SPEAKER DIARISATION AND LONGITUDINAL LINKING IN MULTI-GENRE BROADCAST DATA

P. Karanasou, M. J. F. Gales, P. Lanchantin, X. Liu, Y. Qian, L. Wang, P. C. Woodland, C. Zhang

Department of Engineering, University of Cambridge

ABSTRACT

This paper presents a multi-stage speaker diarisation system with longitudinal linking developed on BBC multi-genre data for the 2015 Multi-Genre Broadcast (MGB) challenge. The basic speaker diarisation system draws on techniques from the Cambridge March 2005 system with a new deep neural network (DNN)-based speech/non speech segmenter. A newly developed linking stage is next added to the basic diarisation output aiming at the identification of speakers across multiple episodes of the same series. The longitudinal constraint imposes an incremental processing of the episodes, where speaker labels for each episode can be obtained using only material from the episode in question, and those broadcast earlier in time. The nature of the data as well as the longitudinal linking constraint position this diarisation task as a new open-research topic, and a particularly challenging one. Different linking clustering metrics are compared and the lowest within-episode and cross-episode DER scores are achieved on the MGB challenge evaluation set.

Index Terms— speaker diarisation, speaker segmentation, agglomerative clustering, longitudinal linking

1. INTRODUCTION

Speaker diarisation is the task of answering to the question "who spoke when", a definition introduced by the DARPA EARS programme for the US National Institute for Standards and Technology (NIST) Rich Transcription (RT) evaluation in 2003 [1]. Determining the speakers and the time periods when each speaker is active can be useful in several speech processing tasks, such as speaker indexing and retrieval, archiving and monitoring large audio sets, movie analysis and rich transcription (i.e. making transcriptions more "readable" by adding punctuation, speaker markers, etc.). In can also be helpful to automatic speech recognition (ASR) for segmentation of the audio to homogeneous blocks. Because of a growing interest towards such applications, diarisation has also received much attention lately with relevant evaluations and open-source toolkits becoming available on the web, such as [2], [3] and [4].

As the available speech corpora move towards more demanding conditions containing multi-genre data, multiple speakers, diverse acoustic conditions or multiple recordings, so does the speaker diarisation domain of application. A way to assess this progress is through evaluations, which became an effective way of comparing existing algorithms and obtaining state-of-the-art systems in a particular domain. The first domain of application of such evaluation for speaker diarisation was Broadcast News (BN) in the NIST RT-03 and RT-04 evaluations [1, 5], followed by more recents evaluations on the more difficult task of meeting data [6] from 2005 to 2009. Similar evaluations have been done on French data including ESTER campaigns on BN [7, 8] and, more recently, the REPERE [9] and the ETAPE [10] corpus including TV shows of different genre, different levels of spontaneous speech and overlapping speech of multiple speakers. Each domain presents unique challenges, although some techniques tend to generalise over several domains [11, 12].

A new challenge is how to process data of very variable type of audio. Such multi-genre data can include documentaries, news, movie trailers, commercials, live sports, etc. In [13], a speaker diarisation system for French multi-genre web videos is presented. The Diarisation Error Rate (DER) for this task degrades severely and varies vastly between different genres (from 12.8% to 53.1%); note that the same system architecture on BN data gives DER $< 10\%$. These results indicate the difficulty of building a generalisable diarisation system on multi-genre data. This is because the audio to recognise may include broadcasts in diverse environments and drama with highly-emotional speech, overlaid background music or sound effects, and also because the different characteristics of each genre make tuning the system challenging. In this paper, a diarisation system is developed for the MGB challenge [14], one of the three official challenges in ASRU 2015. A corpus of BBC TV shows is used, including speech that is mostly British English with a range of regional accents and audio contents covering a broad range of genres, environments and speaking styles [15]. In particular, the genre type varies widely from news, debates and weather reports to drama series, soap operas, comedy shows, documentaries and live sports. See Section 5 for a more detailed description of the data used for training and evaluation. These characteristics of the data make diarisation in this domain a challenging task, not yet fully covered in the existing literature, and surely not yet solved. In Section 3 each stage of the diarisation system is described in detail. It draws on techniques used in [16], with a new Deep Neural Network (DNN)-based Speech/Non-Speech segmenter.

Following the specifications of the MGB challenge, on top of a basic speaker diarisation system, a diarisation system with longitudinal linking was developed. The linking is a newly added stage to the basic diarisation pipeline, described in Section 4. The linking requires the identification of speakers across multiple episodes of the same series. The longitudinal constraint requires the incremental processing of the episodes, where speaker labels for each episode can be obtained using only material from the episode in question, and those broadcast earlier in time. Diarisation with linking is a recently defined task aiming at applications where it may be useful to process a collection of episodes from the same source. This is a frequent situation for digital libraries and multimedia archives where it is likely that some speakers (journalists, actors, frequent guests...) will occur in several episodes. In such cases, having the same speaker associated with the same identifier across all the episodes could be very

This work is in part supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology). C. Zhang is also supported by a Cambridge International Scholarship from the Cambridge Commonwealth, European & International Trust. Supporting data for this paper will be available at the http://www.repository.cam.ac.uk data repository.

convenient for assessing the available audio information. For our incremental linking system, different clustering criteria are compared with and without retraining of the speaker models in an agglomerative clustering framework. In particular, the cross likelihood ratio (CLR) is used to compute the distance between clusters and the cluster merging is done either based on this measure or on the dissimilarity distance measures complete-linkage clustering and single-linkage clustering. The results of the diarisation system with and without linking are presented in Section 6 and are very competitive for the particular multi-genre domain.

