Augmenting ASR for user-generated videos with semi-supervised training and acoustic model adaptation for Spoken Content Retrieval

Yasufumi Moriya and Gareth. J. F. Jones

ADAPT Centre, School of Computing, Dublin City University, Dublin 9, Ireland
{yasufumi.moriya,gareth.jones}@adaptcentre.ie

Abstract. We present an investigation into the use of semi-supervised training and content genre adaptation for improved automatic speech recognition (ASR) of diverse user-generated videos in the task of spoken content retrieval (SCR). Previous work has successfully applied semi-supervised training in single domain ASR tasks. Our focus is on the exploration of the effective use of semi-supervised training of ASR systems for transcription of the spoken content stream of user-generated video data in varied domains and acoustic noise conditions for use in SCR systems. We examine all elements of ASR system development including: data segmentation, data selection, genre labels, acoustic modelling and language modelling using semi-supervised training. We evaluate its effectiveness for ASR and a known-item SCR task using the Blip100000 multimedia collection. Our baseline hybrid ASR system trained out-of-domain produced WERs 31.27% and 44.69% on dev and test sets, respectively. By introducing the techniques outlined above, the WERs are reduced to 26.82% and 39.21% respectively. The improved transcripts increased mean reciprocal rank (MRR) results for the SCR task from 15.59% to 39.38% on dev and 20.98% to 37.23% on test sets.

Keywords: spoken content retrieval · speech recognition · user generated data · semi-supervised training · content genre adaptation

1 Introduction

The growing amount of digital multimedia content such as user-generated videos and podcasts, now widely available on the Internet, is increasing the importance of effective Spoken Content Retrieval (SCR) systems. SCR systems generally operate using speech transcripts created using automatic speech recognition (ASR). It is known that SCR effectiveness is generally impacted by high word error rates (WERs), e.g greater than 30% [4]. High error rates in speech transcripts can cause a "mismatch" between user search queries and document transcripts, even if the documents are highly relevant to the queries. State-of-the-art ASR systems show very low WERs for well controlled transcription tasks such as for the Wall Street Journal (WSJ) and LibriSpeech corpora [2,5]. However, transcription of uncontrolled, highly varied user-generated speech remains a challenging ASR task often with high WERs.

The key challenges of transcribing user-generated spoken video arises from the highly varied speaker characteristics (adult, child and non-native speaker), speaking styles (scripted, formal and informal interviews, sports and video game commentary, and casual conversations), and acoustic conditions (background music, loud audience, applause, and street noise). The goal of our work reported here is to develop a multi-domain ASR system suitable for such challenging user-generated videos, and evaluate its effectiveness both in terms of WER and for an SCR task.

In this paper, we investigate the use of semi-supervised training and video genre tag for development of multi-domain ASR. While past work on multi-domain ASR [8] assumed the existence of manual transcripts or user-uploaded captions, our work focuses on the use of untranscribed speech for semi-supervised training. Further, our ASR system trained using a semi-supervised approach is evaluated in the context of SCR, which has not been the end goal of existing research on semi-supervised ASR training [7,19]. In semi-supervised settings, a seed ASR system is used to generate transcripts or decoded lattices of the data from which a new ASR system is created. The absence of a requirement for manually transcribed training data makes semi-supervised training an attractive option for addressing the challenges of transcribing diverse user-generated videos. User-generated videos are often accompanied by a video genre tag uploaded by the user. We exploit the genre tag information for acoustic model adaptation [1,15]. Our hypothesis is that user-generated videos with the same genre tag will tend to contain the same type of acoustic events (e.g., applauding in "conference" and loud audience in "sports").

The rest of the paper is organised as follows. Section 2 examines ASR system development using untranscribed data. Section 3 presents acoustic model adaptation using content genre, followed by experimental investigations in Section 4 and Section 5. Section 6 provides concluding remarks of this paper.

