THE DEVELOPMENT OF THE CAMBRIDGE UNIVERSITY ALIGNMENT SYSTEMS FOR THE MULTI-GENRE BROADCAST CHALLENGE

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

Cambridge University Engineering Department, Cambridge CB2 1PZ, UK

{pkl27,mjfg,pk407,xl207,yq236,lw519,pcw,cz277}@eng.cam.ac.uk

ABSTRACT

We describe the alignment systems developed both for the preparation of data for the Multi-Genre Broadcast (MGB) challenge and for our participation in the transcription and alignment tasks. Captions of varying quality are aligned with the audio of TV shows that range from few minutes long to more than six hours. Lightly supervised decoding is performed on the audio and the output text is aligned with the original text transcript. Reliable split points are found and the resulting text chunks are force-aligned with the corresponding audio segments. Confidence scores are associated with the aligned data. Multiple refinements - including audio segmentation based on deep neural networks (DNNs) and the use of DNN-based acoustic models - were used to improve the performance. The final MGB alignment system had the highest F-measure value on the evaluation data.

Index Terms— Alignment, Lightly Supervised Training, Multi-genre Broadcast transcription

1. INTRODUCTION

The quantity of multimedia material available on the web has been increasing tremendously over the recent years giving access to a large amount of content. Moreover, the British Broadcasting Corporation (BBC) has a stated aim to open its broadcast archive to the public by 2022. Indexing of such material would give access to historic content, enabling search based on transcriptions, speaker identity and other extracted metadata. The automatic transcription of such material which includes broadcasts in diverse environments, drama with highly-emotional speech, overlaid background music or sound effects remains a challenging task and technologies are still underdeveloped. Recent work has included automatic transcription of podcasts and other web audio [1], automatic transcription of Youtube [2, 3], the MediaEval rich speech retrieval evaluation which used blip.tv semi-professional user created content [4], the automatic tagging of a large radio archive [5] and automatic transcription of multi-genre media archive data [6].

In the scope of the Natural Speech Technology (NST) EPSRC project and in collaboration with BBC Research and Development, we co-organised the Multi-Genre Broadcast (MGB) challenge. This challenge, presented in [7], is an evaluation of speech recognition, lightly supervised alignment and speaker diarization with longitudinal linking using TV recordings from the BBC. Broadcast audio extracted from 7 weeks of BBC output along with associated closed captions (subtitles) was provided to the challenge participants. The original BBC subtitles may be approximate for various reasons, including the caption production process. To facilitate participation in the challenge, we provided a refined aligned version of the original subtitles, including corrected time stamps and confidence scores to spot areas were the transcripts are accurate in order to select data for acoustic model training. Alignment of complete shows, potentially more than 6 hours long, of multi-genre broadcast material can be a difficult task, especially for approximate transcripts. Standard Viterbi-based forced alignment may be inadequate for very long segments of broadcast multi-genre audio which can potentially be contaminated with noise, music and for which the transcript may not be sufficiently accurate. A common approach, based on a lightly supervised approach and first introduced in [8, 9], is to make use of a speech recognizer to produce a time-aligned text transcript of the recording which is then aligned with the original text transcript. Matching sequences of words are then considered as reliable anchor regions to reduce the length of the problem. The same process is done recursively until a forced alignment can be done for each speech segment. The approach was further refined in recent years [10–17]. Alignment is necessary for the creation of a usable corpus for model training in automatic speech recognition but is also useful for other tasks such as automatic subtitling [18, 19], speech synthesis from audiobooks [16, 20], language training [21, 22] or more generally audio/video indexing techniques applied in search engines [23, 24].

In this paper we describe the different alignment systems that were developed for the MGB challenge, on one hand for the refinement of the original transcripts provided to participants, and on the other for our participation in both the transcription and alignment tasks. Our approach is in the same line with the one presented in [9]: lightly supervised decoding is performed on the audio and the output text is aligned with the original text transcript. Split points are then found in matching text regions according to a set of rules and the resulting text chunks are force-aligned with the corresponding audio segments. Confidence scores are computed for selection of the training data. For the transcription task, we ran a second iteration of the alignment process using deep neural network (DNN) based acoustic models as well as a new DNN-based segmenter. For the alignment task, we modified our system to comply with the alignment task rules, used improved acoustic models and a segmenter trained on the new refined transcripts and finally applied different selection schemes in order to maximise the target F-measure value.

The rest of the paper is laid out as follows: Section 2 describes the alignment system developed for the preparation of data for the MGB challenge. The second iteration of alignment performed using an improved version of the system and different selection schemes is described in Section 3. Systems developed for the alignment task are detailed in Section 4 and Section 5 concludes.