A quick overview of the rest of the paper is as follows. Section 2 presents the existing litterature on the diarisation with linking task. Then, the different stages of our basic diarisation system are described in Section 3, with the linking stage separately discussed in more details in Section 4. The experimental setup is given in Section 5 including corpus description, system specifications and definition of evaluation measures. Finally the results of diarisation with and without linking are reported in Section 6 and the conclusions of the current work are summarised in Section 7.

2. BACKGROUND WORK ON DIARISATION WITH LINKING

In the literature, there are different ways of defining and approaching the diarisation with linking task. Linking can be integrated into the main diarisation procedure, or it can be seen as an independent step applied to the output of the diarisation system. Integrating the linking to the main diarisation procedure can be done by simply concatenating all the recordings of the data set and then running a standard speaker diarisation system. This is proposed in [17] and in [18] with a comparison of different linking architectures, some in batch mode and some in incremental mode with different permutations. The experiments were conducted on episodes of the English radio show "The Naked Scientists". The incremental mode increases the DER compared with the batch mode and it is shown that it is sensitive to the order of processing of the episodes. Because of propagation of errors, the best results were achieved when the episodes were sorted by increasing DER.

However, such approaches are not practical or even feasible for larger volumes of data. That is why a two-stage approach has been proposed where the linking is a complementary stage applied on the output of the baseline diarisation system. The number of speaker clusters generated by the main diarisation system is already reduced, and thus the linking clustering can be applied on them more efficiently. This is the so called "speaker attribution" approach in [19], where the proposed complete-linkage clustering is evaluated in terms of cluster purity and cluster coverage to compare Maximum A Posteriori (MAP) and joint factor analysis (JFA) speaker models. Focusing on the linking stage, the complete-linkage clustering is further evaluated in [20] using reference speaker labels as output from the diarisation stage and it is shown to outperform CLR. The entire two-stage system of diarisation and linking is finally evaluated on two-speaker telephone conversations in [21]. That work is the closest to ours concerning the linking clustering criterion as will be explained in Section 4. However, their linking is applied in a batch mode and not incrementally as in our case, and their data corpus is significantly simpler.

Another multi-stage diarisation with linking system was presented in [22]. This work was tested on meeting data and compared different cluster dissimilarity distances for the linking agglomerative clustering stage. The best results were achieved with a two-way Hotteling t-square statistic as the dissimilarity measure between two clusters, with the cross-recording DER approaching the within-recording DER. A more efficient method of linking was proposed in [23] by segmenting long recordings into smaller chunks, which deteriorates the DER but scales better to large data sets. In a more recent work presented in [24], diarisation with linking was applied on a variety of English TV shows. In that work, further gains were achieved by imposing the constraint that speakers locally hypothesised to be distinct must not be assigned to the same cluster during the linking step. Finally, the normalised CLR (NCLR) clustering for linking was compared with an i-vector based approach in [25]. Comparable performance on the ESTER 2 corpus was achieved.

3. DIARISATION SYSTEM

3.1. Speech/Non speech detection

The first stage of the diarisation pipeline is the speech detection (SD) stage. The multi-genre broadcast data exhibit a wide range of environmental conditions, with various sorts of non-speech in it, such as music, applause, laughter, transportation sounds, etc. Thus, a speech/non-speech segmentation front-end has a major impact on automatic segmentation for such data, and is crucial for the following diarisation processing. Since DNN models have been proven to have high accuracy in classifying speech frames, we explore their use in speech/non-speech segmentation of the multi-genre broadcast data. A DNN binary classifier was built to partition the speech signal into regions of speech and non-speech. The DNN classifier is trained in a DNN-HMM hybrid configuration with cross-entropy as the objective function in the back-propagation optimisation. Two softmax units are used in the output layer corresponding to speech and non-speech. Posterior probabilities are estimated by the DNN and converted to log-likelihood. Next, frame-wise decisions are made by Viterbi decoding using 2-state HMMs (speech/non-speech). Further details on the building of the DNN segmenter are given in "System Specifications" Section 5.2.

