Error correction in lightly supervised alignment of broadcast subtitles

Julia Olcoz[1,2]*, Oscar Saz*[2]*, Thomas Hain*[2]

[1]ViVoLab, Aragon Institute for Engineering Research (I3A), Universidad de Zaragoza, Spain
[2]Speech and Hearing Group, Department of Computer Science, University of Sheffield, UK
jolcoz@unizar.es, {o.saztorralba,t.hain}@sheffield.ac.uk

Abstract

This paper presents a range of error correction techniques aimed at improving the accuracy of a lightly supervised alignment task for broadcast subtitles. Lightly supervised approaches are frequently used in the multimedia domain, either for subtitling purposes or for providing a more reliable source for training speech–based systems. The proposed methods focus on directly correcting of the alignment output using different techniques to infer word insertions and words with inaccurate time boundaries. The features used by the classification models are the outputs from the alignment system, such as confidence measures, and word or segment duration. Experiments in this paper are based on broadcast material provided by the BBC to the Multi–Genre Broadcast (MGB) challenge participants. Results, show that the order alignment F–measure improves up to 2.6% absolute (15.8% relative) when combining insertion and word–boundary correction.

Index Terms: Lightly supervised alignment, broadcast media, broadcast subtitles, regression techniques

1. Introduction

The multimedia broadcast domain is increasingly becoming an important area for research topics related to spoken language technologies. Applications such as automatic transcription of broadcast shows or information retrieval from multimedia data require training of acoustic and language models matched to the specific multimedia scenario. Although large amounts of multi–genre data exist and could be exploited as a rich source of training material, it is usually found that only subtitle transcription exist for this data. However, subtitles present a series of difficulties for use in acoustic model training. On one hand, they include only approximate segment–level timing information. On the other hand, subtitles always digress from the actual speech, either as a result of paraphrasing in manual subtitling, or due to errors in automatic subtitling. Hence, the task of aligning an unreliable text to audio has significant relevance to work in the broadcast media domain.

As standard Viterbi–based forced alignment is usually not appropriate for very long segments, lightly supervised approaches [1] are commonly employed. For those scenarios where transcriptions are incomplete, [2] and [3] propose the use of large background acoustic and language models, and [4] implements a method for sentence–level alignment based on grapheme acoustic models. Moreover, if transcript quality is very poor [5] presents an alternative to improve lightly supervised decoding using phone level mismatch information. Other related works such as [6] take also into account situations where transcripts include a mixture of languages.

To provide a benchmark that allows researchers to compare and improve their systems, we proposed lightly supervised alignment to be studied in the context of the Multi-Genre Broadcast (MGB) challenge [7]. The challenge explored spoken language systems performance for general broadcast media, following the steps of previous evaluations. The Hub4 series [8] was organised by NIST to evaluate broadcast news transcription in English. The ESTER campaign [9] studied rich transcription of French broadcast news, while the Albayzin campaigns [10] did the same for audio processing of Spanish broadcast news. The MediaEval campaigns [11] include tasks in automatic retrieval and classification of broadcast data in several languages.

Trying to overcome the lack of complete and reliable reference subtitles, we have addressed the problem by manipulating the lightly supervised alignment output. Such an approach required to study the type of errors that occur in a lightly supervised alignment system. With that knowledge, one can develop a post–processing stage that aims to amend errors in the alignment process. Standard regression and classification techniques based on the boundary correction models presented in [12] are applied to correct word insertions and word–boundary errors, considering different architectures and sets of features.

The rest of the paper is organised as follows: Section 2 describes the state–of–the–art of systems used in lightly supervised alignment tasks, while Section 3 focuses on the post-processing stages developed to correct the alignment output. Section 4 describes the experimental data used and details the baseline systems built, with Section 5 providing analysis of the errors in the alignment output and giving upper bounds for the error correction systems. Finally, Section 6 focuses on the results achieved when correcting alignment errors, and Section 7 concludes this work and proposes future research.