This work is in part supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology). Chao Zhang is also supported by a Cambridge International Scholarship from the Cambridge Commonwealth, European & International Trust. Supporting data for this paper is available at the data repository at http://www.repository.cam.ac.uk/handle/1810/251278

2. MGB CHALLENGE DATA PREPARATION

2.1. Description of the data

The Multi-Genre Broadcast (MGB) challenge, described in detail in [7], is an evaluation of speech-to-text transcription, lightly supervised alignment and longitudinal speaker diarization and linking using TV recordings from the BBC. The speech data is very varied and multi-genre, spanning the whole range of BBC TV output. About 1,600 hours of broadcast audio were provided by the BBC, broadcast on 4 TV channels in 2008 with associated aligned captions divided into a training set and 2 development sets. The *training set* includes 2,193 episodes of 493 unique shows broadcast between April 1st 2008 and May 19th 2008 representing 1,580 hours of audio data with a duration of programmes ranging from 2.3 minutes to 6.4 hours. The *development set* includes 47 episodes of different shows broadcast between May 5th and May 11th 2008 representing 28 hours of audio data. Finally a *longitudinal development set* was provided for the longitudinal transcription and the longitudinal speaker diarization and linking tasks. It includes 19 episodes of 5 unique shows broadcast between May 28th and July 27th 2008 and contains 12 hours of raw audio data. Two evaluation sets were prepared including a *transcription evaluation set* of 16 episodes of different shows representing 11 hours of audio data, and a *longitudinal evaluation* set of 19 episodes of 5 unique shows representing 12 hours of audio. The development and evaluation sets were both hand-transcribed and the longitudinal evaluation and development sets include speaker IDs. Both refined alignment and hand-transcribed transcripts of the two development sets were provided to participants. Finally, all metadata were provided in an XML format designed for the challenge.

The audio material contained in these sets covers different genres: advice, childrens, comedy, competition, documentary, drama, events, news, with a broad range of environments and speaking styles. The original transcripts provided by the BBC were captions for the hearing impaired. They included text transcripts, different text "colours" used for caption display to indicate the different speakers over short periods of time, time stamps, as well as other metadata such as indications of music and sound effects, or indications of the way the text has been pronounced. The quality of the text transcripts varies considerably across genres and shows in terms of precision of the alignment and reliability, due to time-lags that can occur in captions, or to the caption creation process: they may be edited to enhance clarity, paraphrasing, and deletions where the speech is too fast. Transcripts hence need to be refined before being used as training material. To facilitate participation in the challenge, we provided a refined version of the original transcripts to challenge participants. The refinement process was based on an alignment of the original transcripts which is then enriched with confidence scores for selection purposes as described in the next section.

2.2. Alignment system

The goal of an audio alignment system is to find time-stamps for words in the audio transcripts. Our approach follows the same path as that initially proposed in [8, 9] and refined in [10–17, 20]. Text transcripts are first extracted from the BBC closed captions, normalised and tokenised. Each recording is segmented using an baseline *segmenter* described in [25]. This segmenter was initially trained and tuned for broadcast US news and was therefore not well suited for British multi-genre broadcast material. Better segmenters were trained on the resulting aligned data during the challenge and will be described in the following sections. The alignment is based on a *lightly supervised approach* [10] which mainly consists in biasing the recognizer's language model (LM) to the content of the transcript. We estimated one biased language model per week using the transcripts of each of the 7 weeks of data. Each biased LM was then interpolated with a generic LM estimated on a combination of 640M words from BBC captions + 10M words of transcripts for acoustic model training provided to the participants of the challenge, with a 0.9/0.1 interpolation weight ratio. The vocabulary was chosen to ensure coverage of words from the original transcripts. Each speech segment was then decoded using a two-pass recognition framework [25, 26] including speaker adaptation. The decoding used a tandem-SAT system trained on a 200 hour subset of the training dataset using the biased LM for that week's data. Since, the start and end of the supplied episodes are often untranscribed, the decoding was constrained to the region delimited by the first and last time-stamps as indicated in the original captions.

The resulting time-aligned text transcription is then aligned with the original text transcript. The aim is to associate time-stamps with the original text transcript and to partition the text and audio into smaller segments in order to reduce the alignment of a complete episode to the alignment of a set of small segments. The splitting must be performed in areas where we are highly confident that the original text transcript is correct and it should not occur in the middle of an utterance. On the one hand, matching sequences of words (also called *anchor* points [9]) between the decoding output and the original text transcript are a good indicator that the original transcript is correct. On the other hand, captions provide a segmentation of the text into chunks of consecutively spoken words delimited by line breaks. The latter can occur for display constraints but can also indicate the end of an utterance as line breaks are usually inserted before a long silence or when a speaker change occurs. In our approach, line breaks are used as potential split points and a split is performed if a line break is positioned before of after a matching words sequence of at least 3 consecutive words. Each segment of the obtained partition is then force-aligned and a new split is made if the silence duration after a line break exceeds 5s or if a line break occurs at the end of a segment defined by the automatic segmenter. Finally, each segment of the obtained partition is again force-aligned. In the distributed refined transcripts, consecutive segments spoken by a same speaker (having the same text "colour" in the original transcripts) were merged if their inter-silence duration was less or equal that 200ms and with a maximum duration of 30s per segment. Speaker IDs resulting from an automatic clustering [25] were also provided for each segment. The resulting refined transcripts will be referred as v1 in the following sections of the paper.