3.2. Speaker Change Point Detection and Speaker clustering to homogeneous blocks

The next stages are the change point detector (CPD) and an iterative agglomerative clustering (IAC) stage following the best configurations presented in [16]. The CPD stage finds potential changes in audio characteristics within each segment using the symmetric divergence (KL2) distance metric between two adjacent sliding windows of 2 seconds length. The CPD algorithm used finds local maxima in the divergence distance metric between the sliding windows. A left to right search of these peaks is then made removing the smaller of any pairs of neighbouring peaks which occurs within a specified minimum duration. In [16], it was shown that enforcing a minimum length constraint of 1 second on the resulting segments reduces the segment impurity. A full covariance Gaussian is used for each window and the distance threshold is chosen to over-segment the data. At this point, silence portions larger than a tunable threshold are discarded and portions of speech between these silences form the new segments. This internal silence threshold was seen to significantly affect the missed and false alarm speaker rates as will be further shown in Section 6[1].

[1]Note that all the missed and false alarm rates presented in the paper are computed over the total speaker time, following the specifications of the NIST script used for the diarisation evaluation. This distinguishes them from the usual missed and false alarm speech rates used for ASR which are computed over the total speech time.

Then, an iterative agglomerative clustering (IAC) scheme similar to [26] is applied. A single Gaussian model is built for each segment and the likelihood change for each potential merge of segments is calculated. The merge with the smallest likelihood loss is performed and the statistics are recalculated. This is repeated until the potential likelihood loss on merging reaches a certain threshold. These new models are then used to resegment the data using a Viterbi decode. This whole process is repeated until the segmentation converges or a maximum number of iterations are reached. A Bayesian information criterion (BIC) criterion [27] was used for both the stopping decision and the ordering of merges, and it updates the statistics assuming that the data in the cluster has been concatenated. The minimum length constraint on the CPD decreased the number of small segments coming from the CPD stage.This allowed full covariance models to be used throughout the IAC stage. Finally, silence padding of 20 frames is added in the beginning and end of each segment to avoid segment overlaps.

At this point, a segmentation of the data into homogeneous blocks is achieved. The chosen settings heavily under-cluster the data during the IAC stage, but provide a reasonable starting point for the following SID clustering stage.

3.3. SID clustering

An additional agglomerative clustering stage is incorporated which employs speaker identification (SID) techniques as in [28]. In this case, a Universal Background Model (UBM) is first trained with a large amount of audio data. Then, speaker models are derived by adapting the UBM model's parameters with speaker speech data. In particular, a maximum A Posteriori (MAP) adaptation (mean-only) is applied towards each speaker cluster following the variable-prior MAP with multiple iterations as described in [16]. Feature warping as described in [29] using a sliding window of 3 seconds is applied to help reduce the effect of the acoustic environment. The cross likelihood ratio (CLR), first defined in [30], is computed between any two given clusters,

$$\mathrm{CLR}(c_i, c_j) = \frac{1}{N_i} \log \frac{L(x_i|\lambda_j)}{L(x_i|\lambda_{\mathrm{UBM}})} + \frac{1}{N_j} \log \frac{L(x_j|\lambda_i)}{L(x_j|\lambda_{\mathrm{UBM}})}$$

where $L(x_i|\lambda_j)$ is the average likelihood per frame of data x_i given the model λ_j . The pair of clusters with the highest CLR is merged and a new model is created using all the data in the new cluster. The process is repeated until the highest CLR is below a predefined threshold, θ_{CLR} .

4. SPEAKER LINKING

The speaker linking in this paper is a new independent stage applied to the output of the basic diarisation system. The aim of this stage is to cluster together speakers that are the same across different episodes of the same series. Two different architectures are implemented in this paper. The first one concatenates all episodes of each series together and, then, applies a new clustering in a batch mode. In the second case, a causal relationship is imposed between the episodes of each series and the linking is done in incremental mode; speaker labels for each episode can be obtained using only material from the episode in question, and those broadcast earlier in time. This longitudinal linking respects the constraints imposed by the MGB challenge and will be the final result presented in this work.

For both the batch and incremental mode in linking clustering, different similarity metrics between speaker clusters are compared. First, a clustering similar to the one described in Section 3.3 is applied, this time on the output of the basic diarisation system, with CLR being the similarity metric between clusters. The same approach is also tried but without the retraining of new speaker models after each merge. As argued in [31], the retraining phase may be viewed as a 'hard' decision that is not desirable when conducting clustering. This is because an erroneous clustering decision will lead to incorrect new speaker models that are then carried through to the subsequent stages of clustering and propagate the errors through the entire linking procedure.

The other clustering techniques that we compared are two hierarchical clustering methods, complete-linkage clustering and single-linkage clustering. An overview of such methods can be found in [32]. In single-linkage clustering , the similarity of two clusters is the similarity of their most similar members. This single-linkage merge criterion is local, paying attention solely to the area where the two clusters come closest to each other. Other, more distant parts of the cluster and the clusters' overall structure are not taken into account. This leads to a "chaining-effect" resulting in elongated clusters. In complete-linkage clustering on the other hand, the similarity of two clusters is the similarity of their most dissimilar members. This complete-link merge criterion is non-local; the entire structure of the clustering can influence merge decisions. This results in a preference for compact clusters with small diameters over long clusters, but also causes sensitivity to outliers. A single speaker far from the center can increase diameters of candidate merge clusters dramatically and completely change the final clustering.