2. Lightly supervised alignment

A state–of–the–art lightly supervised alignment system [13, 3, 14, 4], typically has a structure as the one shown in Figure 1. The input audio is first processed by Voice Activity Detection (VAD) that identifies boundaries for the segments where speech is present. Meanwhile, standard text normalisation and tokenisation are applied to the input subtitles. A language model adapted to these subtitles is then trained. This can be achieved by interpolating a subtitle–only language model with a larger background language model [14], or by creating word networks constrained by the subtitles [4]. Next, an ASR decoding stage processes the speech segments as obtained previously. Since an adapted language model is used in this decoding, this stage is usually referred to as lightly supervised decoding [13]. Such decoding process can be as complex as necessary, including the use of multiple decoding passes and speaker adaptation. The transcript hypothesis given by the lightly supervised decoding stage is then aligned to the original input text. This can usually be performed via recursive dynamic programming approaches where sequences of words from the subtitles are assigned to

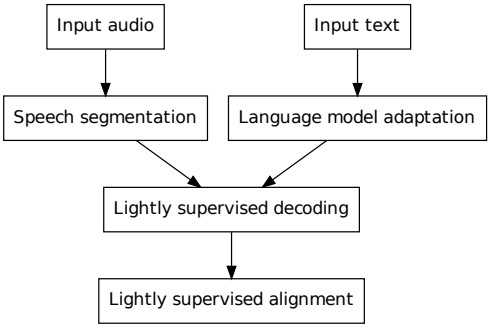

Figure 1: *A standard approach to lightly supervised alignment.*

the speech segments, based on how well they match the output of the lightly supervised decoding, for instance, using the Levenshtein distance [3]. The output of the lightly supervised alignment is a set of speech segments whose transcripts contain words from the original subtitle text. At this stage, a second alignment can be used to provide precise word–boundaries for this output and the process is complete.

This process is liable to several organic errors. First, the system only outputs words contained in the original subtitles, i. e., words not in the subtitles can not be recovered. Second, the system aims to output all words in the original subtitles, so words not present in the audio may appear in the output. And, third, the word–boundaries rely on the quality of the decoding and alignment models used and will not always be accurate. Error recovery stages can be added to the diagram in Figure 1, for instance based on further ASR decodings [15]. This paper proposes new work on the error recovery strategy.

3. Error correction proposal

With the aim of improving the accuracy of lightly supervised alignment systems this paper proposes applying at its output regression techniques to correct misalignments. This approach focuses on recovering two types of errors as mentioned in Section 2: insertions and word–boundary errors. Figure 2 shows a diagram including post–processing stages designed to correct both types of errors. The process takes as input the outcome of a lightly supervised alignment system, as the one in Figure 1. It starts by performing Viterbi forced alignment to provide time–boundaries for the lightly supervised aligned words. At this stage, several instances of Viterbi alignment can be obtained, with different acoustic models, with the purpose of providing distinct alignments. Next, a confidence measure score is computed for each word and segment in the Viterbi aligned output(s). Then, a classification model trained to identify which words in the aligned output are not present in the spoken audio is applied and such words are removed. The final step uses a regression model to provide an estimation of how inaccurate the time boundaries of the words are from one or multiple Viterbi alignments, which is used to correct the word boundaries.

Figure 2 depicts a generic system where error correction can be implemented in different ways. Viterbi alignment and confidence measure calculation can be done with different acoustic models and approaches. The error correction modules can generally employ any type of classification and regression with the desired features of choice. Section 4.2 will discuss the specific choice of models and features used in this paper.

Figure 2 can also lead to simpler or complex error correction configurations by removing some stages from the general system. If insertions are the only target of the correction system, extra Viterbi alignment stages can be discarded, as well as the

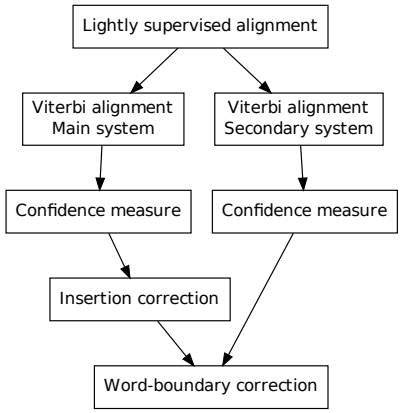

Figure 2: *Post–processing error correction diagram.*

word–boundary correction stage. If only word–boundaries are to be corrected, the insertion correction stage can be removed, and one or several Viterbi alignments can be used (Figure 2 generalises to 1, 2, 3 or more alignments). Different Viterbi alignments are liable to produce distinct word–boundary hypotheses depending on the the type of model used, the features it uses or the input data they were trained on. The use of several hypothesis from multiple Viterbi alignment systems is used in Figure 2 as a way of improving the final boundary estimation. All of these possible subsystems, including the complete system, will be evaluated in Section 6.