2.3. Data selection

Although most of the transcripts were aligned during the alignment process, some might differ significantly from the actual spoken words. As mentioned in the introduction, this can be due to various reasons, including the caption creation process (e.g real-time captioning [27]) but also the alignment process. It is then necessary to provide a way to identify areas where the alignment is not perfect for the selection of training material. Different confidence scores were computed for this purpose and provided to the participants in the MGB challenge.

A relatively small number of segments from the alignment process cancontain large portions of non-speech events. This can happen when no reliable split points are found around a long non-speech event (e.g music). The text transcript present around the event's region might then be wrongly aligned. This is particularly the case

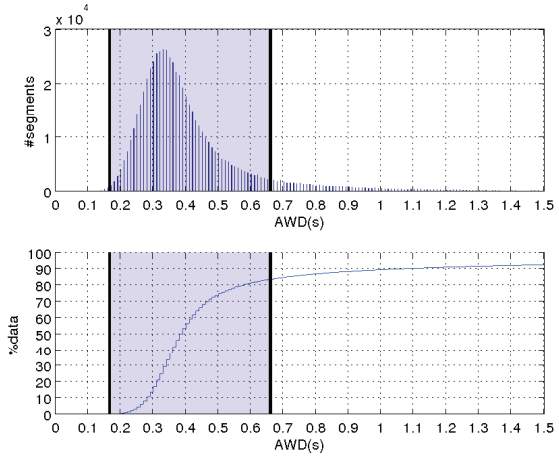

Fig. 1. Average word duration: *top*: segment distribution according to AWD value; *bottom*: cumulative duration in percentage of data selected from the training set according to a threshold on AWD. The shaded portion corresponds to the region (0.165s≤AWD≤0.66s)

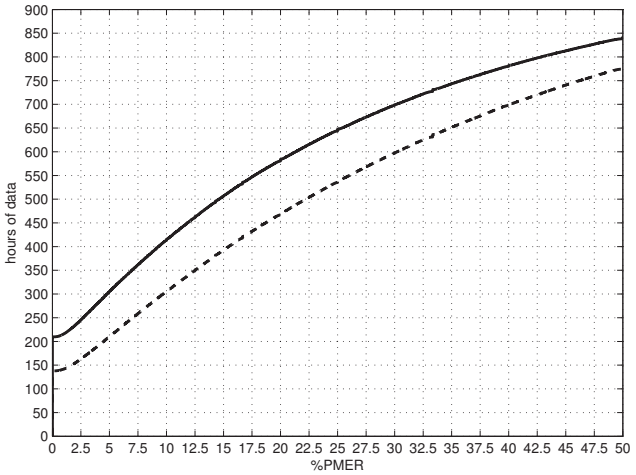

Fig. 2. Cumulative duration of the selected training data according to a threshold on PMER according to the v1 (- - -) and v2 (——) refined transcripts for 0.165s≤AWD≤0.66s. v1 refers to the transcripts provided to the participants and v2 refers to the transcripts refined for the transcription task.

for songs containing lyrics which are not transcribed in the captions. Those segments can be detected and rejected according to the *Average Word Duration* (AWD, in seconds), computed for each aligned segment. At the top of Figure 1 we present the segment distribution for the training set according to the AWD value. At the bottom of the same figure, we present the cumulative distribution of the data selected from the training set according to a threshold on the AWD. According to those plots, 0.165s≤AWD≤0.66s seemed a reasonable range for the AWD and was then adopted for most of our experiments. By doing so on the v1 refined transcription, we rejected 16.3% of the 1197 hours leading to 1001 hours of training data.

Two other confidence scores, the *Phone* and *Word Matched Error Rate* (PMER [28] and WMER), are used to assess the reliability of the transcripts. They are computed by scoring the lightly supervised decoding output of a segment against the corresponding aligned transcripts used as reference[1] which requires a further lightly supervised decoding using the segmentation derived from the refined alignment. If both text transcripts differ strongly, the MER value is high and the original transcript is therefore considered unreliable. This does not necessarily mean that the transcript is incorrect given that the difference could be due to the poor performance of the speech recogniser systems for the particular acoustic conditions. WMER and PMER can be used for data selection: in Figure 2, a dashed line shows the cumulative duration of the selected training data according to a threshold on the PMER for 0.165s≤AWD≤0.66s. This representation can be convenient for data selection. For instance, it can be seen that selecting segments having a PMER value less or equal than 40% leads to 700 hours of training data.

A way of assessing the reliability of the lightly supervised transcript is to use the estimates of the word posterior probabilities encoded in the confusion networks used for minimum word error rate decoding of the aligned segment with the biased LM. Given that these values tend to be over-estimates of the true posteriors, a decision tree is trained on a reference dataset to map the estimates to *confidence scores* [26]. The lightly supervised transcripts, provided to all the participants, can be directly used as training material [10] using the computed confidence scores to select suitable audio. Finally, word and segment level combination between the lightly supervised transcripts and the original text transcripts can be used to yield improved transcriptions [28].

3. SECOND ITERATION OF THE ALIGNMENT PROCESS

For our participation in the transcription task, a new iteration of the alignment process was performed on the MGB data released to participants, using a stronger bias for the language model (episode biased instead of week biased), an improved segmenter and better acoustic models trained on the v1 refined transcript. Those refinements are presented in this section.