The cluster distances in complete and single-linkage clustering are computed based on CLR but transformed to a dissimilarity distance measure:

$$\mathbf{L}_{ij} = \begin{cases} e^{(-\mathbf{A}_{ij})}, & i \neq j \\ 0, & i = j \end{cases}$$

where $\mathbf{A}_{ij}$ are the elements of the upper triangular matrix with the CLR scores between each cluster pair. In the case of hierarchical clustering methods, no retraining of the speaker models is involved. In all cases, the same CLR threshold θ_{CLR} as in the previous SID clustering stage is used as the stopping criterion for the clustering. The disadvantage of all these methods, however, is that they do not scale well; they have a time complexity of $O(n^2)$, where n is the number of initial speaker clusters.

5. EXPERIMENTAL SETUP

In this section, a description of the training and evaluation data sets will be given, as well as some system specifications for the different parts of the diarisation pipeline where needed. A short description of the used evaluation measures is included.

5.1. Corpus description

Training data from 7 weeks of BBC output across all TV channels were made available for the MGB challenge [14]. The provided audio material covers different genres (advice, children shows, comedy, competition, documentary, drama, events and news) and a broad range of environments and speaking styles. The training set includes 493 unique shows and 2193 episodes broadcast in 2008 on 4 BBC TV channels from April 1st through May 19th representing 1580 hours of raw audio data. The duration of the programmes ranges from 2.3 minutes to 6.4 hours. Associated transcriptions were prepared from subtitles for the hearing impaired which were re-aligned with the audio using a lightly supervised approach [33].

Longitudinal development and evaluation sets were hand-transcribed. Transcriptions include start and end time-stamps for each segment as well as the speaker name. The longitudinal development set includes 19 episodes from 5 unique shows with different genres. They were broadcast in 2008, from May 28 through July 27th, representing 12 hours of audio data. Each show include recurring characters. The longitudinal evaluation set includes 19 episodes from 2 shows, a cooking show and a culture news show. They were broadcast in 2008 from June 6th through July 25th, representing 14 hours of raw audio data. This paper is primarily evaluated on the development set. However, the results on the evaluation set of our submitted system to the MGB challenge are also reported.

5.2. System specifications

For the DNN SD stage, the DNN structure used takes as input 40-dimensional filter bank features in a window of 55 frames. It has an input layer with 2200 units, 6 hidden layers, the first with 1000 units and the rest with 200 units, while the output layer has two units (speech/non speech). Performance with different input context windows was investigated as well as different sizes for the first hidden layer to make it more expressive for the very long context windows. The currently employed configuration gave the highest classification accuracy (see [34]).

Different DNN SD segmenters were built based on different selections on the training data. After the state-level alignment on the training set of MGB challenge, audio data of states other than silence and short pause inside a segment is used as 'speech' data to train DNN segmenters. As for the 'non-speech' data in training, two schemes are considered. One is to use only audio data of silence and short pause states inside a segment as 'non-speech' data, and the other is to include the inter-segment non-speech data as well.

The so called "DNN-v1" segmenter was trained using an 100h subset of the original training set, with Average Word Duration (AWD) less than 0.7s and Word Matched Error Rate (WMER) less than 25.0% from lightly supervised alignment. Only intra-segment silence and short pause were used as non-speech data (38h). This segmenter was applied on the whole training set, which then went through another round of lightly supervised decoding (see [33] for more details on the alignment). Then, with this refined alignment, two other segmenters "DNN-v3" and "DNN-v4" were trained using a 209h subset, with AWD between 0.165s and 0.66s and Phone Matched Error Rate (PMER) equal to 0.0 from lightly supervised alignment. "DNN-v3" employed the first scheme of non-speech data keeping only intra-segment silence and short pause non-speech data (37h), while "DNN-v4" employed the second scheme of non-speech data (313h), including inter-segment non-speech data from the 209h subset and adding it to the intra-segment non-speech data from the whole training corpus.

The UBM used for the SID clustering step was built using an 100h subset of the training set. This is the same subset used for the "DNN-v1" segmenter described earlier in the section. A speaker-independent (SI) UBM GMM model with 1024 mixture components was trained on this set. Each component corresponded to a 24-dimensional front-end feature vector, consisting of 12 PLP coefficients appended with the delta coefficients.

