




Lightly supervised alignment of subtitles on multi-genre broadcasts

Oscar Saz¹ · Salil Deena¹  · Mortaza Doulaty¹ ·
Madina Hasan¹ · Bilal Khaliq¹ · Rosanna Milner¹ ·
Raymond W. M. Ng¹ · Julia Olcoz² · Thomas Hain¹

Received: 26 October 2017 / Revised: 13 April 2018 / Accepted: 23 April 2018 /
Published online: 29 May 2018
© The Author(s) 2018

Abstract This paper describes a system for performing alignment of subtitles to audio on multigenre broadcasts using a lightly supervised approach. Accurate alignment of subtitles plays a substantial role in the daily work of media companies and currently still requires large human effort. Here, a comprehensive approach to performing this task in an automated way using lightly supervised alignment is proposed. The paper explores the different alternatives to speech segmentation, lightly supervised speech recognition and alignment of text streams. The proposed system uses lightly supervised decoding to improve the alignment accuracy by performing language model adaptation using the target subtitles. The system thus built achieves the third best reported result in the alignment of broadcast subtitles in the Multi-Genre Broadcast (MGB) challenge, with an F1 score of 88.8%. This system is available for research and other non-commercial purposes through webASR, the University of Sheffield's cloud-based speech technology web service. Taking as inputs an audio file and untimed subtitles, webASR can produce timed subtitles in multiple formats, including TTML, WebVTT and SRT.

Keywords Multigenre broadcasts · Lightly supervised alignment · Language model adaptation · Subtitles

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

✉ Salil Deena
s.deena@sheffield.ac.uk

Thomas Hain
T.Hain@sheffield.ac.uk

¹ Department of Computer Science, University of Sheffield, Sheffield, UK

² Department of Electrical Engineering and Communications, University of Zaragoza, Zaragoza, Spain

1 Introduction

Alignment of subtitles to audio is a task which aims to produce accurate timings of the words within a subtitle track given a corresponding audio track. The alignment task is of significant relevance in the media industry. It is required when producing television shows and motion pictures at several stages, for instance when adding the speech track to the general audio track, when generating dubbed versions in multiple languages [26], and also when preparing captions for final broadcasting [42]. Semi-automated solutions exist, and they involve different degrees of human post-edit based on the automatically generated results. Further task automation would drive down production time and the expensive human labour costs in the production of accurate timing labels.

Spoken language technologies, such as Automatic Speech Recognition (ASR), can provide tools for the automation of these processes. Thanks to deep learning techniques [18], the recent improvements in empirical performance make ASR and relevant technologies suitable for a wide range of tasks in the multimedia domain, including the automatic generation of subtitles [1]. In terms of the alignment task, typically a set of previously trained acoustic models is first used to decode the audio signal into time-marked strings. The Viterbi algorithm, a dynamic programming technique, is then used to pair the reference and the decoded strings. By this method text can be quickly aligned to audio. However, this technique is unreliable with long audio files so other proposals using automatic speech segmentation and recognition have long been proposed [29]. Furthermore, the quality of such alignment degrades significantly when the subtitles deviate substantially from the actual spoken content, which is common in subtitling for media content.

For those scenarios where transcriptions are incomplete, [5] and [20] propose the use of large background acoustic and language models, and [39] implements a method for sentence-level alignment based on grapheme acoustic models. Moreover, if the transcript quality is very poor, [24] presents an alternative to improve lightly supervised decoding using phone-level mismatch information. To deal with complex audio conditions, [11] proposes to add audio markups to the audio file in order to facilitate the later alignment procedure. Other related works such as [4] also take into account situations where transcripts include a mixture of languages.

The main contribution of this paper is to propose a fully automated system for the alignment of subtitles to audio using lightly supervised alignment. The lightly supervised alignment process involves automatic speech recognition of the audio in order to obtain an output and then matching this output, which includes timings, to the subtitles. In order for this process to be effective, the automatic speech recognition output should match the subtitle text as closely as possible, which is known as lightly supervised decoding. This is done by biasing the language model to the subtitles. In this work, both recurrent neural network language models (RNNLMs) and n -gram language models, are biased to the subtitles. Whilst the biasing of n -gram language models to the subtitles is a known procedure and is achieved by merging the n -gram counts, the biasing of RNNLMs to subtitles has not been previously explored for lightly supervised decoding and is a novelty proposed in this paper. Another contribution is an error correction algorithm which helps improve the correctness of word-level alignments and the removal of non-spoken words. The proposed system is made available to the public using the webASR web API, which is free to use by both industrial and academic users. By making this system available for research and demonstration purposes, this paper aims to encourage users operating in the field of broadcast media to investigate lightly supervised approaches to deal with subtitling requirements.

This paper is organised as follows: Section 2 discusses the alignment task and the different ways it can be performed and measured. Section 3 will describe the proposed system of lightly supervised alignment using lightly supervised decoding. Section 4 will present the experimental conditions and results of the proposed system on the Multi-Genre Broadcast (MGB) challenge data [2]. Section 5 will describe the deployment of this system through webASR and how it can be used to build applications using the webASR API. Finally, Section 6 will present the conclusions to this work.

2 The alignment task

Many tasks in the speech technology domain have very clearly defined targets and measures of quality. For instance, in ASR the target is to produce the same sequence of words spoken in the audio; and this can be evaluated by a direct comparison with a manual transcription and counting the number of correct and incorrect words. However, the alignment task is liable to multiple interpretations and its assessment often relies on subjective measurements.

In general terms, there are two approaches to the alignment task. The first one takes as input a list of word sequences and aims to provide a start time and end time for each sequence, from where the input audio corresponds. With the time markings, input audio is truncated into short segments, each linked to the word sequences. Therefore, this approach is the most relevant to subtitling and close captioning. In this approach, a perfect alignment requires the word sequences to be mapped to the audio even when they are not verbatim transcriptions of the audio and may contain paraphrases or deletions and insertion of words for subtitling reasons.

The second approach aims to produce time information at the word level i.e, the precise time at which every given word is pronounced. In this case, either the word sequence to align is an exact verbatim transcription of the audio, or the alignment procedure must discard all words not actually pronounced, as they cannot be acoustically matched to any section of the audio. This approach is of relevance when a finer time resolution in alignment is required, for instance in dubbing procedures.