3.1. System refinements

The alignment presented in section 2.2 was performed using a tandem-SAT system (T200) trained on a 200h subset of the training set[2]. A first refinement of the system was to use a speaker independent sequence-trained DNN hybrid system (H200.v1) based on a random selection of 200 hours with WMER≤25% according to the v1 refined transcript. More details about the model training can be found in the paper describing our participation in the MGB challenge transcription tasks [31]. Using the same biased LM, a reduction in WER of 5.9% absolute was obtained on the development set as shown in the Table 1.

We attempted to increase the bias of the language model by estimating an LM on the transcription for each episode separately instead of for the complete week as in section 2.2. The biased LM was also interpolated with a generic LM estimated on a combination of 640M words from the BBC subtitles and 10M words of acoustic transcripts, with a 0.9/0.1 interpolation weight ratio. In Table 1, it can be seen that by using episode biased instead of week biased LMs, an extra reduction in WER of 2.2% absolute was obtained on

[1]These are described as a matched error rate since there are no accurate transcripts to be used as reference.

[2]HTK V3.5 [29, 30] was used for building most of the models used in the paper.

system	LM	segmenter	%WER(del/ins)
T200	week biased	base-seg	35.0(16.9/4.5)
H200.v1	week biased	base-seg	29.1(16.4/3.3)
H200.v1	episode biased	base-seg	26.9(16.2/3.1)
H200.v1	episode biased	DNN-seg.v1	23.1(10.7/3.4)
H700.v1	episode biased	DNN-seg.v1	22.1(9.3/4.0)

Table 1. System comparison for the second iteration of the alignment. H200.v1 is a speaker independent sequence-trained hybrid system based on a selection of 200h, H700.v1 is a speaker independent sequence-trained hybrid system based on a 700h selection according to the v1 refined transcripts. The decoding used biased tri-gram LM followed by 4-gram expansion.

the development set. This shows that using a stronger bias toward the transcript can significantly improve the performance.

As mentioned in section 2.2, the baseline segmenter was not well suited for British multi-genre broadcast material. A DNN speech/non-speech segmenter was trained and used as a replacement. This segmenter, presented in detail in [31, 32], was trained on a 100h random selection of the training data with AWD≤0.7s and PMER≤25% according to the v1 refined transcript. For non-speech data, intra-segment silences were used and the remaining audio was used as speech data. Audio was parameterised using 40 filter bank features and the feature vector for each frame extended with its preceding 27 and succeeding 27 frames. 6 hidden layers were used in the DNN with 1,000 sigmoid units in the first hidden layer and 200 units in other layers. The DNN was trained with two softmax units in the output layer corresponding to speech and non-speech. During segmentation, the DNN estimated posterior probabilities are converted to log-likelihoods, and decisions are made by Viterbi decoding with an HMM that ensures a minimum 2 frames duration of each class. Finally, the Change Point Detection (CPD) and Iterative Agglomerative Clustering (IAC) stages of the diarisation system described in [33] were applied and the internal silences threshold is set to 50 frames. This segmenter, denoted as DNN.seg-v1 gave a missed speech rate on the development set of 2.6% and a false alarm rate of 4.2%. Using this segmenter, an extra reduction of 3.8% WER absolute was obtained on the development set compared to the baseline segmenter as given in Table 1.

Finally we used a better speaker independent sequence-trained DNN hybrid system (H700.v1) based on a selection of 700h with 0.165s≤ AWD≤0.66s and a PMER≤40.0% from the v1 refined transcripts. Using this system, we reduced the WER by 1% absolute on the development set compared to the H200.v1 acoustic models.

3.2. Comparison with the v1 refined transcripts

The resulting system was used to refine the original transcripts. The refined transcripts will be denoted as v2 in the following. A comparison between the v1 and v2 refined transcripts in terms of quantity of data is presented in Figure 2. It can be seen that the refined alignment system significantly increased the quantity of data having a zero PMER from 140h to 209h. This allowed the training of a better segmenter which will be presented in Section 4. Moreover, keeping an operating point of 700h of data, the PMER decreased from 40% to 30%. A per genre analysis is presented in the top subfigure of Figure 3. Comparing v1 for a PMER=40% with v2 for a PMER=30% corresponding to the 700h operating point, the new alignment significantly changed the distribution of data across genres reducing news data but increasing all others. It also increases the proportion of the harder genres such as drama and for those genres, the refined system