5.3. Evaluation measures

The performance of our systems is evaluated in terms of DER for the diarisation system. As explained in [35], a system hypothesises a set of speaker segments each of which consists of a (relative) speaker ID label and the corresponding start and end times. This is scored against reference "ground-truth" speaker segmentation. Since the hypothesis speaker labels are relative, they must be matched appropriately to the true speaker names in the reference. To accomplish this, a one-to-one mapping of the reference speaker IDs to the hypothesis speaker IDs is performed so as to maximise the total overlap of the reference and (corresponding) mapped hypothesis speakers. Speaker diarisation performance is then expressed in terms of the miss (MS: speaker in reference but not in hypothesis), false alarm (FA: speaker in hypothesis but not in reference), and speaker-error (SpkE: mapped reference speaker is not the same as the hypothesised speaker) rates. The overall DER is the sum of these three components[2]. To assess the impact of speaker linking on the diarisation systems, a within-episode and a cross-episode DER are computed. The within-episode DER is the standard DER used for a diarisation system without linking. To compute the cross-episode DER, the references of all episodes of a series are concatenated and the within-episode speaker identifiers are replaced by unique speaker identifiers across the series. Then, the same algorithm to compute the DER is performed considering now each series as a single episode.

6. RESULTS

In this section, the results on the MGB longitudinal development set (MGB Long dev set) are presented, as well as the results on the MGB longitudinal evaluation set of our primary system submitted to MGB challenge. First, the three segmenters presented in Section 5.2, followed by the next diarisation stages were used in the basic diarisation pipeline and tested on MGB Long dev set. Table 1 presents the DER scores for different system configurations. The column "Insil" presents the internal silence threshold in number of frames; silence portions longer than this threshold are discarded and the current segment is cut into two new segments. The three components that constitute the DER are also presented: Missed Speaker Rate (MS), FA Speaker Rate (FA) and Speaker Error (SpkE). All presented results are after feature warping is applied (see Section 3.3) which improved the DER score by about 17% relative. From this table, it can be seen that the lowest DER score is achieved when using the "DNN-v4" segmenter with internal silence no longer than 30 frames. Also note that the threshold for the SID clustering for all the presented results was experimentally set to $\theta_{CLR} = 0.2$.

S/NSseg	InSil	nSeg	DER			
			MS	FA	SpkE	Total
DNN-v1	50	7826	6.4	1.7	30.4	38.48
	40	8714	6.8	1.5	-	-
	30	9977	7.3	1.4	33.9	42.60
DNN-v4	50	7307	5.5	1.0	34.4	40.87
	40	8129	5.8	1.0	32.5	39.26
	30	9150	6.1	0.9	30.6	**37.48**
DNN-v3	50	7829	4.7	2.8	31.8	39.38
	40	8740	5.1	2.5	35.4	43.05
	30	9972	5.5	2.1	33.0	40.61

Table 1. DER(%) scores on MGB Long dev set

Another observation from Table 1 is that the internal silence threshold is a parameter that significantly influences the result of

[2]A complete description of the evaluation measure and scoring software implementing it can be found at http://nist.gov/speech/tests/rt/rt2004/fall

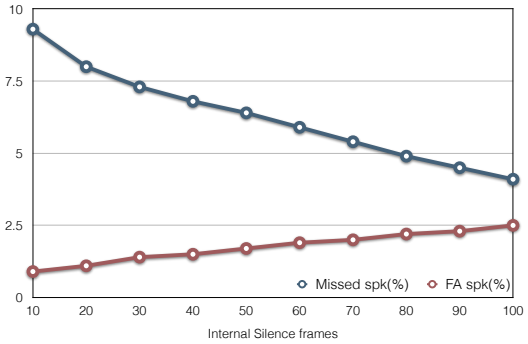

Fig. 1. Miss and FA speaker rates for different internal silence thresholds on the "DNN-v1" segmenter

the segmentation of the audio to homogeneous blocks and, thus, the following clustering stage of the diarisation pipeline. The lower the internal silence threshold, the greater the number of segments and, consequently, the shorter the average segment duration. Ideally, having more segments should avoid propagation of clustering errors from the first diarisation stages and retain the possibility to cluster them correctly in the next diarisation steps. We tested the "DNN-v1" segmenter for values of internal silence threshold varying from 10 to 100 frames with a step of 10 frames, as can be seen in Figure 1. The more segments we have (i.e. the lower the silence threshold), the lower the FA speaker rate, but the higher the Miss speaker rate. The operation points that were found to work better for the diarisation task were the ones with internal silence thresholds 30 to 50 which were, thus, chosen to be presented in Table 1.

Trying to further analyse the presented results, for each DNN segmenter we chose the configuration giving the lowest DER and we present it in Table 2 along with the Ideal DER score, as well as the Word Error Error (WER) when using the segmenters under question to an automatic speech recognition (ASR) system. It is interesting to observe that the segmenter that gives the lowest DER ("DNN-v4" with 30 frames internal silence threshold) also gives the lowest Ideal DER and WER scores. This warrants some further analysis. Concerning the WER scores, the ASR system used is an HTK speaker independent sequence-trained DNN hybrid system trained on the original MGB training set with a 64k 4-gram language model run with fast decoding settings and confusion network output [34].