4. Experimental data and setup

The experiments for this paper are based on the setup for Task 2 of the MGB challenge [7]: *Alignment of broadcast audio to a subtitle file, i. e., lightly supervised alignment*. The data for this goal contained several shows broadcast by the BBC during 2008. The number of shows and the amount of audio for the training, development and evaluation sets were as shown in Table 1. The 1,500 hours of raw audio were the only acoustic data that could be used for training according to the MGB rules. Although no true transcripts were available, lightly supervised transcripts were given to facilitate training of models. Other data provided within the MGB challenge were 640 million words of subtitle text corresponding to shows broadcast from the 1970s to 2008. These subtitles were the only linguistic data available for training language models.

Table 1: *Data for the lightly supervised alignment task.*

Dataset	Number of Shows	Broadcast Time
Training	2,193	1,580.4 h.
Development	47	28.4 h.
Evaluation	16	11.2 h.

4.1. Baseline system

The system used for lightly supervised alignment was as shown in Figure 3. It followed the diagram in Figure 1 in consisting of a first stage of lightly supervised decoding, followed by lightly supervised alignment. The lightly supervised decoding stage operated as follows: first, a DNN–based speech segmentation module identified segments of speech in the show. An initial transcription for these segments was obtained from a speaker independent DNN–HMM system [16] trained on 700 hours of acoustic training data using the Kaldi toolkit [17]. This stage used a background language model trained on the subtitle data using SRILM [18]. This output was used to provide re–segmentation, speaker clustering and speaker adaptation to the

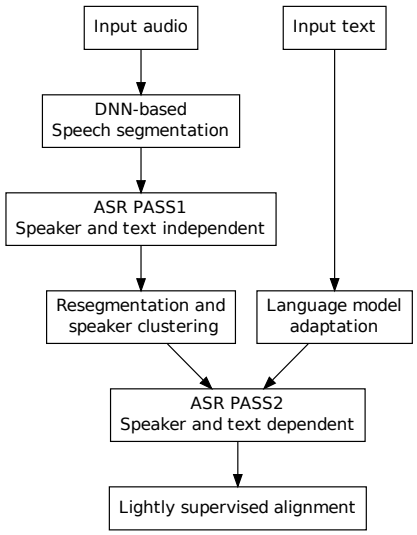

Figure 3: *System for lightly supervised alignment experiments.*

second decoding stage. This decoding was based on a DNN–GMM–HMM system trained on 700 hours of speech using TNet [19] and HTK [20]. This second stage used a language model interpolated from the background language model and the subtitles for the show using the SRILM toolkit [18]. Finally, the output was aligned to the subtitles in a recursive lightly supervised alignment stage. The modules for this baseline system were also part of the systems submitted by the University of Sheffield to Tasks 1 and 3 (transcription and diarisation) of the MGB challenge and further description can be found in their respective system description [21, 22].

The results achieved by this system in the development and evaluation sets of the MGB Task 2 are shown in Table 2. These results were obtained using the official MGB Task 2 scoring script. Performance is assessed in terms of F–measure, calculated as the geometric mean of the precision and recall of the system. For this task, precision is calculated as the number of words correct in the hypothesis, over the total number of words in the hypothesis. Recall is calculated as the number of words correct in the hypothesis, over the total number of words in the manual reference. A word in the hypothesis is considered correct if it matches the same word in the reference with a boundary error of up to 100 milliseconds (10 frames). Only words appearing in the original subtitles were considered for scoring. The results show that the baseline system performed better in terms of recall, and that the F–measure achieved varied from 0.85 in the development data to 0.83 in the evaluation data.