2 Semi-supervised Acoustic and Language Modelling

The overall goal of our investigation is to develop an ASR system which improves SCR effectiveness. We hypothesize that reducing WERs in ASR by optimising data segmentation, data selection, acoustic modelling and language modelling using untranscribed data can help to achieve this. Figure 1 shows a flowchart for the application of the semi-supervised approach to acoustic model and language model training in ASR.

Data Segmentation Natural audio generally consists of a mixture of speech utterances and other audio activities. The curated speech corpora used in existing ASR research are typically pre-segmented into speech utterances. We wish to make use of untranscribed raw video data in our work in which boundaries between speech and other audio data are unknown. We thus seek to identify regions of speech using VAD and use this to segment the video data. To explore

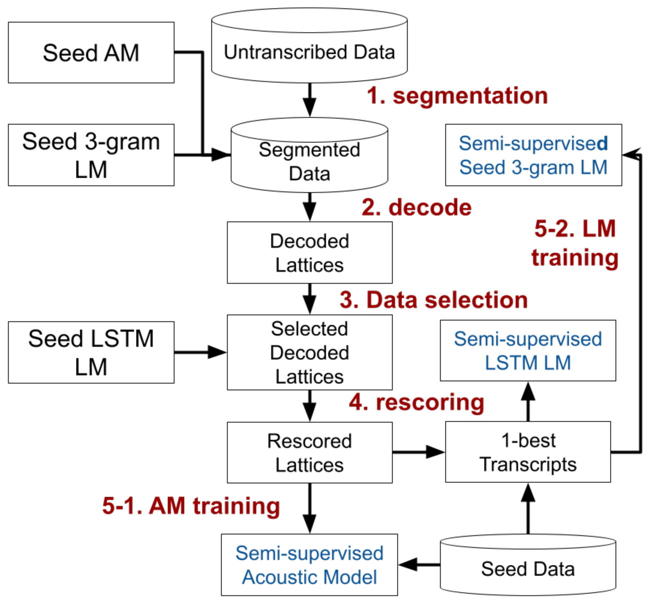

Fig. 1. Flowchart of semi-supervised training for acoustic and language models.

the usefulness of this VAD step, we compare VAD segmented data with simple equal sized data segments. To do this, we compare ASR WER for a system trained on VAD segments with one trained on equal sized segments.

Data Selection Data selection is used to select segments of untranscribed data which are likely to lead to sufficiently accurate ASR for semi-supervised training of an acoustic model and a language model. For our investigation, we use the segment level confidence score described in [22]. This is computed by taking the average of posterior probabilities of speech segments decoded by a seed ASR system. Either VAD or equally segmented untranscribed data below the set confidence level is then excluded from the training data.

Acoustic Model For acoustic model training using untranscribed data, we use the recently proposed semi-supervised lattice-free maximum mutual information (LF-MMI) training method [7]. LF-MMI is discriminative training method where the training objective is to predict a sequence of phone labels as a whole, rather than individual phones in an utterance. In semi-supervised settings, several paths of a decoded lattice of untranscribed data are considered to be the target. When these paths contain lower probabilities (i.e., the seed system is not confident in its prediction), its impact on a new acoustic model is smaller.

Language Model As shown in Figure 1, a seed n-gram language model is used to decode untranscribed data and a seed LSTM language model is used to re-score decoded lattices of untranscribed data [21]. These language models are trained on manual transcripts of a speech corpus. We generate a 1-best transcript of the untranscribed data, and train new n-gram and LSTM language models on a combination of manual transcripts of the seed data with ASR transcripts of untranscribed data. We examine the benefits of incorporating ASR transcripts from varied domains in language model training. Little existing work has studied semi-supervised training of an LSTM language models for lattice re-scoring.