The three compared DNN segmenters in the lower block of Table 2 produce different WER in spite of presenting similar Miss and FA Speaker rates. This is due to a different partition of the audio signal, reinforced by a different detection of points of change of audio characteristics in the signal by the CPD stage and different clustering to homogeneous blocks by the IAC stage. The difference in homogeneity of the output audio blocks of the first stages of the diarisation pipeline is also, and most clearly, shown by the differences in Ideal DER. Substracting the Miss and False Alarm Speaker rates from the Ideal DER gives us the Ideal Speaker Error which is a direct measure of the homogeneity of the generated segments. If the segments were perfectly homogeneous, they should be able to be clustered with Ideal Speaker error equal to 0.0. This is not the case, but the system with the lowest DER has also the lowest Ideal DER and Ideal Speaker Error. Note, finally, that there is a correlation between the Ideal DER and the number of segments; the more segments we have, the lowest the Ideal DER is. This is because more and smaller segments are more probable to be homogeneous.

In Table 2, the DNN segmenters are also compared with the

S/NSseg	InSil	nSeg	MS+FA	DER		WER
				Actual	Ideal	
MGBbase	-	13859	12.7	46.78	-	36.6
CamRT-04	-	6280	18.2	-	-	34.3
DNN-v1	50	7826	8.1	38.48	11.92	27.8
DNN-v4	**30**	9150	7.0	**37.48**	**9.90**	**27.0**
DNN-v3	50	7829	7.5	39.38	11.81	27.6

Table 2. Comparison of different segmenters on MGB Long dev set

Cambridge RT-04 segmenter ("CamRT-04") segmenter and the baseline segmenter provided by the MGB challenge ("MGBbase"). The Cambridge RT-04 segmenter was originally tuned for US English Broadcast News (BN) and used in HTK 2003 BN system [36]. This segmenter was aimed for ASR and that is why it was not further tested for the diarisation task. It reveals, however, the difficulty of the segmentation task on the multi-genre data. This is also supported by the WER and DER scores of the baseline MGB segmenter and diarisation system ("MGBbase") which are significantly worse than our systems.

The muti-genre nature of the presented diarisation task makes it worthwhile having a more detailed analysis of the resulting DER. The DER scores per genre for the "DNN-v4" system with 30 frames internal silence threshold are reported in Table 3. Each series of the dev set correspond to a different genre. It can be seen that the DER varies from 13% to 78%. These results reveal more clearly the difficulty of the data that we are handling in this paper.

series	DER			
	MS	FA	SpkE	Total
sci-fi tv-drama	12.7	1.1	64.38	78.16
sitcom	8.2	1.1	51.90	61.15
documentary	1.9	0.2	10.82	12.90
tv-drama	6.4	1.0	16.27	23.70
sports	5.7	1.6	39.85	47.13
Overall	6.1	0.9	30.6	37.48

Table 3. DER scores per genre on MGB Long dev set for the "DNN-v4" system

Last but not least, the results of diarisation with longitudinal linking are presented in Table 4 on our best diarisation system "DNN-v4" from Table 1 (with 30 frames internal silence threshold). In this table, both the within-episode (NoLinkDER) and the cross-episode DER (LinkDER) are presented. The first row corresponds to the basic diarisation system before applying the linking stage. The rest of the table reports the results of linking in an incremental mode for the different clustering criteria presented in Section 4. The first system uses CLR for clustering with retraining of the merged speaker models ("CLR"), while the retraining stage is removed in the second system ("CLR-noR"). Comparing these two rows, it is observed that removing the retraining step improves the DER scores both in linking and non-linking scoring mode. Then, longitudinal linking with complete-linkage clustering ("CLC") and single-linkage clustering ("SLC") are performed. "SLC" gives the higher DER, probably because of the chaining effect explained in Section 4. On the other hand, "CLC" gives the lowest within-episode and cross-episode DER compared with the other systems with longitudinal linking. This is our primary system submitted to

the MGB challenge.

Link	nSpk		DER	
	NoLink	Link	NoLink	Link
-	640	640	37.48	-
CLR	487	389	39.20	44.35
CLR-noR	533	426	38.91	43.85
CLC	599	473	37.89	**42.72**
SLC	455	378	46.42	51.03

Table 4. Longitudinal linking: DER scores within-episode (NoLinkDER) and cross-episode (LinkDER) on MGB Long dev set for the "DNN-v4" system

The next table (Table 5) presents a comparison of the diarisation systems with linking in batch ("Batch") and in incremental mode ("Longitudinal"). The CLC is chosen as the linking clustering metric as it was shown to outperform the other metrics for the longitudinal linking in Table 4. It also outperforms the other metrics for linking in batch mode, although these experiments are not further presented in this paper as the main focus is the linking with the longitudinal constraint, respecting the MGB challenge specifications. Linking in incremental mode is expected to deteriorate the DER compared to the linking in batch mode because of plausible propagation of clustering errors across episodes. This is indeed observed in Table 5, although the difference between the two modes is only 0.02% relative.