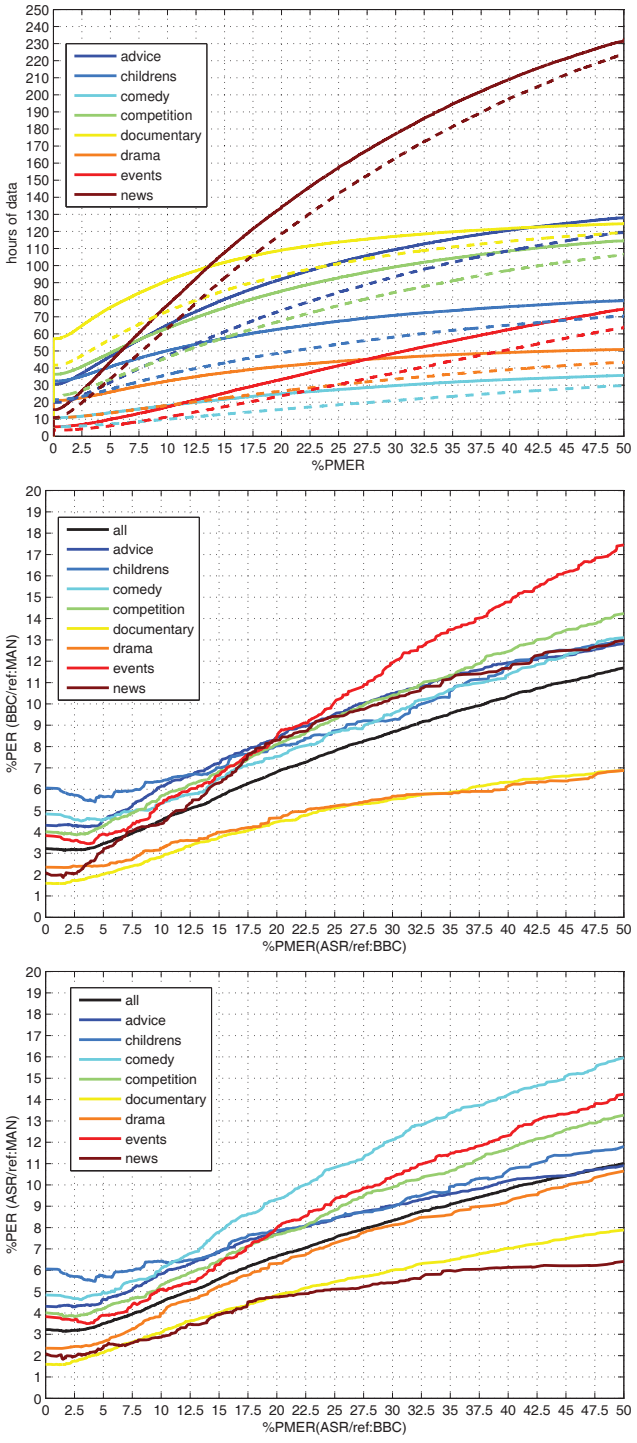

Fig. 3. **top**: Cumulative duration of the selected data on the training dataset per genre according to a threshold on PMER according to the v1 (- - -) and v2 (——) refined transcripts for 0.165s≤AWD≤0.66s. v1 refers to the transcripts provided to the participants and v2 refers to the transcripts refined for the transcription task. Phone Error Rate (PER) of a selection on the development dataset for a given PMER considering **middle**: the original transcript (BBC) **bottom**: the lightly supervised transcript (ASR). The manual transcription of the development set (MAN) was used as reference.

was better at identifying good transcript regions.

Figure 3 shows the Phone Error Rate (PER) for a given PMER considering the original transcript (BBC) in the middle subfigure and the lightly supervised transcript (ASR) in the bottom one. This used the development set following the same lightly supervised process as for the training data and using the manual transcription as reference. Globally the quality of transcripts is similar for both BBC and ASR. However, for some genres the difference can be striking. This is the case for the news for which lightly supervised transcripts have a much better quality than the original transcripts and it is also the case for the advice and events. One possible explanation is that for news broadcasts and live sports, real-time captioning [27] is increasingly common since 2001: respeakers (or voice writers) uses a mask or speech silencer to repeat what they hear into a speech recogniser to estimate the corresponding text. Thus, they might reformulate what they hear to enhance clarity and some errors can be due to the speech recognition software. However, for drama and comedy, the original transcripts have better quality than lightly supervised ones, which is probably because they are created offline and the audio often includes noisy environments and highly expressive speaking styles for which acoustic model performance may be poor. From these observations, we plan to explore transcripts combination according to genre in future work.

4. DEVELOPMENTS FOR THE ALIGNMENT TASK

4.1. Description of the task

In the alignment task, participants were supplied with single stream audio of the development set, and transcripts without time-stamps. Text transcripts did not include line breaks except when a speaker change occurred. Those were produced from the original BBC captions by removing time-stamps and merging consecutive tokenized text chunks spoken by the same speaker. As the transcripts can be approximate in some cases, participants were required to identify those words that were actually spoken as well as when they are spoken. Hence, the alignment task required to find the precise start and end time of each word in the audio. The alignment task was *script-constrained* [7]: words could be removed from the original provided transcript but no words added. Overlapping sections of the audio where two or more speakers talk at the same time were ignored during scoring.