3 Adaptation of Acoustic Model using Content Genre

As outlined in Section 1, we hypothesize that user-generated content with the same genre tag (e.g., "conference") will contain similar types of acoustic activities ("applause"). By providing an acoustic model with content genre, it may become more robust to acoustic information contained in a given content genre. We propose two approaches for adapting an acoustic model using content genre: genre code and genre embedding. The core idea is to transform a user-provided genre tag into a single digit (genre code) or into an embedding vector (genre embedding) extracted from an acoustic feature using a genre classification network for the input of an acoustic DNN model.

Figure 2 shows extraction of content genre information from user-generated video data. A genre code is generated by converting each of the unique genre tags to a digit. This is similar to the domain ID used in [13], however, we apply genre codes in the semi-supervised settings. This is treated as adaptation information and concatenated with an acoustic feature vector.

A more sophisticated approach is inspired by an x-vector system designed for speaker recognition [17]. The x-vector, which is useful to identify a speaker from speech, can be extracted from a DNN model trained to classify a speaker given an acoustic feature vector. The x-vector extractor consists of five layers operated on several speech frames with specified context, followed by a statistics pooling layer and two layers operated on speech segments. The x-vector is extracted from the first segment layer of the extractor. To extract genre embedding, we train a DNN model with the same structure as the x-vector extractor which can classify content genre given an acoustic feature vector, genre embedding can be extracted from the first segment layer of the extractor. This embedding is concatenated with an acoustic feature vector as input to an acoustic model.

4 ASR Experiments

4.1 Creation of Manual Transcripts for Blip10000

To evaluate our proposed use of semi-supervised training and content genre adaptation on diverse user-generated videos we use the Blip10000 corpus [16]. The Blip10000 corpus is a collection of user-generated videos of diverse qualities and genres crawled from the internet. It contains 14,838 videos (3,288 hours) released under Creative Commons. The corpus is partitioned into 5,288 videos for dev set and 9,550 videos for test set. The Blip10000 corpus contains videos of 26 different genres; its content includes materials such as vlogs, conferences,

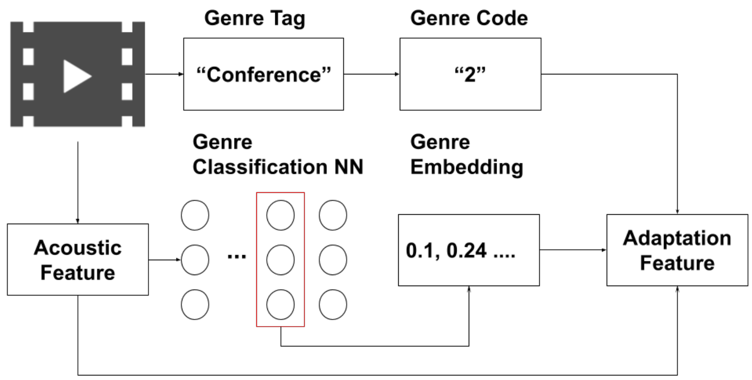

Fig. 2. Generation steps of genre code and genre embedding for content genre adaptation of an acoustic model.

street interviews, semi-professional broadcasts, technology reviews and so on. The spoken language is mainly English, but non-English videos can be found. To use blip10000 for ASR research, we created manual transcripts of a subset of data as follows: 670 videos of dev set (20 hours) and 566 videos of test set (15 hours)[1]. This amount of manually transcribed data enables us to study ASR behaviour on a much wider range of data than is typically the case in ASR research. Videos were selected from shorter ones to increase the number of documents and the diversity of content. The selected videos were manually transcribed by crowd-sourcing using Amazon Mechanical Turk (AMT).

4.2 Experimental Setup

Baseline Systems We built two baseline systems: a hybrid DNN-HMM system using Kaldi [10] and an end-to-end ASR system using espresso [20].