CLC Link	nSpk		DER	
	NoLink	Link	NoLink	Link
Longitudinal	599	473	37.89	**42.72**
Batch	636	532	37.55	42.07

Table 5. Longitudinal vs batch linking: DER scores within-episode (NoLinkDER) and cross-episode (LinkDER) on MGB Long dev set for the "DNN-v4" system

Table 6 briefly presents the results of our primary system ("DNN-v4" with CLC linking) submitted to MGB challenge and tested on the MGB longitudinal evaluation set. The DER scores on both series of the evaluation set are presented. It should be noted that these series are of different genre than the ones included on the development set used to tune the system. They also include more episodes (11 and 8 episodes while in the development set the maximum number of episodes per series was 6). This is an extra difficulty for the longitudinal linking because the more episodes we have, the bigger the risk of clustering error propagation from one to the other. The linking indeed degrades the DER by about 7% absolute (compared to 5% absolute in the dev set). This is consistent for both series and for the overall cross-episode DER score (LinkDER). The overall scores are the lowest achieved among the participants of the MGB challenge, both in terms of within-episode and cross-episode DER.

A final observation can be made on the number of speakers identified by the systems in Table 4. Our primary system ("DNN-v4" with CLC linking) results in more speakers (473 speakers) than the reference (351 speakers) even after the linking stage (see Column "LinkSpk"). As mentioned earlier, the stopping criterion for the linking clustering was $\theta_{CLR} = 0.2$, kept the same as for the SID clustering. It was found, though, that changing this value to $\theta_{CLR} = 0.15$ does not influence the DER scores, but drops the speaker number to 384, which is very close to the real number of reference speakers.

Series	NoLinkDER	LinkDER
competitive cuisine show	44.59	51.92
culture show	33.07	40.21
Overall	40.2	47.46

Table 6. DER scores within-episode and cross-episode for the system "DNN-v4" with CLC linking on MGB Long eval set

7. CONCLUSIONS

This paper presented a diarisation system augmented with a longitudinal linking stage, and applied to the difficult domain of multi-genre data. Each part of the system was detailed with a focus on a new DNN-based speech segmenter and on the longitudinal linking stage. Different DNN segmenters were compared and the best operation points were found. A further analysis of the homogeneity of the segments produced by each segmenter was also attempted. For the longitudinal linking, different clustering metrics were presented with complete-linkage clustering outperforming the rest. Results were reported for the basic diarisation system and for the diarisation with linking on the MGB longitudinal development set. The best of these systems with longitudinal linking was our primary submission to the MGB challenge and achieved the lowest within-episode and cross-episode DERs on the MGB evaluation set.

8. REFERENCES

[1] NIST, "Fall 2003 rich transcription (RT-03F) evaluation plan," http://www.itl.nist.gov/iad/mig/tests/rt/2003-fall/docs/rt03-fall-eval-plan-v9.pdf, 2003.

[2] M. A. H. Huijbregts, *Segmentation, diarization and speech transcription: Surprise data unraveled*, Ph.D. thesis, University of Twente, 2008.

[3] S. Meignier and T. Merlin, "LIUM SpkDiarization: an open source toolkit for diarization," in *CMU SPUD Workshop*, 2010.

[4] Anthony Larcher, Jean-François Bonastre, Benoit G. B. Fauve, Kong-Aik Lee, Christophe Lévy, Haizhou Li, John S. D. Mason, and Jean-Yves Parfait, "ALIZE 3.0 - open source toolkit for state-of-the-art speaker recognition," in *Interspeech*, 2013, pp. 2768–2772.

[5] NIST, "Fall 2004 rich transcription (RT-04F) evaluation plan," http://www.nist.gov/speech/ tests/rt/rt2004/fall/docs/rt04f-eval-plan-v14.pdf, 2004.

[6] NIST, "Fall 2005 rich transcription (RT-05F) evaluation plan," http://www.itl.nist.gov/iad/mig/tests/rt/2005-spring/rt05s-meeting-eval-plan-V1.pdf, 2005.

[7] Sylvain Galliano, Edouard Geoffrois, Djamel Mostefa, Khalid Choukri, Jean-François Bonastre, and Guillaume Gravier, "The ESTER phase II evaluation campaign for the rich transcription of French broadcast news," in *Interspeech*, 2005, pp. 1149–1152.

[8] Sylvain Galliano, Guillaume Gravier, and Laura Chaubard, "The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts," in *Interspeech*, 2009.

[9] Aude Giraudel, Matthieu Carr, Valrie Mapelli, Juliette Kahn, Olivier Galibert, and Ludovic Quintard, "The REPERE corpus : a multimodal corpus for person recognition," in *LREC*, 2012.

[10] Guillaume Gravier, Gilles Adda, Niklas Paulsson, Matthieu Carr, Aude Giraudel, and Olivier Galibert, "The ETAPE corpus for the evaluation of speech-based TV content processing in the French language," in *LREC*, 2012.

[11] Y. Moh, P. Nguyen, and J.-C. Junqua, "Towards domain independent speaker clustering," in *ICASSP*, 2003, vol. 2, pp. II–85–8.