Table 2: *Lightly supervised alignment baseline results.*

Dataset	Precision	Recall	F–measure
Development	0.8451	0.8655	0.8551
Evaluation	0.8179	0.8559	0.8365

4.2. Alignment correction models

The implementation of the error correction strategy as proposed in Figure 2 used two different DNN–GMM–HMM systems for producing a Viterbi forced alignment of the baseline lightly supervised alignment output. The main system was trained on 700 hours of speech with sequence–trained DNN bottleneck features, while the secondary system was trained on 500 hours of speech with cross–entropy trained bottlenecks [21]. The confidence measure estimation was based on DNN posteriors [23]. Finally, the error correction modules were based on regression trees using the scikit–learn machine learning toolkit [24]. Regression trees were successfully used before by [12] to correct phone–boundary errors in the TIMIT database [25].

The set of features used in the regression trees were chosen from the following: confidence score of the word, duration of the word, number of phones in the word, confidence measure of the segment in which the word is included, duration of the segment, and number of words in the segment. The features to be used, as well as other hyperparameters for training of the regression trees (tree depth and minimum numbers of examples per leaf) were optimised on the development set. The predicted variable depends on the kind of error to recover. In case of insertions, it is a value between 0 and 1, where a low value indicates a higher confidence of the word being an insertion. When dealing with word–boundaries, it is an unbounded integer that refers to the signed deviation in frames between Viterbi forced alignment and manual reference time–stamps. Two different regression trees are required to infer the boundary correction, one for the beginning and one for the end of the word respectively.

5. Error analysis and oracle systems

Taking into account the baseline results presented in Table 2 for the development dataset, it was first studied which types of errors occurred in the lightly supervised alignment output. Four categories of errors were observed: deletions (*D*, words in subtitles and in manual reference but not in the alignment output), insertions (*I*, words in the alignment output and in the subtitles but not in manual reference) and word–boundary errors (*T1* and *T2*, differentiating whether the words in the manual reference and the alignment output present some overlap in time or not). Table 3 presents the number of deletions, insertions and word–boundary errors found in the output of the baseline system considering the development set.

Table 3: *Number of errors per category in the lightly supervised alignment system in the development set.*

	D	I	T1	T2
Number of errors	8,875	11,842	3,429	6,980

It can be seen that the most common errors are insertions (11,842), followed by word–boundary errors (10,409 adding *T1* and *T2*) and finally deletions are the least common (8,875). Using these oracle outputs, regression trees were trained to predict insertions and boundary errors. When dealing with word–boundary errors, sparsity in the examples used for training becomes an issue. Figure 4 shows the distribution of words with boundary errors in the baseline system regarding the distance in frames from the manual reference to the alignment output (from 11 up to 100 frames). It can be seen that there were very few examples of words with a given value of frame distance from the reference to the alignment output. Word–boundary errors of 10 or less are not errors in the task and were not studied.

At this stage, oracle error correction stages were designed in order to provide upper boundaries to the error correction system presented. The oracle output was obtained by manually removing a specific category of errors in the baseline system output. The results in the development set are shown in Table 4. The largest improvement in F–measure provided by correcting a single type of error is 0.0646 with the correction of word–boundary errors (*T*), due to an increase in both precision and recall. Correcting all deletion errors (*D*) highly increases the recall, but the global F–measure only improves by 0.0270. Insertion correction (*I*) produces a large effect in precision improvement with an improvement in F–measure of 0.0356. Finally, the

Table 5: *Lightly supervised alignment error correction results.*

	Development set				Evaluation set			
Correction stage	Precision	Recall	F–measure	Rel. Impr.	Precision	Recall	F–measure	Rel. Impr.
I	0.9135	0.8537	0.8826	18.6%	0.8888	0.8418	0.8647	17.2%
T (single Viterbi)	0.8476	0.8671	0.8573	1.5%	0.8209	0.8579	0.8390	1.5%
T (two Viterbis)	0.8478	0.8661	0.8568	1.2%	0.8178	0.8541	0.8355	-0.6%
I+T (single Viterbi)	0.9018	0.8624	0.8817	18.4%	0.8732	0.8518	0.8624	15.8%
I+T (two Viterbis)	0.9010	0.8621	0.8811	17.9%	0.8723	0.8511	0.8616	15.4%

