be used to perform this stage in a given episode. Episodes from future dates were not allowed under any situation to affect the diarisation of any episode.

Diarisation error rate (DER) was the metric identified by the MGB challenge to measure the speaker diarisation results. DER is a commonly used metric that is defined as the sum of three frame error values: miss (MS), false alarm (FA) and speaker error (SE) [1, 2, 16]. Missed speech refers to reference speech detected as silence, false alarm is reference silence detected as speech, and speaker error measures the percentage of scored time in which a speaker label is assigned to the wrong speaker. All miss and false alarm numbers were calculated using the total speaker scored time, as the scoring was set to ignore overlap. A standard collar of 0.25 seconds around the boundaries was applied. It is important to note that DER does not penalise for the creation of many short segments directly. Hence there is typically no direct relation between ASR outcome and DER outcomes, which also justifies diarisation to be a separate task.

3. SPEECH ACTIVITY DETECTION

Speech and non–speech segmenters were built using deep neural networks (DNNs). Further to this, the output segmentation was improved before being used in adapting the new segments to the original DNN model. The final segmentation was achieved by decoding using a language model setup for speech and non–speech states.

3.1. DNN based segmentation

All DNNs were trained with TNet[1] [17] using filterbank features of 23 dimensions with a context window of 15 frames on both sides. Log Mel-filterbanks were used as opposed to Mel frequency cepstral coefficients (MFCCs) or perceptual linear predictive (PLPs) features as they are found to have better performance with DNNs in this setting [18]. There were 368 input nodes, 2 hidden layers of 1000 nodes and an output layer consisting of 2 nodes only, for speech and non–speech.

Training data for speech and non–speech segmenter was selected from the complete training set. Forced alignment with the transcript was first performed to refine the timings of phoneme segments, in order to better separate silence and non-speech portions. All phoneme segments were considered as speech, while all other portions, including the gaps between transcribed segments, were considered as non-speech. This resulted in 759 hours of speech and 793 hours of non–speech. Using this data, an initial segmenter, *SNS1*, was trained.

For an alternative segmenter, *SNS2*, an attempt was made to constrain training data to those portions of relatively good annotation quality. The training segments were decoded and the hypothesis words and phones were compared with the reference to obtain word and phone matching error rates. Data selection was carried out to reject segments with both matching error rates greater than 40%. Furthermore, word alignment should give a word duration between 0.3 and 0.7 seconds, otherwise the segment was rejected. The programmes were then further split into chunks of 60 seconds containing speech and non–speech segments. If a chunk contained a segment which was rejected under the previous constraints, all the segments within that chunk were considered unreliable to be used for training. This resulted in 116 hours of speech and 363 hours of non–speech. A number of alternative selection methods were tried where selection criteria and speech / non–speech balance varied, but this method of data selection gave the optimum result.

Table 2 shows the results on the full development set. The results have been tuned for the lowest DER (which for segmentation is the sum of MS and FA), by considering the num-

[1] http://speech.fit.vutbr.cz/software/neural-network-trainer-tnet

ber of states (fixing the state duration to enforce a minimum duration), the prior probability for non–speech and the grammar scale factor. Padding, or widening, of the output segments was also applied. This added 0.25 seconds to the beginning and end of every speech segment. A segmenter with high miss rate is detrimental in initial passes because these errors cannot be easily recovered in subsequent steps. With the lowest miss rate, *SNS2* together with padding has been used for the rest of the paper.

Table 2. *Results for SNS1 and SNS2 on the development data, where MS refers to the missed speech error and FA refers to the false alarm error.*

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

3.2. SAD adaptation and duration language models

Before performing adaptation and decoding, the segments were refined. This is done by first employing a speech recognition system to obtain a sentence hypothesis for each segment. A 144-monophone-state-target DNN[19] was then used to obtain confidence measures for each word in the hypothesis. The raw posterior values were mapped to confidence scores using a decision tree trained on the development data. The decision targets were either 1 if the word was in an area of speech as defined in the reference, or 0 if the word was in an area of non-speech as given by human annotation. The raw confidence score of each word, the confidence score of the segment, the length of the word (in seconds), the length of the word (in phonemes) and the length of the segment (in seconds) were used as input features to this decision tree. Once the confidences were calculated, words with confidence score below a threshold were removed from the transcript. The remaining words were used to define the new segments.

Using *SNS2*, these new speech segments are used to train a new DNN model. Since it is important to adjusts to both the specific speech and the specific noise, the gaps between the redefined speech segments are now also considered as non–speech. Further iteration of DNN training using this data, on a per show basis is performed, using a standard cross-entropy criterion and stochastic gradient descent.