The way the quality of an alignment is measured differs depending whether sequence-level or word-level alignment is required. For sequence-level alignments, it is possible to manually come up with a ground truth labelling of where the segments should be aligned to and then compare this to the sequence boundaries automatically derived. In applications such as subtitling this manual ground truth depends on several subjective elements, such as the speed at which viewers can plausibly read the subtitles, or the way sentences are paraphrased, which makes measuring the quality of the alignment very subjective.

For word-level alignments, objective measurements are more feasible, as the start time and end time of any given word in an audio can always be identified manually. In this case, the alignment task turns into a classification task where the target is to correctly determine the timings for each word in the ground truth. As any other classification task, it can then be measured in terms of True Positive (TP) rate, the rate of words for which the times are correctly given, False Negative (FN) rate, the rate of words for which no time or incorrect times are given, and False Positive (FP) rate, the rate of words for which a time is given when no such word exists. From these values, standard classification metrics such as accuracy, precision, recall, sensitivity, specificity or F1 score can be computed.

Figure 1 visualises the differences between sequence-level and word-level alignment. In this example, the utterances “*Not far from here is a modern shopping mall. There are all kinds of shops there.*” is subtitled as “*A modern shopping mall is nearby. There are all kinds*

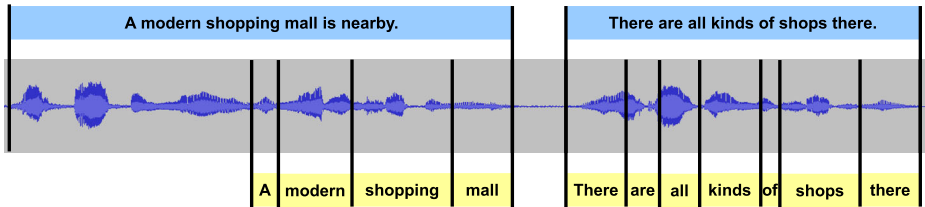


Fig. 1 Sequence-level (top, blue) and word-level (bottom, yellow) alignment for the subtitle text “A modern shopping mall is nearby. There are all kinds of shops there.”. Actual utterance is “Not far from here is a modern shopping mall. There are all kinds of shops there”

of shops there.”. Looking into the first half of the subtitles “A modern shopping mall is nearby.”, when performing sequence-level alignment the subtitles should be aligned to the utterance “Not far from here is a modern shopping mall.” since they are paraphrases of each other. In the word-level alignment, only the words actually spoken “A modern shopping mall” could be aligned. Moving on to the second half of the subtitles “There are all kinds of shops here”, the sequence-level and word-level alignment output the same words, just with sequence or word timings as required.

3 Lightly supervised alignment system

The system proposed in this paper follows the concept of lightly supervised alignment, i.e., of an alignment system where the input subtitles are used to train lightly supervised models that can be used to inform the alignment procedure. The main building blocks that are required for this setup are speech segmentation, lightly supervised decoding and the alignment itself.

3.1 Speech segmentation

Speech segmentation is the process in which a large piece of audio, i.e. with long duration is split into many short utterances, normally delimited by speech pauses or non-speech areas. The process consists initially of a Voice Activity Detection (VAD) procedure, where areas containing purely speech are identified in the audio signal. From these sections of pure speech, speech utterances are created by merging several of these sections into larger chunks. The final goal is to generate acoustically coherent segments of continuous speech that can then be used independently in downstream processes, like speech recognition.

VAD is a well studied problem in the speech community, where several solutions have long been proposed [34]. Previous approaches used acoustic properties of the speech signal to identify speech areas. The most basic VAD systems are based on detecting areas of higher energy, usually associated with speech; while more complex approaches performed an online estimation of the spectral characteristics of speech and non-speech areas to perform this separation [35].

Statistical approaches to VAD have produced improved performance, including the use of Neural Networks (NNs) to learn the classification of speech and non-speech [10, 14]. Deep Neural Networks (DNNs) have provided further improvements in this task [37] and are the basis of the VAD proposed in this system. In this setup, the neural networks are trained to classify each frame in one of two classes, one corresponding to speech being present and the other one representing speech not being present.

During VAD implementation, a DNN provides an estimation of the posterior probability for each audio frame, on whether or not it contains speech. Subsequently, a Hidden Markov Model (HMM), that takes as input the posteriors from DNN, determines the optimal sequence of speech and non-speech chunks by considering the speech/non-speech likelihood in a smoothed version over multiple frames. The final output is a speech segmentation which corresponds to segments of continuous speech.

Figure 2 provides an example of the speech segmentation process on a 25-second audio clip. The red chunks correspond to areas of speech as detected by the VAD, which has identified 8 speech areas. The green chunks are the speech segments obtained after agglomerating the VAD segments. In this case, speech segments with short pauses in between are merged into a single segment, resulting in only 4 output segments from the initial 8.

3.2 Lightly supervised decoding

Once a set of speech segments has been identified, decoding is the process in which ASR is run in order to provide a hypothesis transcript for each segment in the set. The decoding process proposed in this system employs a 3-stage procedure based on a standard setup using the Kaldi toolkit [33]. First, it performs decoding using a set of previously trained hybrid DNN-HMM acoustic models [19] and a Weighted Finite State Transducer (WFST) calculated from a previously trained 3-gram language model. This generates a set of lattices which are then rescored using a 4-gram language model. Finally, the 25-best hypotheses for each segment are rescored again using a Recurrent neural network language model (RNNLM) [7, 27]. The hypothesis with the best final score after this process is given as the output for each segment.

Lightly supervised decoding involves using the input subtitles that have to be aligned to the audio to adapt the language model components of the previous decoding system in order to improve the quality of the decoding hypothesis. Figure 3 shows the block diagram of the proposed system. On the left-hand side of the diagram, the decoder works in 3 stages as described above: Decoding, lattice rescoring and N-best RNN rescoring. On the right-hand side, the lightly supervised procedure is depicted.

First, the subtitles are tokenised — the text is normalised on a set of constraints used by the decoding procedure. This includes capitalisation, punctuation removal, numeral-to-text conversion and acronym expansion. In this procedure, for instance, the subtitle text “*Today is 14th of July, you are watching BBC1.*” is converted to “*TODAY IS FOURTEENTH OF JULY YOU ARE WATCHING B. B. C. ONE*”.