The hybrid system consists of a DNN acoustic model, an n-gram language model and an LSTM language model for lattice re-scoring. The acoustic DNN model consists of 17 time-delayed layers with 1,024 units each trained using LF-MMI [11]. The n-gram is a 3-gram model built using the SriLM toolkit [18]. The LSTM language model consists of 2 layers of LSTM layers with 256 units each. For the hybrid system, we trained one acoustic model on approximately 500 hours of How2 data containing instruction videos [14] and another model on 960 hours of LibriSpeech audio-book data [9]. The How2 data is a collection of user-generated instructional videos, its acoustic conditions are more similar to those of Blip10000 than LibriSpeech. Nevertheless, the domain of How2 videos is limited to instruction and most of the How2 videos contain a single speaker, whereas the number of speakers varies in the Blip10000 data. The n-gram and

[1] We plan to make these manual transcripts publicly available.

LSTM language models were trained on manual transcripts of How2 and 960 hours transcripts of the LibriSpeech corpus.

The end-to-end ASR system is an encoder-decoder architecture with 4 convolutional layers followed by 4 LSTM layers as an encoder and 4 LSTM layers as a decoder. The sub-word language model [3] was incorporated into the end-to-end system using shallow fusion. Similar to the hybrid system, the end-to-end system was trained on How2 data and the sub-word language model was trained on transcripts of How2 and LibrisSpeech. The vocabulary size was approximately 100,000 words.

The How2 test set consisting of roughly 5 hours of data was used to evaluate the hybrid DNN-HMM and the end-to-end ASR trained on How2 data.

Semi-supervised Training For semi-supervised training, the ASR system was trained on 500 hours of manually transcribed How2 data combined with untranscribed Blip10000 dev data consisting of 1,050 hours of data. For VAD, we used the NeMo toolkit [6] trained on the Google Speech Commands and Freesound datasets. This tool is claimed to classify speech and non-speech frames with 99% accuracy[2]. Untranscribed data were split into segments when non-speech frames were longer than 2 seconds. We added 0.5 of non-speech frames to the beginning and end of each segment to avoid abrupt cut-offs. Equal segments were created by segmenting untranscribed data into 30 second chunks with 5 second overlap with adjacent segments. Segments of 30 seconds were quick to process with the seed system, but 5 seconds of overlap ensures no abrupt cut-offs. We found that rejecting segments of untranscribed data with confidence score lower than 80% was the most effective both for VAD and equal segmentation.

For content genre adaptation of the acoustic model, we generated a genre code by transforming a genre tag of each Blip10000 video into a unique digit (e.g., 1: "technology", 2: "documentary"). Since How2 videos are not classified into different genres and all are instruction videos, How2 speech segments share the same genre code (i.e., 0). In the case of genre embedding, the genre tags were used for supervised training of an genre embedding extractor which classified a genre tag given acoustic features. Following the original paper on xvector [17], we trained a DNN genre embedding extractor with the embedding size set to 512. Genre embedding was the output of the first segment layer of the extractor. Along with the untranscribed dev set of Blip10000, audio from How2 data was added to the training data of the extractor. Acoustic features used in this paper are 40 dimensional MFCCs.

4.3 Experimental Results

Baseline Results Table 1 shows WER results on the How2 test set and on the Blip10000 dev and test sets for the baseline hybrid ASR system and the end-to-end system. The results show that an acoustic model trained on How2 (instruction videos) is more suitable than LibriSpeech (audiobooks). There is a relatively small difference of 3.5% in WERs between the hybrid and the end-to-end system

[2] `https://ngc.nvidia.com/catalog/models/nvidia:vad_matchboxnet_3x1x1`

Table 1. WERs of the baseline hybrid DNN-HMM systems and the end-to-end system on How2 test set and Blip10000. (lr) the seed LSTM LM re-scored lattices of the Blip10000 dev and test sets.

	how2 test	blip dev	blip test
hybrid How2	13.19	31.27	44.69
hybrid How2(lr)	N/A	28.92	42.23
hybrid Libri	N/A	35.94	51.42
end2end	16.77	63.05	78.23