The adapted DNN is used in hybrid decoding. Typically, the DNN has two output states: speech and non–speech. A state graph with adjustable minimum state duration, prior probabilities, transition probabilites and graph (grammar) scale factor sits on top of the DNN to give decoding results. This framework is similar to the one described in [20]

The combination of dictionary and grammar (language model) serves to control the duration patterns in such settings. Assuming balanced distribution between speech and nonspeech segments or assuming equal duration between these appears to be inappropriate. For this reason the use of duration models was explored. For this purpose the full training data (as used to train the *SNS1* models) was categorised into "duration words". A set of duration boundaries was defined, in our case 4 seconds, 7 seconds and 10 seconds. If a segment of speech or nonspeech in the training set was shorter than 4 seconds then it was translated into a unique tag, here D400. Segments between 4 and 7 seconds received tag D700 and finally segments with duration of more than 10 seconds received tag D1000. Given such labels a duration class language model (bigram) can be trained. Each of the duration words can then be match with segments of length with the duration bounds, i.e. 0-4 seconds, etc. Matching HMMs are constructed in a way to allow exit in those time ranges. In practice this was implemented using a standard Viterbi decoder and a dictionary with different pronunciation variants and granularity of 0.2 seconds (i.e. only segments of multiples of that duration can be produced).

Experiments where conducted using different duration boundaries and different number of duration classes. The aforementioned boundaries of 4, 7 and 10 seconds gave the best performance on the development set, although by only small margins compared to many other settings. Experiments also investigated genre-dependence of such language models but no significant perplexity gain was obtained and hence such models were not considered further.

4. SPEAKER DIARISATION

Speaker segmentation and clustering was initially performed using a standard toolkit. The clustered output (speaker homogeneous segments with cluster labels) was then used to adapt a DNN trained for speaker classification.

4.1. SHoUT

SHoUT[1] was originally designed for the diarisation of meetings and uses BIC segmentation and a BIC stopping criterion in an unsupervised model training regime [21, 22, 23]. As the complete system is unsupervised (i.e. not trained on other data), it was usable within this challenge. For diarisation, SHoUT conducts an initial pass using the speech only segments. These are first randomly split into clusters of speaker-pure segments. Models for the clusters are iteratively trained and realigned to the speech data to produce speaker models. BIC is used to find the two most similar models which are then merged and the retraining repeated. BIC is also used to stop the clustering process.

[1] http://shout-toolkit.sourceforge.net/

4.2. Speaker Clustering Adaptation

We introduce a novel approach to improve speaker clustering. A speaker separation DNN [24] is trained on data from the training set. Again, log Mel-filterbanks are used and the structure is an input layer of 368 nodes, three hidden layers with 1000 nodes, and a bottleneck layer with 26 nodes. The number of nodes in the final layer is the number of speakers in the training data. Speaker separation DNNs need to be trained on speaker homogeneous segments which have a cluster (speaker) label. This was not available for the training set in the MGB challenge. Only the official development set contained speaker labels, which is a limited 9.3h of data and ideally should not be used for training.

For the training set, the speaker names in the original files contained the subtitle colour (as displayed on TV screens) as a way to distinguish speaker changes. These cannot be considered as speaker clusters as there are only four colours per show, and one speaker may be covered by different colours throughout one programme. Furthermore, sometimes the colours are used to emphasize words from the same speaker. The segments were aligned and then clustered automatically, yielding new hypothesized speaker labels. These were then matched up with the original subtitle speaker colours. This allowed us to derive segments which were speaker-pure and and therefore to select clusters which were spread across only one colour.

Next, these segments were reclustered using our BIC-based clusterer which was tuned to the development set for the lowest DER. The clusters were then filtered to only keep those with at least 40 seconds duration and then every segment was taken in at the beginning and end by 20ms with the aim to remove silence – the opposite of padding. Finally, the resulting segments were split into smaller chunks to help improve the DNN training by having more data of each speaker spread across the training list. This resulted in 53501 segments and 2495 speakers over 33.4 hours, roughly 50 seconds per speaker on average. The main issue with the data was the small amount of speech available for each speaker.

The speaker separation DNN has a final output layer of 2495, the number of speakers in the dataset. To perform adaptation on the clustering, the final layer was removed and a final layer was randomly initialised to the size of the number of speakers in the SHoUT output. An iteration of DNN training was performed to resegment and cluster.