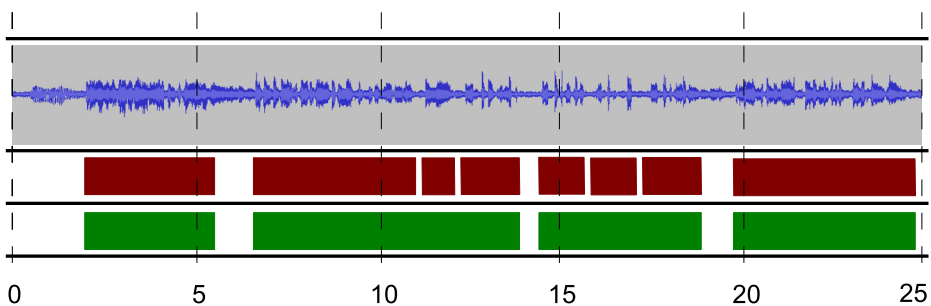


Fig. 2 Example of speech segmentation process

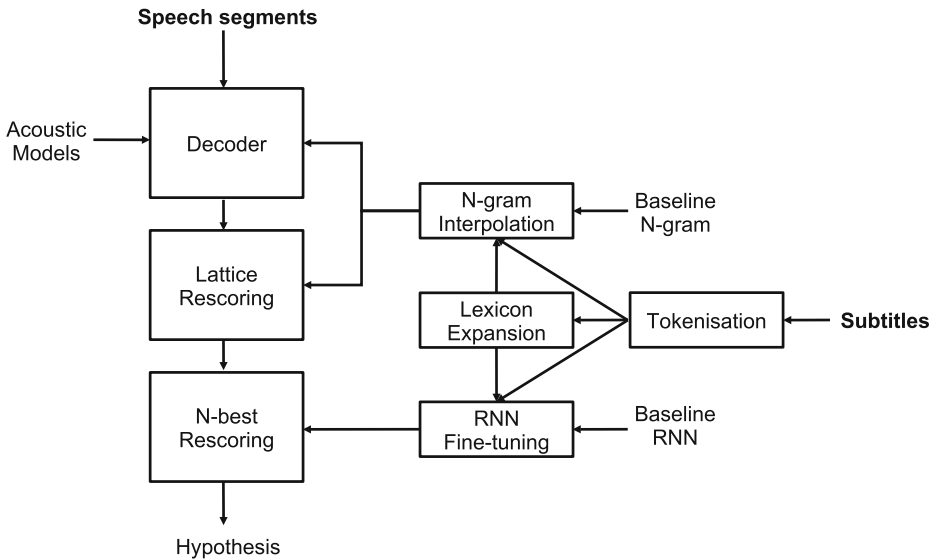


Fig. 3 Block diagram of lightly supervised decoding system

After tokenisation, the decoding lexicon, which contains the set of words that can be recognised is expanded with out-of-vocabulary (OOV) words. OOV are words from the subtitles in the training set that are not covered by the decoding lexicon. In this process, a phonetic transcription of the words is either extracted with some carefully crafted dictionary or generated by automatic phonetisation. In the proposed system, the Combilex dictionary [36] is used to extract manual pronunciations and the Phonetisaurus toolkit is used to derive new pronunciations [30] when not covered by Combilex.

The next step involves n -gram language model (LM) interpolation between a previously trained baseline n -gram LM that provides a full coverage of the target language and, an n -gram LM trained exclusively on the subtitles. In this work, this is achieved using the SRILM toolkit [40] and biases the decoder and the lattice rescoring towards producing hypotheses which are closer to the words and language used in the subtitles. Such interpolation of n -grams [21] has been shown in the past to help improve accuracy in ASR systems when interpolating a large out-of-domain n -gram model with a smaller in-domain n -gram model.

Finally, a previously trained baseline RNNLM is fine-tuned [6, 9] using the subtitles by further training the RNNLM using the subtitle text as input for a given number of iterations in order to make the RNNLM model closer to the linguistic space of the subtitles. Fine-tuning of RNNLMs has also been shown to produce better accuracies in the ASR task [9, 41]. Once the adapted n -grams and RNNLMs are trained, they can be used in the decoding procedure instead of the baseline language models, i.e., n -grams/WFSTs for decoding and lattice rescoring and RNNLM for N -best rescoring.

3.3 Alignment

Finally, when an ASR hypothesis transcript is available for the audio file, a Dynamic Time Warping (DTW) alignment is performed comparing the hypothesis and the input subtitles.

The aim of this alignment is to assign words in the subtitles to segments in the hypothesis. This alignment is performed in several stages. First, sequences of words from the hypothesis and the subtitles with high matching content are matched together. For each of these matching word/sequence pairs, the timing of the subtitle is derived from that of the corresponding ASR hypothesis. When all the best matches are found, the residual words in the subtitles not already matched will have their timings assigned to fill the time gaps left behind by previous matching.

Table 1 presents an example of this procedure in a 34-second audio clip. The speech segmentation identifies two segments from 1.47 seconds to 17.59 seconds and from 21.96 seconds to 34.15 seconds; and the lightly supervised decoding gives the hypothesis presented in rows 2 and 3 of the Table. The original and tokenised subtitles for this clip are shown in rows 4 and 5. Then, the output of the lightly supervised alignment system is presented, which gives 3 segments, different to the original ones. The first segment matches the subtitles in the range “Justice, wombats ... one, go!” with the hypothesis in the range “JUSTICE WOMBATS ... ONE GO”. The first word of the hypothesis (“PENDULUM”) is deleted as it does not match the subtitles. The next match occurs between the subtitles in “looking forward ... Chipmunk. Chipmunk.” and the hypothesis in “LOOKING FORWARD ... CHIPMUNK CHIPMUNK”. The remaining subtitle words cannot be matched with any remaining hypothesis so they are then assigned to a new intermediate segment covering “I’m Anthony. Who are you”.