Table 2. WER results of semi-supervised training for acoustic and language modelling. The top three rows show results of semi-supervised acoustic model, seed LSTM re-scoring lattices of untranscribed data and seed LSTM re-scoring lattices of evaluation data. The bottom two rows show results of enhancing n-gram and LSTM language model with automatic transcripts of untranscribed data.

	dev		test	
	eq	vad	eq	vad
AM-semisup	29.85	30.06	42.65	43.02
+seedLSTM-rescore	29.54	29.59	42.39	42.50
+seedLSTM-eval	27.70	27.67	40.44	40.42
seedLSTM-rescore+ngram	28.99	28.89	41.61	41.67
+semisupLSTM-eval	**27.28**	**27.07**	**39.68**	**39.71**

on the How2 test set. However, there is a much larger gap between the hybrid and end-to-end systems between the Blip10000 dev and test set. Two explanations are as follows. Firstly, the How2 data is spoken videos, but its domain is limited to instructional videos. Both the ASR systems experienced worse WERs on the Blip10000 dev and test due to the systems being trained on the out-of-domain data. Secondly, the end-to-end system produced much worse results than the hybrid system, most likely because it requires a much greater amount of training data to achieve satisfactory recognition output. For example, recent work on an end-to-end system used 162,000 hours of transcribed data or data with user-uploaded captions which is roughly equal to 18.5 years of data [8]. For the remainder of our experiments, our baseline system is the hybrid system trained on How2 audio, since this system produced the best WERs. All results after this section are only evaluated on the transcribed Blip10000 dataset.

Semi-supervised Training Results Table 2 shows WER results achieved using semi-supervised training for acoustic and language models. Note that lattice re-scoring can be applied to lattices of untranscribed data and to lattices of evaluation data. Training an acoustic model on re-scored lattices of untranscribed data led to a 0.3-0.5% reduction in WERs. Re-training the n-gram on manual transcripts and 1-best transcripts of untranscribed data led to 0.3-0.8% reduction in WERs. Applying the new LSTM language model to re-scoring the lattices of the evaluation set produced further gain in WERs.

We found that there was no difference between the ASR systems trained on equal segments and VAD segments. This was surprising since applying VAD for

Table 3. WER results of semi-supervised trained systems using content genre adaptation. Top rows (1–3) show WERs without lattice rescoring by the seed LSTM language model, while middle rows (4–6) show WERs by the seed LSTM language model applied to the evaluation set. Bottom row shows the n-gram and the LSTM language model trained on the 1-best transcripts of unsupervised data and transcripts of seed data to the genre embedding system.

	dev		test	
	eq	vad	eq	vad
noadapt	29.54	29.59	42.39	42.50
genreCode	29.17	29.56	42.27	42.58
genreEmb	29.10	29.22	41.69	41.99
+seed-LSTM-eval				
noadapt	27.70	27.67	40.44	40.42
genreCode	27.42	27.63	40.25	40.45
genreEmb	27.29	27.35	39.91	39.95
+semisupLMs				
genreEmb	**26.82**	**26.94**	**39.21**	**39.29**

data segmentation was expected to generate cleaner segments containing less background noise and silence. After rejecting segments lower than 80% confidence, segments created by equal segmentation retained 732 hours of data, while segments created by VAD segmentation were 415 hours. There were 2,874 videos in equal segments while 2,331 videos in VAD segments. 634 videos were observed only in equal segments, while 91 were only in VAD segments. This shows that output of VAD segments was almost a subset of output of equal segments. The VAD system used to create segments was trained on different data domains (Section 4.2) from Blip10000. This may explain why the VAD system filtered out too many speech frames from the untranscribed data.