[12] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, O. Friedland, and O. Vinyals, "Speaker diarization : A review of recent research," *"IEEE Transactions On Audio, Speech, and Language Processing" (TASLP), special issue on "New Frontiers in Rich Transcription"*, vol. 20, no. 2, 2012.

[13] Pierre Clément, Thierry Bazillon, and Corinne Fredouille, "Speaker diarization of heterogeneous web video files: A preliminary study," in *ICASSP*, 2011, pp. 4432–4435.

[14] P. Bell, M.-J.-F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester, and P.-C. Woodland, "The MGB challenge: Evaluating multi-genre broadcast media transcription," in *ASRU*, 2015.

[15] P. Lanchantin, P.-J. Bell, M.-J.-F. Gales, T. Hain, X. Liu, Y. Long, J. Quinnell, S. Renals, O. Saz, M.-S. Seigel, P Swietojanski, and P.-C. Woodland, "Automatic transcription of multi-genre media archives," in *SLAM Workshop*, 2013.

[16] R. Sinha, S. E. Tranter, M. J. F. Gales, and P. C. Woodland, "The Cambridge University March 2005 speaker diarisation system," in *Interspeech*, 2005.

[17] Viet-Anh Tran, Viet Bac Le, Claude Barras, and Lori Lamel, "Comparing multi-stage approaches for cross-show speaker diarization," in *Interspeech*, 2011, pp. 1053–1056.

[18] Qian Yang, Qin Jin, and Tanja Schultz, "Investigation of cross-show speaker diarization," in *Interspeech*, 2011, pp. 2925–2928.

[19] Houman Ghaemmaghami, David Dean, Robbie Vogt, and Sridha Sridharan, "Extending the task of diarization to speaker attribution," in *Interspeech*, 2011.

[20] Houman Ghaemmaghami, David Dean, and Sridha Sridharan, "Speaker linking using complete-linkage clustering," in *SST*, 2012.

[21] Houman Ghaemmaghami, David Dean, Robbie Vogt, and Sridha Sridharan, "Speaker attribution of multiple telephone conversations using a complete-linkage clustering approach," in *ICASSP*, 2012, pp. 4185–4188.

[22] Marc Ferras and Herve Bourlard, "Speaker diarization and linking of large corpora," in *SLT*, 2012, pp. 280–285.

[23] Marijn Huijbregts and David A. van Leeuwen, "Large-scale speaker diarization for long recordings and small collections.," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 20, no. 2, pp. 404–413, 2012.

[24] Xavier Bost and Georges Linarès, "Constrained speaker diarization of TV series based on visual patterns," in *SLT*, 2014, pp. 390–395.

[25] Grgor Dupuy, Mickael Rouvier, Sylvain Meignier, and Yannick Estve, "I-vectors and ILP clustering adapted to cross-show speaker diarization," in *Interspeech*, 2012, pp. 2174–2177.

[26] Jean luc Gauvain, Lori Lamel, and Gilles Adda, "Partitioning and transcription of broadcast news data," in *ICSLP*, 1998, pp. 1335–1338.

[27] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 127–132.

[28] Claude Barras, Xuan Zhu, Sylvain Meignier, and Jean-Lluc Gauvain, "Improving speaker diarization," in *Fall 2004 Rich Transcription Workshop(RT-04)*, 2004.

[29] Jason Pelecanos and Sridha Sridharan, "Feature warping for robust speaker verification," in *2001: A Speaker Odyssey - The Speaker Recognition Workshop*, 2001, pp. 213–218.

[30] Douglas A. Reynolds, Elliot Singer, Beth A. Carlson, Gerald C. O'Leary, Jack McLaughlin, and Marc A. Zissman, "Blind clustering of speech utterances based on speaker and language characteristics.," in *ICSLP*, 1998.

[31] David A. van Leeuwen, "Speaker linking in large data sets," in *Odyssey 2010: The Speaker and Language Recognition Workshop*, 2010.

[32] William H. Day and Herbert Edelsbrunner, "Efficient algorithms for agglomerative hierarchical clustering methods," *Journal of Classification*, vol. 1, no. 1, pp. 7–24, 1984.

[33] P. Lanchantin, M.J.F. Gales, P. Karanasou, X. Liu, Y. Qian, L. Wang, P.C. Woodland, and C. Zhang, "The development of the Cambridge University alignment systems for the Multi-Genre Broadcast challenge," in *ASRU*, 2015.

[34] P.C. Woodland, X. Liu, Y. Qian, C. Zhang, M.J.F. Gales, P. Karanasou, P. Lanchantin, and L. Wang, "Cambridge University transcription systems for the Multi-Genre Broadcast challenge," in *ASRU*, 2015.

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

[36] M.J.F. Gales, Do Yeong Kim, P.C. Woodland, Ho Yin Chan, D. Mrva, R. Sinha, and S.E. Tranter, "Progress in the cu-htk broadcast news transcription system," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1513–1525, 2006.