From here on, the system provides word-level time information by performing Viterbi forced alignment of the words in each segment. At this step, some of the segments may be dropped from the output if the alignment procedure cannot find an acoustic match in the audio, resulting in the loss of some words in the output. On the other hand, it will usually be

Table 1 Lightly supervised alignment example

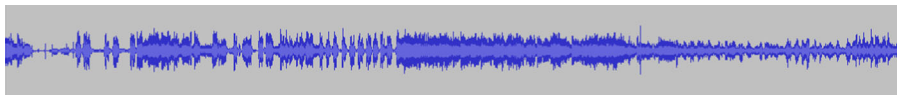

Hypothesis [1.47–17.59]: PENDULUM JUSTICE WOMBATS PIGEON DETECTIVES THE KURDS THE ENEMY RECOGNISE GOLDFRAPP RADIO ONE’S BIG WEEKEND MAIDSTONE KENT DAY TWO FIVE FOUR THREE TWO ONE GO
Hypothesis [21.96–34.15]: I DON’T LOOKING FORWARD SAME PENDULUM SAY THE KINKS CAN’T WORRY CHIPMUNK ON THE STAGE BUT MY FAVOURITE MY MIND’S CHIPMUNK
Subtitles: Justice, wombats, the cooks, the Enemy, Goldfrapp - Radio 1’s Big Weekend, Maidstone five, four, three, two, one, go! I’m Anthony. Who are you looking forward to seeing? Pendulum. Who Pendulum. Who are you looking forward to seeing? The Kooks. Chipmunk. Chipmunk.
Subtitles (tokenised): JUSTICE WOMBATS THE COOKS THE ENEMY GOLDFRAPP RADIO ONE’S BIG WEEKEND MAIDSTONE FIVE FOUR THREE TWO ONE GO I’M ANTHONY WHO ARE YOU LOOKING FORWARD TO SEEING PENDULUM WHO PENDULUM WHO ARE YOU LOOKING FORWARD TO SEEING THE KOOKS CHIPMUNK CHIPMUNK
Output [3.68–15.14]: Justice, wombats, the cooks, the Enemy, Goldfrapp - Radio 1’s Big Weekend, Maidstone five, four, three, two, one, go!
Output [15.14–24.60]: I’m Anthony. Who are you
Output [24.60–34.15]: looking forward to seeing? Pendulum. Who Pendulum. Who are you looking forward to seeing? The Kooks. Chipmunk. Chipmunk.

Table 2 Datasets in the MGB challenge

Data	Shows	Audio	Speech
Training	2,193	1580.4 h.	1196.7 h.
Development	47	28.4 h.	19.6 h.
Evaluation	16	11.2 h.	8.6 h.

the case that words which are not pronounced in the audio but are in the subtitles will still appear in the output. To improve the correctness of word-level alignments and remove non-spoken words, an algorithm [31] has been proposed to find such cases and remove them. This algorithm uses a previously trained binary regression tree to identify these words based on some acoustic values of each aligned word, like duration or confidence measure.

The alignment procedure generates an output on tokenised form, and in order to recover the original textual form of the subtitles a re-normalisation procedure is performed to recover punctuations, cases, numerals and acronyms in their initial form. This can be easily done as a hash table is generated during the tokenisation procedure linking each original word in the subtitles to one or more tokens in the normalised form.

4 Experiments and results

The experimental setup in which the proposed system was evaluated was based on Task 2 of the MGB challenge 2015[3]. This task was defined as “Alignment of broadcast audio to a subtitle file” and was one of the four core tasks of the challenge¹. The MGB challenge aimed to evaluate and improve several speech technology tasks in the area of media broadcasts, extending the work of previous evaluations such as Hub4 [32], TDT [8], Ester [12], Albayzin [45] and MediaEval [23]. MGB was the first evaluation campaign in the media domain to propose lightly supervised alignment of broadcasts as a main task.

The focus of the MGB challenge was on multi-genre data. Most previous work on broadcast audio has focused on broadcast news and similar content. However, the performance achieved on broadcast data dramatically degrade in the presence of more complex genres. The MGB challenge thus defined 8 broadcast genres: Advice, children’s, comedy, competition, documentary, drama, events and news.

4.1 Experimental setup

The experimental data provided on the MGB challenge 2015 consisted of more than 1,600 hours of television shows broadcasts on the BBC through April and May of 2008. It was divided into training, development and evaluation sets as shown in Table 2.

The only transcription available for the 1,200 hours of training speech were the original BBC subtitles, aligned to the audio data using a lightly supervised approach [25]. No other audio data could be used to train acoustic models according to the evaluation conditions. More than 650 million words of BBC subtitles, from the 1970s to 2008, were also provided for language model training. As with the acoustic model training data, no other linguistic materials could be used for training language models. For building lexical models, a

¹www.mgb-challenge.org

version of the Combilex dictionary [36] was distributed and was the only available source for developing the lexicon.

The system used for decoding [38] was based on acoustic models trained on 700 hours of speech extracted from the available 1200 hours using a segment-level confidence measure based on posterior estimates obtained with a DNN [46] to select only segments with an accurate transcription. A 6-hidden-layer DNN with 2,048 neurons was trained using Deep Belief Network (DBN) pretraining and then fine-tuned using first the Cross-Entropy (CE) criterion, followed by the state-level Minimum Bayes Risk (sMBR) criterion. The input to the DNN are 15 spliced Perceptual Linear Prediction (PLP) acoustic frames. The vocabulary used for decoding was a set of common 50,000 words from the linguistic training data and n-gram language models were trained from the available linguistic resources and then converted to WFSTs. For the rescoring, an RNNLM was trained also using the available language training data.

For speech segmentation, two strategies based on 2-hidden-layer DNNs were explored [28]. In the first one, all the available acoustic training data was separated into speech and non-speech segments, providing 800 hours of speech and 800 hours of non-speech for training the DNN using the CE criterion. This is referred to as DNN VAD 1. In the second strategy, data selection was applied to yield 400 hours of audio with 300 hours and 100 hours of speech and non-speech content respectively. Identical training was performed on the carefully selected data set to give a 2-layer DNN (DNN VAD 2).