The alignment task was framed as a word detection task and the assessment of alignment quality was based on *Precision* (P), *Recall* (R) and F*-measure*. Words were individually assessed to be correctly or incorrectly aligned. The correctly assigned words were then counted and the measures computed.

$$P = \frac{N_{\mathrm{match}}}{N_{\mathrm{hyp}}},\ R = \frac{N_{\mathrm{match}}}{N_{\mathrm{ref}}},\ F = 2 \times \frac{P \times R}{P + R} \qquad (1)$$

where N_{hyp} is the number of words in the scored automatic alignment system output (*hypothesis*) that are also in the original transcript. The system output contains words from the original transcript that were found in the audio, and their precise timings. N_{ref} is the number of words in the reference, that are also in the original transcript and N_{match} denotes matching word counts. The *reference* is the ground truth against which the system output of participants were measured. It was based on manual transcripts of the data, to the same standards that is common for ASR transcription tasks. The reference also contains the timings of the words derived by forced alignment, under the constraints of manually generated segment timings. To determine word matches the scoring used the boundary method with a 100ms window so that words are considered matched if the start and end times fall within a range of 100ms of the respective reference times. During the scoring, the reference and the hypothesis are first aligned against the original transcript. Regions of overlapped speech are then removed from both reference and hypothesis and word are matched with the original transcripts. Finally the hypothesis is aligned against the reference and metrics are computed.

system	F	Precision	Recall	N_{hyp}	N_{match}
A1	0.8847	0.8577	0.9134	155,951	133,761
A2*	0.8974	0.8594	0.9389	159,990	137,503
B1*	0.9120	0.9283	0.8936	141,404	131,260
B2	0.9149	0.9266	0.9036	142,809	132,324
B3*	0.9158	0.9282	0.9037	142,580	132,343
B4	0.9160	0.9311	0.9013	141,754	131,991

Table 2. Assessment of alignment quality of the different system developed for the MGB challenge alignment task on the development data. Systems with a * were the ones submitted for the evaluation data (primary and contrastive submissions). **B1*** was the system used for the primary submission. N_{ref} =146449.

4.2. Maximisation of recall

In the successive developments of our system for the alignment task, we first focused on maximizing the recall. Results obtained on the development set are presented in Table 2 and the corresponding systems are described below.

The A1 system was the one presented in section 3 used for the production of the v2 refined transcripts for the transcription task. Segmentation was done using the DNN.seg-v1 system and the H700.v1 speaker independent sequence-trained DNN hybrid system was used for both the lightly supervised decoding and forced-alignment. Lightly supervised decoding was performed using episode level biased LMs interpolated with the general LM estimated on the 640M of captions provided for the challenge and it was the case for all systems described in the following of the paper. Performance of the A1 system is given in Table 2 with an F-measure value of 0.8847

Compared to the original transcripts aligned in the previous sections, the transcripts provided for the task didn't include time-stamps. It was therefore not possible to constrain the decoding to the region spanned by the first and last time stamps of the caption files. Furthermore, transcripts didn't include line breaks, except when a speaker change occured. Given that our previous systems were using line breaks as the only potential split points, splitting rules described in section 2.2 had to be modified for the task. Thus, in this system, a split can occur on a line break (due to speaker change) positioned before 3 consecutive matching words or on the border of a segment defined by the automatic segmenter positioned before or after a matching word sequence of at least 2 consecutive words. Each segment of the obtained partition was then force-aligned and a new split was obtained according to the two following modified rules: a split can occur on a line break not followed by a deletion or on the limit of a segment defined by the automatic segmenter if the following silence duration is greater than 0.25s. An improvement of 0.87% absolute in recall was obtained by applying these adapted splitting rules.

A new DNN speech/non-speech segmenter was also trained on a selection of 209h of data with 0.165s≤AWD≤0.66s and a PMER=0% using the v2 refined transcripts. Only the speech was used for the

speech data (173h). For non-speech data (313h), both intra-segment silences and inter-segment silences >1s were used. Any speech in the inter-segment audio was filtered out by a previously trained DNN speech/non-speech segmenter. The same DNN topology as the DNN.seg-v1 was used. During the segmentation, both the CPD and IAC stages of the diarisation pipeline were applied and a 30s threshold on internal silences was used. On the development set, the resulting missed-speech was equal to 2.5% and false alarm to 1.9%. An extra improvement of 0.33% absolute in recall was obtained using this new DNN.seg-v2 segmenter.

Finally, a speaker independent sequence-trained DNN hybrid system (H700.v2) based on a selection of 700h with 0.165s≤ AWD≤0.66s and a PMER≤30.0% using the v2 refined transcripts was trained and then used for the lightly supervised decoding. A speaker independent tandem system (T700.v2) trained on the same selection was used for the forced-alignments. Using this better system, an extra improvement of 1.35% absolute in recall was obtained. The resulting system denoted A2 was our system that maximised the recall on the development set with a value equal to 0.9389 with 137,503 matching words giving a improvement of 2.55% absolute in recall compared to the A1 system. This system was used for one of our contrastive submissions for the evaluation. Moreover, the good performance of both the segmenter and the acoustic models trained on the new v2 refined transcripts shows the advantage of having run a second iteration of the alignment process on the training data.