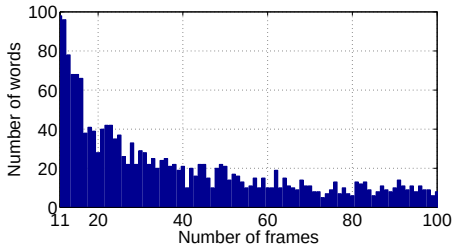

Figure 4: *Distribution of words with incorrect boundaries in the lightly supervised alignment system in the development set.*

combined correction of *I* and *T*, the target of the experiments, has an upper bound of 0.9586 in F–measure, given by the oracle results. The results are also expressed as relative improvement over the baseline, using an F–measure of 1.0 as the upper value.

$$Rel.Impr. = \frac{F_{system} - F_{baseline}}{1.0 - F_{baseline}} \quad (1)$$

Table 4: *Results and relative improvement over the baseline using oracle error correction on the development set.*

Error	Precision	Recall	F–measure	Rel. Impr.
D	0.8480	0.9191	0.8821	18.6%
I	0.9165	0.8663	0.8907	24.6%
T	0.9089	0.9308	0.9197	44.6%
I and *T*	0.9864	0.9323	0.9586	71.4%

6. Results

A complete set of experiments were performed to optimise hyperparameters in the regression models used. The optimal value for tree depth was found to be 10, and the optimal value for minimum samples per leaf was found to be 500 for *I* error correction, and 1000 for *T* recovering. Regarding features, presented in Section 4.2, the optimal set was found to be the 5–feature set for the *I* correction, i. e., confidence score, duration and number of phones of the word and confidence measure and duration of the segment of the word, and the 3–feature set for the *T* correction, this is, confidence score, duration and number of phones of the word. Finally, a threshold had to be found in order to decide whether a word was removed or not following the insertion correction. The optimal threshold value was found to be 0.4. Words with regression output of 0.4 or lower were discarded from the lightly supervised output.

Table 5 presents the results for different error correction scenarios as described in Section 3. The best result on the development set was obtained when using only insertion correction, with an F–measure improvement of 0.0275. This was an 18.6% relative improvement over the baseline and it improved 77.2% compared to the upper bound given by the oracle (0.8907). The improvement obtained by correcting only word–boundary errors was less significant (0.0022, or 1.5% relative), and the combination of insertion and word–boundary correction produced smaller gains than insertion error correction alone (0.0266 or 18.4% relative). Furthermore, the use of two Viterbi alignment systems produced a very small drop in gain compared to using a single Viterbi system.

A similar pattern of results was observed on the evaluation set, also shown in Table 5. F–measure values indicated the difficulty of correcting errors of the word boundaries, where only small gains of 1.5% relative were observed. As shown when studying the distribution of boundary errors in words, it is possible that regression trees were not the most appropriate model for this task. Nevertheless, regression trees were shown to give significant gains in the detection of insertions (0.0282, or 17.2% relative). The combination of both systems produced smaller gains (0.0259 or 15.8% relative). The use of multiple Viterbi alignments to improve the correction of word boundaries was found to produce very small change in the results. This can be due to the shown weakness of regression trees for this task; also the two systems used for Viterbi alignment might not have been distinct enough to provide a gain to each other, since both were trained on MGB acoustic model training data.

7. Conclusions and future work

This paper presented a set of error correction methods to be considered for post–processing of the output of a lightly supervised alignment system. Insertions and word–boundary errors were amended by applying regression and classification techniques based on boundary correction models. Results shown a large improvement when using this setup to remove insertions, with 17–18% relative improvement in F–measure, and up to 77% of errors recovered compared to a manual oracle. However, word–boundary error correction became a more challenging task due to the sparsity of input data to the correction models training. Although small gains of 1.5% relative were achieved, the combination of insertion and word–boundary did not manage to improve over the insertion–only correction results.

Future research could focus on considering new sets of input features, including some information about neighbouring words and silences, as well as using confidence scores obtained applying newer techniques. Alternative regression approaches for word–boundary correction might help improve the performance of this system. These improved alignment output can then be used to train better acoustic models or to improve the output of automatic captioning systems.

8. 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 system outputs and scoring results are available with DOI 10.15131/shef.data.3437426

9. Acknowledgements

Julia Olcoz is supported by the Spanish Government and the European Union under BES–2012–056894 FPI Grant and project TIN2014–54288–C4–2–R, and also by the European Commission FP7 IAPP Marie Curie Action GA–610986. Oscar Saz and Thomas Hain are supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

10. References

[1] P. J. Moreno, C. F. Joerg, J.-M. Van Thong, and O. Glickman, "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, Australia, 1998, pp. 2711–2714.