4.2 Results

Task 2 of the MGB challenge was a word-level alignment task and thus it was evaluated as a classification task. The evaluation metrics were the precision and recall of the system, with the final metric being the F1 score (or F-measure). The precision is measured as the number of TPs divided by the total number of words in the system output, which is the sum of TPs and FPs. Recall is measured as the number of TPs divided by the total number of words in the reference, which is the sum of TPs and FNs. With these two measures, the F1 score is computed as the geometrical mean of precision and recall:

$$F1 = 2 \frac{Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FN + FP} \quad (1)$$

For scoring purposes of the MGB challenge, a word is considered correct if the output start time and end time are less than 100 milliseconds away from the ground truth start and end times of that word. A set of experiments were performed using the MGB development set to investigate the optimal setup of the proposed lightly supervised alignment system, in terms of getting the best F1 scores. Table 3 presents the results for four different configurations. The first two rows show the differences achieved using an unadapted decoding system with the two speech segmentation strategies with DNN VAD 1 and DNN VAD 2. The use of the

Table 3 Results in the MGB development set

Configuration	SER	WER	Precision	Recall	F1 score
DNN VAD 1 + DNN-HMM decoding	11.9%	32.6%	90.8%	86.3%	88.5%
DNN VAD 2 + DNN-HMM decoding	16.7%	31.3%	90.5%	88.9%	89.7%
+ n-gram adaptation	—”—	24.9%	90.5%	90.1%	90.3%
+ RNNLM fine-tuning	—”—	23.2%	90.5%	90.6%	90.5%

Table 4 Results in the MGB evaluation set

System	F1 score
University of Cambridge [22]	90.0%
University of Cambridge [3]	89.3%
Proposed system	88.8%
Quorate/University of Edinburgh [3]	87.7%
Computer Research Institute of Montreal (CRIM) [3]	86.3%
Vocapia/Laboratory for Mechanics and Engineering Sciences (LIMSI) [3]	84.6%
University of Sheffield [3]	83.4%
Japan Broadcasting Corporation (NHK) [3]	79.7%

DNN VAD 2 leads to an increase in Segmentation Error Rate (SER), but reduces the Word Error Rate (WER) and improves the F1 score by a significant 1%. This is due to the fact that DNN VAD 2 only misses 1.2% of speech frames, which helps the alignment procedure to identify matches between the hypothesis and the subtitles. The use of lightly supervised decoding using n-gram adaptation reduces WER to 24.9% and increases F1 score to 90.3%. Finally, RNNLM fine-tuning provides an extra 1.7% reduction in WER and 0.2% increase in F1 score.

The final proposed system, achieving 90.5% F1 score on the MGB development set, was then run on the MGB evaluation set, where it achieved 88.8% F1 score. Table 4 presents the result for this system compared to other systems reported previously. In terms of the systems officially submitted to the MGB challenge, and reported in [3], the proposed system would achieve second place, only 0.5% below the University of Cambridge system, and substantially improving the original submission by the University of Sheffield.

5 System deployment

The automatic alignment system described in this paper has been made available through webASR². webASR was setup as a free cloud-based speech recognition engine in 2006 [15, 17, 43, 44] and was redeveloped in 2016 as a multi-purpose speech technology engine [16]. It allows research and non-commercial users to freely run several speech technology tasks, including automatic speech recognition, speech segmentation, speaker diarisation, spoken language translation and lightly supervised alignment. It runs as a cloud service on servers located at the University of Sheffield using the Resource Optimization ToolKit (ROTK) [13] as its backend. ROTK is a workflow engine developed at the University of Sheffield that allows the running of very complex systems as a set of smaller asynchronous tasks through job scheduling software in a grid computing environment.

The web interface of webASR allows new users to register for free and, once registered, to submit their audio files to one of the available tasks. Once the processing in the backend is finished, the users can retrieve the result files directly from their accounts in webASR. As processes run asynchronously, users can run multiple systems at the same time and wait for the results of each one as they happen.

²www.webasr.org

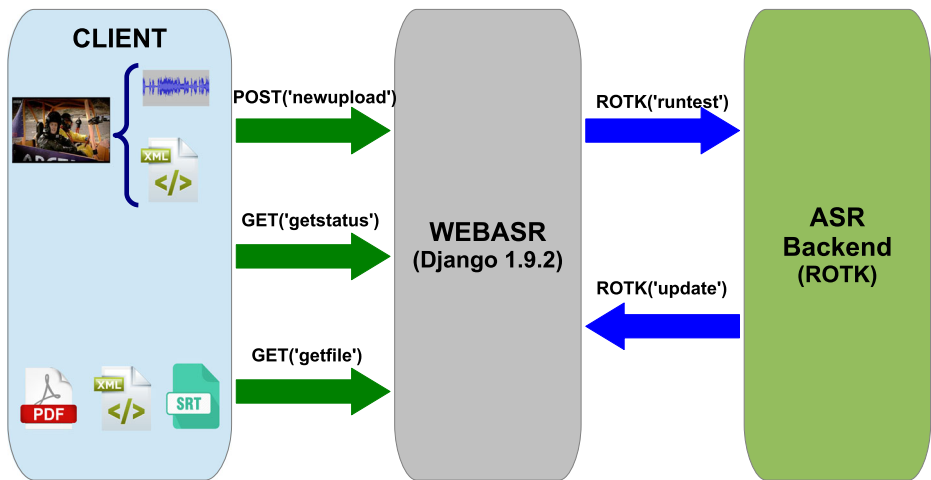


Fig. 4 Use of the webASR API for lightly supervised alignment of subtitles

In order to facilitate the building of applications using the webASR cloud service, an API was implemented using the Django web framework. Figure 4 depicts the integration of an ASR backend system into webASR using the API. The API acts as a backend system wrapper and handles all post and query requests from the user. Taking the alignment system as an example, a user first submits an audio file and an untimed subtitle file to webASR through a POST command (<http://webasr.org/newupload>). This will trigger webASR to connect to the ASR backend and run the alignment system in ROTK. The user can poll the status of the system through a GET command (<http://webasr.org/getstatus>), which will return whether the backend has finished processing the file or not. When ROTK is finished it updates webASR with the outcome. At that point, the user can use a final GET command (<http://webasr.org/getfile>) to retrieve a set of files containing the aligned subtitles. These files are PDF, XML, TTML, WebVTT and SRT formats.