5. LONGITUDINAL DIARISATON

The proposed method in this work, to perform diarisation across episodes from the same series, is based on performing the diarisation within a given episode independently of all previous episodes in the series. This is followed by a post-processing stage where speakers in the current episode are matched to speakers in previous episodes (i.e. linked). The alternative option, where the diarisation within the current episode is already informed of the speakers found in previous episodes, was not explored here.

The speaker linking stage was performed as follows: speakers in the current episode were ranked according to the amount of speaker time assigned to them, as well as speakers in previous episodes. Speakers with an amount of speaking time below a certain threshold were discarded, as short–timed speakers were found to be more likely to be non–recurrent speakers. BIC measures were calculated from each speaker in the current episode with respect to all the speakers in the previous episodes. If the lowest of these values was under the defined threshold, both speakers were given the same tag and, thus, considered the same speaker. If two speakers from the current episode were linked to the same speaker, they were effectively merged, being this the only instance where the within–episode diarisation was affected by the longitudinal diarisation.

Table 3 shows the effect on the threshold of speaker time when performing the diarisation across episodes. The original within–recording diarisation was based on SHoUT and had a SE of 37.7% and a DER of 46.4%. This Table shows how reducing the amount of speaking time to allow speakers to be linked across episodes increased the number of speakers eventually linked. However, this did not always provided a reduction in linked DER.

Table 3. *Effect of the threshold in speaker duration on the diarisation across episodes.*

Threshold	Speakers linked	SE	linked DER
1000 sec.	0	65.7%	74.5%
750 sec.	2	58.3%	67.1%
500 sec.	3	46.5%	55.3%
400 sec.	5	42.1%	50.9%
300 sec.	6	42.2%	51.0%
200 sec.	8	41.9%	50.7%
100 sec.	18	46.3%	55.1%
10 sec.	38	57.8%	66.6%

When no speakers were linked the SE increased to 66%, 30% more than scoring the series as individual shows. SE reduced as the speakers with the highest amount of speaking time were linked. For instance, when 8 common speakers were found across episodes, the SE was 41.9%, a mere 4% more than scoring series as individual shows. But when more speakers were attempted to be linked, the SE dramatically increased, when 38 speakers where linked, to 57.8%. The reasons for this could be twofold: first, due to the nature of the broadcast shows, only a small number of speakers may recur from episode to episode, with a large number of speakers appearing only in an episode. Second, errors in the initial within–recoding clustering could degrade very quickly the ability of the proposed system to correctly link speakers.

6. SYSTEM DESCRIPTION

The final system as implemented for the MGB challenge submission followed the diagram pictured in Figure 1. Each node in the diagram was implemented as a composition of separate modules, each performing specific computation on the speech data. The input audio was split into speech segments using a DNN segmenter based on the $SNS2$ strategy, as defined in Section 3.1. These segments were decoded by an initial, unadapted $Hybrid$ ASR system: *ASR-P1*. The ASR system used a DNN consisting of 6 hidden layers of 1,000 neurons each, and an output layer of 6,000 triphone state targets. State-level Minimum Bayes Risk (sMBR) [25] criterion and Stochastic Gradient Descent (SGD) was used for training. Decoding with $Hybrid$ systems was performed in two stages; in the first stage, lattices were generated using a highly pruned 3-gram. Afterwards the lattices were rescored using a complete 4-gram and the 1-best obtained. This 1-best output was then used for resegmenting the audio. The segmentation was then refined using confidence measures in the ASR output as in Section 3.1. This was followed by the SAD adaptation with LM decoding described in section 3.2. Subsequently speaker clustering using SHoUT was performed to assign a speaker label to each speech segment. This is followed by speaker clustering adaptation and, finally, speaker linking.

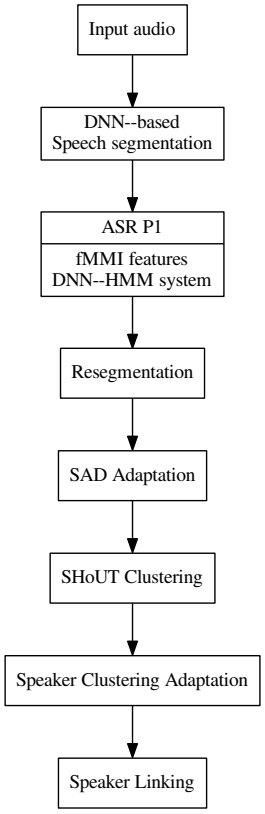

Fig. 1. System diagram

Table 4. *Segmentation using SNS2+Padding, where DER is simply the sum of missed and false alarm speech.*