4.3. Data selection

Starting from the A2 system, we then aimed to increase the precision and F-value by selectively removing words from the aligned data. Lightly supervised decoding on the segmentation derived from the aligned transcripts using the H700.v2 system was performed for this purpose. We removed words in the aligned transcript corresponding to deletions (the aligned transcript being the reference). By doing so, we improved the precision by 3.48% while reducing the recall by 2.59% absolute, leading to an improvement in F-measure by 0.61% absolute. We also removed words for which the absolute time difference Δt of the start positions with the matching word in the lightly supervised transcripts was greater than a given threshold. Different values were tested in the range 0-1s and it was found the the F-value is maximised for $\Delta t \leq 250$ms on the development set. Doing so greatly improved the precision (3% absolute) leading to an extra improvement of 0.75% absolute in F-measure value. Confusion networks (CNs) are used for minimum word error rate decoding of the aligned segment considering the biased LM and allow estimates of word posterior probabilities to be computed. As mentioned in Section 2.3, these posterior estimates need to be mapped using a decision tree trained on a reference dataset (here the development set) as they tend to be over-estimate the true posteriors. We used those confidence scores per word of the lightly supervised transcripts to remove substituted words having a confidence score ≤ 0.90. This selection gave an extra small gain in precision leading to an F-measure value equal to 0.9120. This system denoted B1 was used for our primary submission for the evaluation and achieved the highest F-measure of 0.9001 on the MGB evaluation set.

Another approach was also based on confidence scores but for words in the aligned transcripts. The reference transcription was first aligned with the CNs. Where there was an alignment of the reference word with a word in the CN, the confidence score from the word in the CN was assigned to the reference word. If the reference word was not aligned then a floored value (or NULL) was assigned to the reference word. By simply removing all words that did not appear

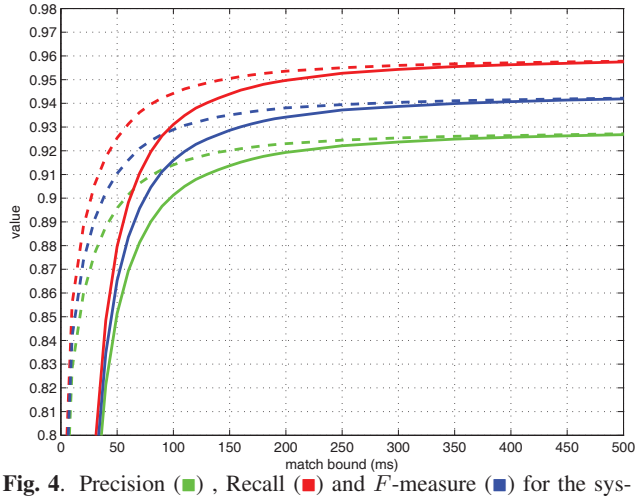

Fig. 4. Precision (■) , Recall (■) and F-measure (■) for the system (B4) maximising the F-measure value for different matching window sizes considering the official scoring reference (—) and the reference aligned using the T700.v2 system (- - -).

in the CN (no other selection criterion was applied), a F-value on the development set of 0.9149 was obtained for our B2 system. The time-stamps provided by the lightly-supervised decoding (re-aligned using the T700.v2 system at the phone level in order to remove silences) were used for matching words as replacement of the aligned original transcript ones. By doing so, an improvement of 0.09% absolute was obtained for our B3 system which was used for one of our contrastive submissions for the evaluation. Using $\Delta t \leq 1s$, our B4 system reached an F-measure value of 0.9160

Finally, low precision can be due to the alignment procedure but also to the acoustic models used for the alignment. In Figure 4 we plotted the precision, recall and F-measure for the B4 system for different matching window sizes considering the official scoring reference (—) and the one re-aligned using our T700.v2 system (- - -) on the development set. An improvement of 2.35% (F=0.9395) could be obtained in F-measure by extending the matching window size to 250ms. Moreover, at 100ms, which was the precision required for the task, the difference due to the acoustic models used to produce the reference is significant, while the difference is negligible for a matching window size of 250ms.

5. CONCLUSIONS

We described the different alignment systems developed for the preparation of data for the Multi-Genre Broadcast challenge and for our participation in the transcription and alignment tasks. Multiple refinements, including audio segmentation based on deep neural networks (DNNs) and the use of DNN-based acoustic models, were used to improve performance. It was shown that multiple iterations of the alignement process were useful for the improvement of both the segmenter and acoustic models. Different schemes for the selection of training data were presented, including a selection according to genre which will be assessed in future work. For the alignment, we first aimed at maximising the recall and then improved the F-measure value by using different selection schemes. It was observed that for a matching window size of 100ms, the difference related to the acoustic models used to prepare the reference can be significant. Finally, our system participated in the MGB challenge and achieved the highest F-measure value on the MGB evaluation set.

6. REFERENCES

[1] J. Ogata and M. Goto, "Podcastle: Collaborative training of acoustic models on the basis of wisdom of crowds for podcast transcription," in *Proc. Interspeech*, 2009.

[2] C. Alberti, M. Bacchiani, A. Bezman, C. Chelba, A. Drofa, H. Liao, P. Moreno, T. Power, A. Sahuguet, M. Shugrina, and O. Siohan, "An audio indexing system for election video material," in *Proc ICASSP*, 2009, pp. 4873–7876.