6 Conclusions

This paper has presented a lightly supervised alignment system of subtitles to broadcast audio. A thorough description of the steps required to implement such a system has been given, from speech segmentation, lightly supervised decoding to text alignment. Results show that a minimum missed rate of speech in the upstream speech segmentation is essential to downstream performance improvements.

In terms of methodologies proposed in this work and in contrast to other systems for lightly supervised alignment that were proposed for the MGB challenge 2015 and whose results are given in Table 4, we must note two main novelties. The first is the use of RNNLM adaptation, achieved by fine-tuning the RNNLM on the subtitle text in order to bias the RNNLM towards the subtitles, which was shown to both reduce recognition errors as well as improve the accuracy of the alignment output. The second is the use of the error correction algorithm proposed by same authors [31], which deals with improving the correctness of word-level alignments and the removal of non-spoken words, using a binary regression tree to identify these words based on acoustic values such as duration and confidence measures.

From the point of view of lightly supervised decoding, the experiments have shown how RNNLM rescoring helps not only to reduce the recognition errors but, more importantly to improve the accuracy of the alignment output. Adaptation of n -grams and RNNLMs produce a significant reduction in recognition errors and an associated increase in alignment accuracy. In general, the lightly supervised approach has shown how it can significantly improve the outcome of the alignment task.

The current state-of-the-art system for the MGB challenge alignment task is the University of Cambridge system for lightly supervised alignment [22]. The steps for lightly supervised decoding and alignment are very similar to the one presented in this paper, except for the two novel contributions detailed above. The reason why the Cambridge system produced the best results in the challenge, was because their audio segmentation and lightly supervised decoding systems were better, making use of enhanced Deep Neural Network (DNN) architectures. In this work, the improvements proposed to both the lightly supervised decoding and alignment stages help us achieve results close to the state-of-the-art.

The proposed alignment system achieves F1 scores of 90.5% and 88.8% in the development and evaluation sets, respectively, in Task 2 of the MGB challenge. The evaluation results are the third best reported results on this setup and would achieve the second place on the official challenge results behind the Cambridge System [22]. In order to improve these results, even larger improvements in acoustic modelling and language modelling of the lightly supervised decoding stage would be necessary. While the presented system achieves 23.2% WER on the development set, it is expected that reducing this error rate to below 20% would increase the F1 score further, getting it closer to the best reported results of 90.0% in the evaluation set.

In order to facilitate the use of this system, for research and non-commercial purposes, this system has been implemented in webASR. Through its API, webASR allows an easy integration into any given workflow. Given an audio file and its corresponding untimed subtitles, timed subtitles are produced and can be used in further processing. This can greatly facilitate the work on subtitling, close captioning and dubbing.

7 Data access management

All the data related to the MGB challenge, including audio files, subtitle text and scoring scripts is available via special license with the BBC on <http://www.mgb-challenge.org/>. All recognition outputs and scoring results are available with <https://doi.org/10.15131/shef.data.3495854>.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Alvarez A, Mendes C, Raffaelli M, Luis T, Paulo S, Piccinini N, Arzelus H, Neto J, Aliprandi C, del Pozo A (2015) Automatic live and batch subtitling of multimedia contents for several European languages. *Multimedia Tools and Applications*. <https://doi.org/10.1007/s11042-015-2794-z>

2. Bell P, Gales M, Hain T, Kilgour J, Lanchantin P, Liu X, McParland A, Renals S, Saz O, Webster M, Woodland P (2015) The MGB challenge: Evaluating multi-genre broadcast media transcription. In: ASRU'15: Proc Of IEEE workshop on automatic speech recognition and understanding, Scottsdale
3. Bell P, Gales MJF, Hain T, Kilgour J, Lanchantin P, Liu X, McParland A, Renals S, Saz O, Wester M, Woodland PC (2015) The MGB Challenge: evaluating multi-genre broadcast media recognition. In: Proceedings of the 2015 IEEE automatic speech recognition and understanding workshop, Scottsdale, pp 687–693
4. Bordel G, Peñagarikano M., Rodríguez-fuentes LJ, Varona A (2012) A simple and efficient method to align very long speech signals to acoustically imperfect transcriptions. In: Proceedings of the 13th annual conference of the international speech communication association (Interspeech), Portland, pp 1840–1843
5. Braunschweiler N, Gales MJF, Buchholz S (2010) Lightly supervised recognition for automatic alignment of large coherent speech recordings. In: INTERSPEECH, pp 2222–2225
6. Chen X, Tan T, Liu X, Lanchantin P, Wan M, Gales MJF, Woodland PC (2015) Recurrent neural network language model adaptation for multi-genre broadcast speech recognition. In: INTERSPEECH'15: Proc Of the 16th annual conference of the international speech communication association, pp 3511–3515
7. Chen X, Wang Y, Liu X, Gales MJF, Woodland PC (2014) Efficient GPU-based training of recurrent neural network language models using spliced sentence bunch. In: Proceedings of the 15th annual conference of the international speech communication association (Interspeech), Singapore, pp 641–645
8. Cieri C, Graff D, Liberman M, Martey N, Strassel S (1999) The TDT-2 text and speech corpus. In: Proceedings of the 1999 DARPA broadcast news workshop, Herndon
9. Deena S, Hasan M, Doulaty M, Saz O, Hain T (2016) Combining feature and model-based adaptation of rnnlms for multi-genre broadcast speech recognition. In: Interspeech'16: proceedings of 17th annual conference of the international speech communication association
10. Dines J, Vepa J, Hain T (2006) The segmentation of multi-channel meeting recordings for automatic speech recognition. In: Proceedings of the 7th annual conference of the international speech communication association (Interspeech), Pittsburgh, pp 1213–1216
11. Federico M, Furini M (2012) An automatic caption alignment mechanism for off-the-shelf speech recognition technologies. *Multimedia Tools and Applications* 72(1):21–40
12. Galliano S, Geoffrois E, Gravier G, Bonastre JF, Mostefa D, Choukri K (2006) Corpus description of the ESTER evaluation campaign for the rich transcription of french broadcast news. In: Proceedings of the 5th international conference on language resources and evaluation (LREC), Genoa, pp 139–142
13. Hain T, Burget L, Dines J, Garner P, Grezl F, Hannani A, Huijbregts M, Karafiat M, Lincoln M, Wan V (2012) Transcribing meetings with the AMIDA systems. *IEEE Trans Audio Speech Lang Process* 20(2):486–498
14. Hain T, Burget L, Dines J, Garner PN, Grezl F, El Hannani A, Huijbregts M, Karafiat M, Lincoln M, Wan V (2012) Transcribing meetings with the AMIDA systems. *IEEE Trans Audio Speech Lang Process* 20(2):486–498
15. Hain T, Christian J, Saz O, Deena S, Hasan M, Ng RWM, Milner R, Doulaty M, Liu Y (2016) webasr 2 – improved cloud based speech technology. In: Interspeech'16: Proceedings of 17th annual conference of the international speech communication association
16. Hain T, Christian J, Saz O, Deena S, Hasan M, Ng RWM, Milner R, Doulaty MM, Liu Y (2016) webASR 2 - improved cloud based speech technology. In: Proceedings of the 17th annual conference of the international speech communication association (Interspeech), San Francisco
17. Hain T, El-Hannani A, Wrigley SN, Wan V (2008) Automatic speech recognition for scientific purposes - webASR. In: Proceedings of the 9th annual conference of the international speech communication association (Interspeech), Brisbane, pp 504–507
18. Hinton G, Deng L, Yu D, Dahl G, rahman Mohamed A, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath T, Kingsbury B (2012) Deep neural networks for acoustic modeling in speech recognition. *Signal Processing Magazine*
19. Hinton G, Deng L, Yu D, Dahl GE, Mohamed A, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN, Kingsbury B (2012) Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Proc Mag* 29(6):82–97
20. Katsamanis A, Black MP, Georgiou PG, Goldstein L, Narayanan S (2011) Sailalign: Robust long speech-text alignment. In: Proceedings of the workshop on new tools and methods for very large scale research in phonetic sciences (VLSP), Philadelphia, pp 44–47
21. Klakow D (1998) Log-linear interpolation of language models. In: Proceedings of the 5th international conference on spoken language processing (ICSLP), Sydney