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

Table 5. *Results for the speaker segmentation, clustering and linking stages.*

Stage	Spkrs	MS	FA	SE	unlinked DER	linked DER
SHoUT	409	3.2	4.2	41.1	48.4	-
+Adaptation	333	4.6	4.1	37.7	46.4	-
SpkrLink	312	4.6	4.1	42.0	-	50.8

6.1. System implementation

The implementation of the system is based on the Resource Optimisation Toolkit (ROTK), which is developed by the team at the University of Sheffield and was presented initially in [23]. ROTK allows the formulation of functional modules that can be executed in asynchronous fashion using computing grid infrastructure. Systems are defined as a set of modules linked together by directed links transferring data of specific types. This is informally depicted in a graph in Figure 1, the actual modules used are more specific. The system uses metadata to organise how data is processed. Each module can split its own tasks into several subtasks based on data, which then can be processed in parallel. The overall dependency structure of these sub-tasks is then automatically inferred. The ROTK system allows for simple repeatability of the experiments as the same graph can be executed on multiple datasets such as development and evaluation sets.

7. RESULTS

Table 4 shows the performance for the different stages of the speech activity detection. Refining the segmentation gives a considerable reduction in DER, a relative reduction of 25.4%. However, it changes the balance of MS and FA errors. Ideally, lower miss rates are better than lower false alarm rates as missed speech is usually harder to recover. The adaptation, where decoding is performed using the duration LM, helps to reduce the DER further and reduces the miss error at the expense of the false alarm error.

The speaker segmentation and clustering results are displayed in Table 5. The SHoUT toolkit resegments the input which is the cause for the different miss and false alarm to the best segmentation result. It reduces the segmentation rate further (the sum of the miss and false alarm is now 7.4%) but unfortunately it gives a high speaker error, probably because of the difficult nature of the data. The DNN adaptation at this stage both re-segments and clusters the data. The segmen-

Table 6. *Final results per series, where the number of shows is listed in brackets.*

Series	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

tation score increases slightly to 8.9% but the speaker error reduces by 3.4% absolute, improving the unlinked DER to 46.4%. Despite the high unlinked DERs, the results show that the clustering adaptation helps performance.

The speaker linking result is also displayed in Table 5 and this increases the unlinked DER to 50.8% linked DER. The number of speakers changes from 333 down to 312. This means 21 speakers have been found to occur on more than one programme. Unfortunately, it is the speaker error which increases to give a result higher than SHoUT both with and without the adaptation.

Finally, Table 6 shows the final results for the five series of shows which were part of the development set. Here it can be seen how the Documentary and Political Drama series achieved the lowest unlinked DER and linked DER values, and that there is no loss in the diarisation across series, which indicates that recurrent speakers have been found. A significant degradation in performance occurs in the Sci-fi Drama, the Sitcom and the Sports Event series, which manifests the large difficulty in diarising these shows which have a large diversity in recording conditions and existing speakers.

8. CONCLUSION

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

This paper introduced several new methods to improve the performance of speech segmentation: we improved segment generation for broadcast media by use of speech recognition, confidence scores and decision trees; we introduced show based DNN adaptation for segmentation; and new duration class LMs where used in decoding. We further introduced a new method for speaker clustering using DNNs and proposed a simple although effective method for speaker linking. All of the techniques were combined into a single system of processing stages. Each stage reduced the overall error rate. For initial clustering, SHoUT was used along with adaptation on a DNN trained to separate speakers. This clustering adaptation again helped to reduce the unlinked DER. The final stage, speaker linking, was performed after diarisation on each recording, achieving a final result of 50.77% linked DER. The results per series showed the large variability of the results across the different series.

9. ACKNOWLEDGEMENT

We thank members of the MINI group at Sheffield as well as our colleagues on the NST programme for the constructive discussions. We also thank the BBC for providing support to work on this data, beyond the NST programme.

The audio and subtitle data used for these experiments was distributed as part of the MGB Challenge (www.mgb-challenge.org) and was made available through a licence with the BBC. System output and complete results for the presented system is also available as part of the challenge results to participants.

10. REFERENCES

[1] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[2] X. A. Miró, S. Bozonnet, N. W. D. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[3] S. S. Chen and P. S. Gopalakrishnan, "Clustering via the Bayesian information criterion with applications in speech recognition," in *ICASSP*, (Seattle, WA), pp. 645–648, 1998.