Content Genre Adaptation Results Table 3 shows results of using content genre for acoustic model adaptation. Results of "noadapt" correspond to "+seedLSTM-rescore" and "+seedLSTM-eval" in Table 2. A simple addition of genre code to an acoustic feature vector led to a small gain in WERs. Using the classifier to generate genre embedding resulted in the best WER among the systems 29.10% on dev and 41.69% on test without lattice rescoring, and 27.29% on dev and 39.91% on test sets with lattice rescoring. Further reduction in WERs of 0.45% on dev and 0.70% on test was obtained by decoding Blip dev and test using the n-gram re-trained on combination of the original training data with semi-supervised data and re-scoring lattices using the LSTM language model re-trained on combination of the original training data with semi-supervised data.

Overall, our experiments demonstrate that both semi-supervised training and content genre adaptation of an acoustic model can be effective for transcription of highly varied user-generated videos. The best configuration is to use equal segmentation of raw video data with removal of segments with confidence score less than 80%, addition of 1-best transcripts of untranscribed Blip10000 dev data to the n-gram and LSTM language model training, and genre embedding adaptation of the acoustic model together with semi-supervised training.

5 SCR Experiments

5.1 Creation of Known-Item Queries for Blip10000

To evaluate utility of the transcripts created using our alternative ASR systems for SCR we created 15 known-item search queries for dev set and 35 queries for test set using AMT. A known-item search seeks to re-find a previously observed relevant item. 15 documents from transcribed dev set and 35 documents from transcribed test set were randomly selected. AMT Workers were presented with the manual transcript of each document and a video. The workers were asked a question "Suppose you would like to find this video content on YouTube or another video sharing platform, enter minimum 3 words you would put in the search box". The workers were asked to create queries for not more than 3 documents to ensure that the query set was created by a diverse range of workers

5.2 Experimental Setup

We created search indexes of Blip10000 dev and test from ASR transcripts of the baseline ASR system (hybrid How2) in Table 1 and the augmented ASR system using semi-supervised acoustic model, language model and genre embedding (genreEmb) in Table 3. The dev document collection and the test document collection were indexed separately. In addition, the indexes of dev and test set were created using the manual transcripts described in Section 4.1. Since only 670 videos of dev set and 566 videos of test set have been manually transcribed, the indexed collections here were smaller than the actual document collections. The manually transcribed indexes were used as the oracle. The standard probabilistic BM25 information retrieval model was used to rank documents for each search query [12]. BM25 computes a relevance score given a document and a query by analysing frequency and inverse document frequency (IDF) of each query term in a document.

We report results for the known-item search task using the standard Mean Reciprocal Rank (MRR) metric conventionally used for known-item search tasks, where each query has a single relevant document. MRR is defined as follows:

$$MRR = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{rank_i} \tag{1}$$

where N is the number of user queries and $rank_i$ is the rank of the document relevent for the ith query.

5.3 Experimental Results and Analysis

Table 4 summarises MRR scores using the ASR baseline and augmented transcripts. The augmented ASR transcripts produced very large relative improvements in MRR of 152.6% on dev and by 77.45% on test sets over the baseline transcripts. Improvement of the MRR scores are statistically significant

Table 4. MRR results for known-item search using BM25 applied to the baseline and to the genre embedding transcripts.

	baseline	genreEmb	% change	oracle
dev	15.59	39.38	+152.6%	82.95
test	20.98	37.23	+77.45%	76.68

($p < .05$). This demonstrates that the 4-4.5% improvement of WERs gained from semi-supervised training and acoustic model adaptation led to significant improvement in SCR effectiveness for this task. For the dev set, MRR scores for 8 queries improved for 3 queries decreased and for the remaining 4 queries did not change. For test set MRR scores improved for 24 queries, decreased for 6 queries, and did not change for remaining 5 queries. These results are though still far below results achieved using error free manual transcripts. Despite being error free, manual transcripts do not achieve perfect results since queries can still score higher against non-relevant documents in cases where using the BM25 algorithm when they match the query better than the relevant item.