[3] R.C. van Dalen, J. Yang, and M.J.F. Gales, "Generative kernels and score-spaces for classification of speech: Progress report," in *Tech. Rep. , Cambridge University Engineering Department*, 2012.

[4] M. Larson, M. Eskevitch, R. Orderlman, C. Kofler, S. Schmiedeke, and G.J.F. Jones, "Overview of Mediaeval 2011 Rich speech retrieval task and genre tagging task," in *Working Notes Proceedings of the MediaEval 2011 Workshop*, 2011.

[5] Y. Raimond, C. Lowis, R. Hodgson, and J. Tweed, "Automatic semantic tagging of speech audio," in *Proc. WWW 2012*, 2012.

[6] 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 *Proc of the first Workshop on Speech, Language and Audio in Multimedia*, 2013.

[7] 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 *Proc. IEEE ASRU*, 2015.

[8] J. Robert-Ribes and R. Mukhtar, "Automatic generation of hyperlinks between audio and transcript," in *Proc. Eurospeech*, 1997.

[9] P.J. Moreno, C. Joerg, J.M.V. Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *International Conference on Spoken Language Processing*, 1998, vol. 8.

[10] L. Lamel, J.L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," in *Computer Speech and Language*, 2002, vol. 16, pp. 115–129.

[11] A. Venkataraman, A. Stolcke, W. Wang, D. Vergyri, V.R.R. Gadde, and J. Zheng, "An efficient repair procedure for quick transcriptions," in *Proc. ICSLP*, 2004.

[12] H.Y. Chan and P.C. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training," in *Proc. ICASSP*, 2004, vol. 1, pp. 737–740.

[13] L. Mathias, G. Yegnanarayanan, and J. Fritsch, "Discriminative training of acoustic models applied to domains with unreliable transcripts," in *Proc. ICASSP*, 2005, vol. 1, pp. 109–112.

[14] B. Lecouteux, G. Linares, P. Nocera, and J.F. Bonastre, "Imperfect transcript driven speech recognition," in *Proc. InterSpeech'06*, 2006, pp. 1626–1629.

[15] A. Haubold and J. Kender, "Alignment of speech to highly imperfect text transcriptions," in *IEEE International Conference on Multimedia and Expo*, 2007, pp. 224–227.

[16] N. Braunschweiler, M.J.F Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Proc. Interspeech*, 2010, pp. 2222–2225.

[17] S. Hoffman and B. Pfister, "Text-to-speech alignment of long recordings using universal phone models," in *Proc Interspeech*, 2013, pp. 1520–1524.

[18] G. Bordel, S. Nieto, M. Penagarikano, L.J. Rodriguez-Fuentes, and A. Varona, "A simple and efficient method to align very long speech signals to acoustically imperfect transcriptions," in *Proc Interspeech*, 2012.

[19] A. Alvarez, H. Arzelus, and P. Ruiz, "Long audio alignment for automatic subtitling using different phone-relatedness measures," in *Proc ICASSP*, 2014.

[20] O. Boeffard, L. Charonnat, S. L. Maguer, D. Lolive, and G. Vidal, "Towards fully automatic annotation of audiobooks for TTS," in *International Conference on Language Resources and Evaluation*, 2012.

[21] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, "Automatic text-independent pronunciation scoring of foreign language student speech," in *Proc. of ICSLP 96*, 1996.

[22] J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub, "Automatic evaluation and training in english pronunciation," in *Proc. of ICSLP*, 1990, pp. 1185–1188.

[23] S. Dharanipragada, M. Franz, and S. Roukos, "Audio-indexing for broadcast news," in *Proc. of TREC6*, 1997.

[24] J. Foote, "An overview of audio information retrieval," in *ACM Multimedia Systems*, 1999.

[25] M.J.F. Gales, D.Y. Kim, P.C. Woodland, D. Mrva, R. Sinha, and S.E Tranter, "Progress in the CU-HTK broadcast news transcription system," in *IEEE Transactions on Audio Speech and Language Processing*, September 2006.

[26] G. Evermann and P.C. Woodland, "Design of fast LVCSR systems," in *Proc. ASRU Workshop*, 2003.

[27] C. Eugeni, "Respeaking the BBC news: a strategic analysis of respeaking on the BBC," *The Sign Language Translator and Interpreter*, vol. 3, no. 1, pp. 29–68, 2009.

[28] Y. Long, M.J.F. Gales, P. Lanchantin, X. Liu, M. S. Seigel, and P.C. Woodland, "Improving lightly supervised training for broadcast transcriptions," in *Proc. Interspeech*, 2013.

[29] HTK 3.5, http://htk.eng.cam.ac.uk.

[30] C. Zhang and P.C. Woodland, "A general artificial neural network extension for HTK," in *Proc. Interspeech*, 2015.

[31] 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 *Proc. ASRU*, 2015.

[32] P. Karanasou, M.J.F. Gales, P. Lanchantin, X. Liu, Y. Qian, L. Wang, P.C. Woodland, and C. Zhang, "Speaker diarisation and longitudinal linking on multi-genre broadcast data," in *Proc. ASRU*, 2015.

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