22. Lanchantin P, Gales MJF, Karanasou P, Liu X, Qian Y, Wang L, Woodland PC, Zhang C (2015) The development of the Cambridge University alignment systems for the multi-genre broadcast challenge. In: Proceedings of the 2015 IEEE automatic speech recognition and understanding (ASRU) Workshop, Scottsdale, pp 647–653
23. Larson M, Anguera X, Reuter T, Jones GJF, Ionescu B, Schedl M, Piatrik T, Hauff C, Soleymani M (eds) (2013) Proceedings of the MediaEval 2013 multimedia benchmark workshop, Barcelona
24. Long Y, Gales MJF, Lanchantin P, Liu X, Seigel MS, Woodland PC (2013) Improving lightly supervised training for broadcast transcription. In: Proceedings of the 14th annual conference of the international speech communication association (Interspeech), Lyon, pp 2187–2191
25. Long Y, Gales MJF, Lanchantin P, Liu X, Seigel MS, Woodland PC (2013) Improving lightly supervised training for broadcast transcriptions. In: Proceedings of the 14th annual conference of the international speech communication association (Interspeech), Lyon, pp 2187–2191
26. Matousek J, Vit J (2012) Improving automatic dubbing with subtitle timing optimisation using video cut detection. In: ICASSP'12: Proc Of IEEE international conference on acoustics, speech and signal processing, pp 2385–2388
27. Mikolov T, Kombrik S, Burget L, Cernocky J, Khudanpur S (2011) Extensions of recurrent neural network language model. In: Proceedings of the 2011 international conference on acoustic, speech and signal processing (ICASSP), Prague, pp 5528–5531
28. Milner R, Saz O, Deena S, Doulaty M, Ng RWM, Hain T (2015) The 2015 sheffield system for longitudinal diarisation of broadcast media. In: Proceedings of the 2015 IEEE automatic speech recognition and understanding (ASRU) Workshop, Scottsdale, pp 632–638
29. Moreno PJ, Joerg CF, Van Thong JM, Glickman O (1998) A recursive algorithm for the forced alignment of very long audio segments. In: Proceedings of the 5th international conference on spoken language processing (ICSLP), Sydney, pp 2711–2714
30. Novak JR, Minematsu N, Hirose K (2012) WSFT-based grapheme-to-phoneme conversion: open source tools for alignment, model-building and decoding. In: Proceedings of the 10th international workshop on finite state methods and natural language processing, San Sebastián
31. Olcoz J, Saz O, Hain T (2016) Error correction in lightly supervised alignment of broadcast subtitles. In: Proceedings of the 17th annual conference of the international speech communication association (Interspeech), San Francisco
32. Pallett D, Fiscus J, Garofalo J (1996) Przybocki, m.: 1995 hub-4 dry run broadcast materials benchmark test. In: Proceedings of 1996 DARPA speech recognition workshop, Harriman
33. Povey D, Ghoshal A, Boulianne G, Burget L, Ondrej G, Nagendra G, Hanneman M, Motlicek P, Yanmin Q, Schwarz P, Silovsky J, Stemmer G, Vesely K (2011) The kaldi speech recognition toolkit. In: Proceedings of the 2011 IEEE automatic speech recognition and understanding (ASRU) workshop, Big Island
34. Ramírez J, Górriz JM, Segura JC (2007) Voice activity detection. fundamentals and speech recognition system robustness. In: Grimm M., Kroschel K. (eds) Robust speech recognition and understanding, chap. 1, I-Tech, Vienna, pp 1–22
35. Ramírez J, Segura JC, Benítez C, de la Torre A, Rubio A (2004) Efficient voice activity detection algorithms using long-term speech information. *Speech Comm* 42(3–4):271–287
36. Richmond K, Clark R, Fitt S (2010) On generating combilex pronunciations via morphological analysis. In: Proceedings of the 11th annual conference of the international speech communication association (Interspeech), Makuhari, pp 1974–1977
37. Ryant N, Liberman M (2013) Speech activity detection on youtube using deep neural networks. In: Proceedings of the 14th annual conference of the international speech communication association (Interspeech), Lyon, pp 728–731
38. Saz O, Doulaty M, Deena S, Milner R, Ng RWM, Hasan M, Liu Y, Hain T (2015) The 2015 sheffield system for transcription of Multi-Genre broadcast media. In: Proceedings of the 2015 IEEE automatic speech recognition and understanding (ASRU) Workshop, Scottsdale, pp 624–631
39. Stan A, Mamiya Y, Yamagishi J, Bell P, Watts O, Clark RAJ, King S (2016) Alisa: an automatic lightly supervised speech segmentation and alignment tool. *Comput Speech Lang* 35:116–133
40. Stolcke A (2002) SRILM – An extensible language modeling toolkit. In: Proceedings of the 7th international conference on spoken language processing (ICSLP), Denver, pp 901–904
41. Wen TH, Heidel A, Lee HY, Tsao Y, Lee LS (2013) Recurrent neural network based personalized language modeling by social network crowdsourcing. In: Proceedings of the 14th annual conference of the international speech communication association (Interspeech), Lyon, pp 2703–2707
42. Williams GF (2009) Online subtitling editorial guidelines v1.1. Tech. rep., British Broadcasting Corporation. http://www.bbc.co.uk/guidelines/futuremedia/accessibility/subtitling_guides/online_sub_editorial_guidelines.vs1.1.pdf