Table 5 shows an analysis of success and failure cases using the augmented ASR transcripts for the known-item search. The three queries 2, 5 and 38 show dramatic improvement in MRR score. This is brought about by the improvement of the ASR transcripts. We note that the augmented ASR system correctly transcribed "Julia Morris" for query 2, while the baseline system replaced it for "Egeria Moa". Similarly, the surname "Marcy" was replaced for "Mercy" by the baseline system for query 5 and "Hansel's Affair" was replaced for "Humps Hills" for query 38. This demonstrates that semi-supervised training and acoustic model adaptation using genre embedding help to improve search effectiveness. The more interesting cases are the bottom three queries at Table 5. Despite improvement in WER of documents relevant to the queries, MRR scores were not better. For query 10 is due to the fact that transcription of the proper noun "Mail Chimp" did not succeed. The retrieval model de-ranked the relevant document for query 24, since the augmented ASR system correctly transcribed the term "revamp" in another document irrelevant to the query and this document was ranked higher than the relevant document. Retrieval failure for query 47 occurred because the baseline transcript of the relevant document contained the term "Rocco", while the augmented transcript did not. These failure cases show the importance of correctly transcribing named entities for the search task, which could not be addressed by our enhancements to ASR system.

6 Conclusions and Further Work

In this paper, we reported our investigation into the use of semi-supervised training for ASR on the Blip10000 corpus., including data segmentation, data selection, acoustic modelling and language modelling. We found that: (i) surprisingly equal segmentation was slight better than VAD segmentation of data due to its larger amount of useful training data, and (ii) that further gain in

Table 5. Example queries for which MRR score increased or decreased when using augmented ASR transcripts. MRR scores of queries are shown with WERs of the relevant document.

queryID	query	baseline		genreEmb	
		MRR	WER	MRR	WER
2	Ancient craft film, Julia morris gallery ...	3.33	87.5	100	24.81
5	time for change, marcy winograd ...	0.76	50.0	100	13.70
38	Hansel Affair political ad	0.0	81.82	33.3	44.81
10	Mail Chimp Advertisement	16.67	29.59	12.5	18.93
24	Sonata, revamp, fuel, mileage, US	100	87.5	50	21.46
47	Vito Rocco Faintheart	100	57.14	0.0	40.62

WERs can be obtained by adding 1-best transcripts of untranscribed data to n-gram and LSTM language model training data. We found that use of content genre embedding can add useful information of acoustic conditions to adapted acoustic model. Overall we achieve a 4% WER reduction on the dev set and 4.5% on the test set. However, these improvements increased SCR effectiveness by approximately 150% on the dev set and 77% on the test set.

Acknowledgement

This work was supported by Science Foundation Ireland as part of the ADAPT Centre (Grant 13/RC/2106) at Dublin City University.

References

1. Abdel-Hamid, O., Jiang, H.: Fast speaker adaptation of hybrid nn/hmm model for speech recognition based on discriminative learning of speaker code. In: Proceedings of ICASSP 2013. pp. 7942–7946 (2013)
2. Hadian, H., Sameti, H., Povey, D., Khudanpur, S.: End-to-end speech recognition using lattice-free mmi. In: Proceedings of Interspeech 2018. pp. 12–16 (2018)
3. Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: Conference on Empirical Methods in Natural Language Processing (EMNLP 2018). pp. 66–71 (2018)
4. Larson, M., Jones, G.J.F.: Spoken Content Retrieval: A survey of techniques and technologies. Foundations and Trends in Information Retrieval **4**(4-5), 235–422 (2012)
5. Lüscher, C., Beck, E., Irie, K., Kitza, M., Michel, W., Zeyer, A., Schlüter, R., Ney, H.: RWTH ASR Systems for LibriSpeech: Hybrid vs Attention. In: Proceedings of Interspeech 2019. pp. 231–235 (2019)
6. Majumdar, S., Ginsburg, B.: MatchboxNet: 1D Time-Channel Separable Convolutional Neural Network Architecture for Speech Commands Recognition. In: Proceedings of Interspeech 2020. pp. 3356–3360 (2020)
7. Manohar, V., Hadian, H., Povey, D., Khudanpur, S.: Semi-supervised training of acoustic models using lattice-free mmi. In: Proceedings of ICASSP 2018. pp. 4844–4848 (2018)