[4] D. A. van Leeuwen, "Speaker linking in large data sets," in *Odyssey 2010, Brno, Czech Republic, June 28 - July 1, 2010*, p. 35, 2010.

[5] N. Brümmer and E. de Villiers, "The speaker partitioning problem," in *Odyssey 2010, Brno, Czech Republic, June 28 - July 1, 2010*, p. 34, 2010.

[6] C. Vaquero, A. Ortega, and E. Lleida, "Partitioning of two-speaker conversation datasets," in *INTERSPEECH, Florence, Italy, August 27-31, 2011*, pp. 385–388, 2011.

[7] H. Ghaemmaghami, D. Dean, R. Vogt, and S. Sridharan, "Speaker attribution of multiple telephone conversations using a complete-linkage clustering approach," in *ICASSP 2012, Kyoto, Japan, March 25-30, 2012*, pp. 4185–4188, 2012.

[8] M. Ferras and H. Boudard, "Speaker diarization and linking of large corpora," in *IEEE SLT, Miami, FL, USA, December 2-5, 2012*, pp. 280–285, 2012.

[9] V. Tran, V. B. Le, C. Barras, and L. Lamel, "Comparing multi-stage approaches for cross-show speaker diarization," in *INTERSPEECH, Florence, Italy, August 27-31, 2011*, pp. 1053–1056, 2011.

[10] Q. Yang, Q. Jin, and T. Schultz, "Investigation of cross-show speaker diarization," in *INTERSPEECH Florence, Italy, August 27-31, 2011*, pp. 2925–2928, 2011.

[11] 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, August 25-29, 2013*, pp. 1477–1481, 2013.

[12] S. Galliano, E. Geoffrois, G. Gravier, J. F. Bonastre, D. Mostefa, and K. Choukri, "Corpus description of the ESTER evaluation campaign for the rich transcription of french broadcast news," in *LREC*, (Genoa, Italy), pp. 139–142, 2006.

[13] O. Galibert and J. Kahn, "The first official REPERE evaluation," in *SLAM*, 2013.

[14] M. Zelenak, H. Schulz, and J. Hernando, "Speaker diarization of broadcast news in Albayzin 2010 evaluation campaign," *EURASIP Journal on Audio, Speech and Music Processing*, vol. 19, pp. 1–9, 2012.

[15] P. Bell, M. J. F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Webster, and P. Woodland, "The MGB Challenge: Evaluating Multi–genre Broadcast Media Transcription," in *ASRU 2015, Scottsdale, AZ, 2015*, 2015.

[16] "Diarisation error rate scoring code, NIST." http://www.itl.nist.gov/iad/mig/tests/rt/2006-spring/code/md-eval-v21.pl. Accessed: 08-07-2015.

[17] "Neural Network Trainer TNet, Brno University." http://speech.fit.vutbr.cz/software/neural-network-trainer-tnet. Accessed: 08-07-2015.

[18] H. Hermansky and S. Sharma, "TRAPS - classifiers of temporal patterns," in *The 5th International Conference on Spoken Language Processing, Incorporating The 7th Australian International Speech Science and Technology Conference, Sydney Convention Centre, Sydney, Australia, 30th November - 4th December 1998*, 1998.

[19] P. Zhang, Y. Liu, and T. Hain, "Semi–supervised DNN training in meeting recognition," in *Proceedings of SLT*, (South Lake Tahoe, CA), 2014.

[20] J. Dines, J. Vepa, and T. Hain, "The segmentation of multi-channel meeting recordings for automatic speech recognition," in *Interspeech'06*, 2006.

[21] M. Huijbregts, R. Ordelman, L. van der Werff, and F. M. G. de Jong, "SHoUT, the university of twente submission to the n-best 2008 speech recognition evaluation for dutch," in *INTERSPEECH, Brighton, United Kingdom, September 6-10, 2009*, pp. 2575–2578, 2009.

[22] M. Huijbregts, *Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled*. Phd thesis, University of Twente, 2008.

[23] T. Hain, L. Burget, J. Dines, P. Garner, F. Grezl, A. Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, "Transcribing meetings with the AMIDA systems," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 20, no. 2, pp. 486–498, 2012.

[24] Y. Liu, P. Zhang, and T. Hain, "Using neural network front-ends on far field multiple microphones based speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 5542–5546, May 2014.

[25] B. Kingsbury, T. N. Sainath, and H. Soltau, "Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization," in *INTERSPEECH*, 2012.