43. Wrigley SN, Hain T (2011) Making an automatic speech recognition service freely available on the web. In: Proceedings of the 12th annual conference of the international speech communication association (Interspeech), Florence, pp 3325–3326
44. Wrigley SN, Hain T (2011) Web-based automatic speech recognition service - webASR. In: Proceedings of the 12th annual conference of the international speech communication association (Interspeech), Florence, pp 3265–3268
45. Zelenak M, Schulz H, Hernando J (2012) Speaker diarization of broadcast news in albayzin 2010 evaluation campaign. EURASIP Journal on Audio Speech and Music Processing 19:1–9
46. Zhang P, Liu Y, Hain T (2014) Semi-supervised DNN training in meeting recognition. In: Proceedings of the 2014 IEEE spoken language technology (SLT) workshop, South Lake Tahoe, pp 141–146



Oscar Saz received his B.Sc. in Telecommunications Engineering from the University of Zaragoza, Spain in 2004 and his Ph.D. in 2009 from the same institution. From 2010 to 2012 he was a Fulbright scholar at Carnegie Mellon University and has been a Research Associate at the University of Sheffield until 2016. His interests are acoustic modelling and adaptation for Automatic Speech Recognition and his most recent publications include articles at Computer, Speech and Language and conference papers at Interspeech and ASRU.



Salil Deena is a Research Associate in the Speech and Hearing (SpandH) group at the University of Sheffield where he has been a member of the Natural Speech Technology (NST) project. He received a PhD in Computer Science from the University of Manchester with a thesis on visual speech synthesis, in 2012. From 2012 to 2014, he worked as Research Engineer at Image Metrics researching Computer vision and Machine Learning techniques for facial analysis and animation. His research interests are in Deep Learning for Automatic Speech Recognition, Computer Vision and Natural Language Processing.



Mortaza Doulaty received his B.Sc. (Hons) and M.Sc. in Computer Science from the University of Tabriz, Iran in 2009 and 2011. He recently received his Ph.D. candidate in Computer Science at the Machine Intelligence for Natural Interfaces (MINI), Speech and Hearing Group (SPandH), University of Sheffield, UK. His research interests are domain discovery and domain adaptation in speech recognition.



Madina Hasan received her B.Sc. in Electronics Engineering from The University of Bahrain, Bahrain 2004. Between 2004 and 2006 She worked as a lecturer at Bahrain Training Institute. Madina received her MSc in Data Communications and Ph.D in Computer Science, with a thesis in Scientific Computation, from The University of Sheffield, Uk 2007 and 2012 respectively. Since 2012 She is a research associate at the University of Sheffield. Her research interests are in applying machine learning techniques for natural language processing problems.

Bilal Khaliq was a Research Associate in the Speech and Hearing (SpandH) group at the University of Sheffield in 2016. He received his PhD in Informatics from the University of Sussex in 2015 and his MPhil in Computer Speech, Text and Internet Technologies from the University of Cambridge in 2009. He is currently employed as Data Scientist at Teradata in Pakistan.



Rosanna Milner received her PhD in Computer Science from the University of Sheffield in 2017. Her work focused on speaker diarisation of broadcast archives under the supervision of Prof. Thomas Hain. She studied my MSc in Computer Science with Speech and Language Processing at the University of Sheffield and achieved her BSc in Mathematics with Linguistics from the University of York.



Raymond W. M. Ng was a research associate with the University of Sheffield until June 2017. His research focussed on spoken language translation, and he was also involved in research related to speaker and language recognition. Raymond received his BEng, MA and PhD from the Chinese University Hong Kong.



Julia Olcoz is research member of the ViVoLab (Voice input Voice output Laboratory) at the University of Zaragoza (UZ), in Zaragoza (Spain). She holds a BSc in Telecommunication Engineering (2011) and a MSc in IT & Mobile Networks (2012), both from UZ. She has been a visiting researcher at the L2f (INESC ID's Spoken Language Systems Laboratory) in Lisbon (Portugal) from September to December 2014, at the MLDC (Microsoft Language Development Center) in Lisbon (Portugal) from January to May 2015, and at the MiNi – SPandH (Machine Intelligence for Natural Interfaces – Speech and Hearing Group) in the University of Sheffield, in Sheffield (UK) from September to December 2015. Her research interests focus on Automatic Speech Recognition (ASR), Acoustic and Language Modeling in ASR, Unsupervised Learning Techniques for ASR, Second Language Learning (L2), among others.



Thomas Hain is Professor for Computer Science at the University of sheffield. He holds the degree 'Dipl.-Ing' in Electrical and Communication Engineering from the University of Technology, Vienna, and a PhD in Information Engineering from Cambridge University (2002). After work at Philips Speech Processing, Vienna he joined the Cambridge University Engineering Department in 1997, where he was appointed to Lecturer in 2001. He then moved to Sheffield in 2004, and was promoted to Professor in 2012. Prof Hain leads the 15-strong subgroup on Machine Intelligence for Natural Interfaces and has more than 140 publications on machine learning and speech recognition topics.