8. Narayanan, A., Misra, A., Sim, K.C., Pundak, G., Tripathi, A., Elfeky, M., Haghani, P., Strohman, T., Bacchiani, M.: Toward domain-invariant speech recognition via large scale training. In: Proceedings of SLT 2018. pp. 441–447 (2018)
9. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: An ASR corpus based on public domain audio books. In: Proceedings of ICASSP 2015. pp. 5206–5210 (2015)
10. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The Kaldi speech recognition toolkit. In: Proceedings of ASRU 2011. pp. 1–4 (2011)
11. Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., Khudanpur, S.: Purely sequence-trained neural networks for asr based on lattice-free mmi. In: Proceedings of Interspeech 2016. pp. 2751–2755 (2016)
12. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at TREC-3. In: Proceedings of TREC 3. vol. 500-225, pp. 109–126 (1994)
13. Sainath, T.N., He, Y., Li, B., Narayanan, A., Pang, R., Bruguier, A., Chang, S.y., Li, W., Alvarez, R., Chen, Z., Chiu, C.C., Garcia, D., Gruenstein, A., Hu, K., Kannan, A., Liang, Q., McGraw, I., Peyser, C., Prabhavalkar, R., Pundak, G., Rybach, D., Shangguan, Y., Sheth, Y., Strohman, T., Visontai, M., Wu, Y., Zhang, Y., Zhao, D.: A streaming on-device end-to-end model surpassing server-side conventional model quality and latency. In: Proceedings of ICASSP 2020. pp. 6059–6063 (2020)
14. Sanabria, R., Caglayan, O., Palaskar, S., Elliott, D., Barrault, L., Specia, L., Metze, F.: How2: a large-scale dataset for multimodal language understanding. In: Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL). NeurIPS (2018)
15. Saon, G., Soltau, H., Nahamoo, D., Picheny, M.: Speaker adaptation of neural network acoustic models using i-vectors. In: Proceedings of ASRU 2013. pp. 55–59 (2013)
16. Schmiedeke, S., Xu, P., Ferrané, I., Eskevich, M., Kofler, C., Larson, A.M., Estève, Y., Lamel, L., Jones, G.J.F., Sikora, T.: Blip10000: a social video dataset containing spug content for tagging and retrieval. In: Proceedings of ACM MMSys 2013 (2013)
17. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X-vectors: Robust dnn embeddings for speaker recognition. In: Proceedings of ICASSP 2018. pp. 5329–5333 (2018)
18. Stolcke, A.: SRILM-an extensible language modeling toolkit. In: Proceedings of International Conference on Spoken Language Processing (ICSLP 2002) (2002)
19. Veselỳ, K., Hannemann, M., Burget, L.: Semi-supervised training of deep neural networks. In: Proceedings of ASRU 2013. pp. 267–272 (2013)
20. Wang, Y., Chen, T., Xu, H., Ding, S., Lv, H., Shao, Y., Peng, N., Xie, L., Watanabe, S., Khudanpur, S.: Espresso: A fast end-to-end neural speech recognition toolkit. In: Proceedings of ASRU 2019 (2019)
21. Xu, H., Chen, T., Gao, D., Wang, Y., Li, K., Goel, N., Carmiel, Y., Povey, D., Khudanpur, S.: A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition. In: Proceedings of ICASSP 2018. pp. 5929–5933 (2018)
22. Yu, K., Gales, M., Wang, L., Woodland, P.C.: Unsupervised training and directed manual transcription for lvcsr. Speech Communication **52**(7), 652–663 (2010)