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

[3] A. Katsamanis, M. Black, P. Georgiou, L. Goldstein, and S. Narayanan, "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, PA, 2011, pp. 44–47.

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

[5] Y. Long, M. J. Gales, P. Lanchantin, X. Liu, M. S. Seigel, and P. C. Woodland, "Improving lightly supervised training for broadcast transcription," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech)*, Lyon, France, 2013, pp. 2187–2191.

[6] G. Bordel, M. Peñagarikano, L. J. Rodríguez-Fuentes, and A. Varona, "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, OR, 2012, pp. 1840–1843.

[7] P. Bell, M. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, and M. Wester, "The MGB Challenge: Evaluating multi-genre broadcast media recognition," Scottsdale, AZ, 2015, pp. 687–694.

[8] D. Pallett, J. Fiscus, J. Garofalo, and M. Przybocki, "1995 Hub–4 Dry Run Broadcast Materials Benchmark Test," in *Proceedings of 1996 DARPA Speech Recognition Workshop*, Harriman, NY, 1996.

[9] S. Galliano, E. Geoffrois, G. Gravier, J. F. Bonastre, D. Mostefa, and K. Choukri, "Corpus Description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News," in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, 2006, pp. 139–142.

[10] D. Castán, D. Tavarez, P. Lopez-Otero, J. Franco-Pedroso, H. Delgado, E. Navas, L. Docio-Fernández, D. Ramos, J. Serrano, A. Ortega *et al.*, "Albayzín-2014 evaluation: audio segmentation and classification in broadcast news domains," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–9, 2015.

[11] M. Larson, X. Anguera, T. Reuter, G. Jones, B. Ionescu, M. Schedl, T. Piatrik, C. Hauff, and M. Soleymani, Eds., *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain, 2013.

[12] A. Stolcke, N. Ryant, V. Mitra, J. Yuan, W. Wang, and M. Liberman, "Highly accurate phonetic segmentation using boundary correction models and system fusion," in *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 5552–5556.

[13] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech & Language*, vol. 16, no. 1, pp. 115–129, 2002.

[14] P. Lanchantin, M. J. Gales, P. Karanasou, X. Liu, Y. Qian, L. Wang, P. C. Woodland, and C. Zhang, "The Development of the Cambridge University Alignment Systems for the Multi-Genre Broadcast Challenge," Scottsdale, AZ, 2015, pp. 647–654.

[15] T. J. Hazen, "Automatic alignment and error correction of human generated transcripts for long speech recordings," in *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech–ICSLP)*, Pittsburgh, PA, 2006, pp. 1606–1609.

[16] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Processing Magazine*, vol. 29 (6), no. 6, pp. 82–97, 2012.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," 2011.

[18] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, Denver, CO, 2002, pp. 901–904.

[19] K. Vesely, L. Burget, and F. Grezl, "Parallel Training of Neural Networks for Speech Recognition," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech)*, Makuhari, Japan, 2010, pp. 2934–2937.

[20] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. J. Odell, D. G. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book version 3.4*, Cambridge, UK, 2006.

[21] O. Saz, M. Doulaty, S. Deena, R. Milner, R. W. M. Ng, M. Hasan, Y. Liu, and T. Hain, "The 2015 Sheffield System for Transcription of Multi–Genre Broadcast Media," Scottsdale, AZ, 2015, pp. 624–631.

[22] R. Milner, O. Saz, S. Deena, R. W. M. N. M. Doulaty, and T. Hain, "The 2015 Sheffield system for longitudinal diarisation of broadcast media," Scottsdale, AZ, 2015, pp. 632–639.

[23] P. Zhang, Y. Liu, and T. Hain, "Semi–Supervised DNN Training in Meeting Recognition," in *Proceedings of the 2014 IEEE Spoken Language Technology (SLT) Workshop*, South Lake Tahoe, CA, 2014, pp. 141–146.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[25] J. Garofalo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1," Philadelphia, PA: Linguistic Data Consortium, 1993.