SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network

William Chan, Daniel S. Park, Chris A. Lee, Yu Zhang, Quoc V. Le, Mohammad Norouzi

Google Research, Brain Team

{williamchan,danielspark,chrisalee,ngyuzh,qvl,mnorouzi}@google.com

Abstract

We present SpeechStew, a speech recognition model that is trained on a combination of various publicly available speech recognition datasets: AMI, Broadcast News, Common Voice, LibriSpeech, Switchboard/Fisher, Tedlium, and Wall Street Journal. SpeechStew simply mixes all of these datasets together, without any special re-weighting or re-balancing of the datasets. SpeechStew achieves SoTA or near SoTA results across a variety of tasks, without the use of an external language model. Our results include 9.0% WER on AMI-IHM, 4.7% WER on Switchboard, 8.3% WER on CallHome, and 1.3% on WSJ, which significantly outperforms prior work with strong external language models. We also demonstrate that SpeechStew learns powerful transfer learning representations. We fine-tune SpeechStew on a noisy low resource speech dataset, CHiME-6. We achieve 38.9% WER without a language model, which compares to 38.6% WER to a strong HMM baseline with a language model.

Index Terms: end-to-end speech recognition, multi-domain speech recognition

1. Introduction

End-to-end speech recognition models [1, 2, 3] have seen remarkable success in recent years [4]. The success of these methods have often been attributed to the abundance of training data [5] and the use of large deep models [6]. However, on noisy, low resource speech recognition datasets, such as CHiME-6 [7], where overfitting is a significant problem, end-to-end methods tend to struggle relative to HMM-based baselines [8]. For example, the best previously published end-to-end model achieved 49.0% WER on the CHiME-6 dev set [8], while the best HMM model achieves 36.9% WER [9].

Multi-lingual training [10, 11], multi-domain training [12, 13], unsupervised pre-training [14, 15], semi-supervised learning [16, 17] and transfer learning [18, 19] are some techniques proposed in the literature to enhance generalization. These methods optimize speech recognition models on data from related tasks (typically of high resource), to help the specific task of interest (typically of low resource). For example, in multi-lingual training, the knowledge from a high resource language may transfer to a low resource language [20]. In multi-domain training, combining different domain datasets of the same language, could facilitate the cross-sharing of knowledge across the domains [13]. In unsupervised pre-training, the knowledge from the pre-training task may transfer to the supervised task [21]. In transfer learning, a general model is trained with a large amount of data. Subsequently, its knowledge is transferred via fine-tuning on training data from a downstream task that is typically of low resource [22].

This paper presents SpeechStew. SpeechStew is a simple approach to end-to-end speech recognition, which leverages both multi-domain training and transfer learning. SpeechStew follows the following simple recipe:

1. Combine all available speech recognition data without any domain-dependent re-balancing or re-weighting.
2. Train a single large neural network (a 100M or 1B parameter model) on the combined data.

Our method does not utilize any domain labels, or introduce any additional hyperparameters for combining the data. We do not incorporate an external language model during inference, yet our result compares favourably to prior work that utilize strong language models, achieving SoTA or near SoTA results across various tasks (AMI, Common Voice, LibriSpeech, Switchboard, Tedlium, and WSJ).

We also demonstrate that SpeechStew has strong transfer learning capabilities. When presented with a new unseen low resource dataset (CHiME-6 in our setup), we merely:

3. Fine-tune SpeechStew on the new labelled dataset.

We find that this straightforward pre-training and fine-tuning procedure yields near-SoTA results on CHiME-6. This is encouraging since CHiME-6 is a particularly challenging task [7] for end-to-end speech recognition models, which suffer from over-fitting issues [8]. We also demonstrate that our method is complementary to other pre-training methods, in particular unsupervised wav2vec pre-training [21] which we use in conjunction with SpeechStew training.

2. SpeechStew

In this section, we describe the model and training data setup of SpeechStew. We also describe our transfer learning setup for fine-tuning on new unseen tasks.

2.1. Model

In our implementation, SpeechStew uses the Conformer [31] RNN-T [32] architecture. We experiment with both the 100M parameter [31] and the 1B parameter configuration [6]. We find that wav2vec pre-training [15] is needed to train the 1B parameter model [6]. We apply the default hyperparameters from prior work [31, 6] including the learning rate schedule. We do not incorporate an external language model.

2.2. Multi-domain Training

We combine the following datasets without any form of re-weighting or resampling to construct the training set for SpeechStew:

1. AMI [33]. AMI is approximately 100 hours of meeting recordings.
2. Common Voice [34]. Common Voice is a crowd-sourced open licensed speech dataset. We use the version 5.1

Task	AMI		Common Voice	LibriSpeech		Switchboard/Fisher		Tedlium	WSJ
	IHM	SDM1		clean	other	SWBD	CH		eval92
Prior Work (no LM)									
Single domain			16.9† [23]	1.5 [6]	2.7 [6]			7.5 [23]	9.3 [24]
Multi-domain	12.2‡ [25]	**21.2**‡ [25]	15.5† [23]	3.0 [23]	7.3 [23]	6.3 [23]	10.7 [23]	6.9 [23]	3.4 [23]
Prior Work (with LM)									
Single domain	17.5 [26]	36.4 [27]	13.6† [23]	**1.4** [6]	**2.6** [6]	4.9 [28]	9.5 [28]	5.6 [29]	2.9 [30]
Multi-domain			10.6† [23]	2.1 [23]	4.4 [23]	5.5 [23]	9.1 [23]	**5.2** [23]	2.0 [23]
Our Work (no LM)									
Single Domain Baseline (100M)	26.1	40.5	16.3 (13.8†)	2.1	4.4	5.6	9.7	7.6	28.2
SpeechStew (100M)	**9.0**	**21.7**	12.1 (9.7†)	2.0	4.0	**4.7**	**8.3**	**5.3**	**1.3**
SpeechStew (1B)	9.5	22.7	**10.8 (8.4†)**	**1.7**	**3.3**	4.8	10.6	5.7	**1.3**

Table 1: *Speech recognition word error rates (%) across multiple tasks. SpeechStew achieves SoTA or near SoTA across many tasks. Our SpeechStew 1B model uses wav2vec pre-training on LibriLight. SpeechStew does not use a separate language model.* †*We follow [23] and remove punctuations during evaluation.* ‡*Concurrent work [25].*

Model	Dev	Eval
Prior Work (with LM)		
Official HMM Baseline [7]	51.8	51.3
HMM [9]	**36.9**	**38.6**
RNN-T [8]	49.0	
Our Work (no LM)		
Zero-Shot (never seen CHiME-6)		
SpeechStew (100M)	54.9	57.2
SpeechStew (1B)	39.2	53.7
Fine-tuned with CHiME-6		
Baseline (100M)	70.0	66.7
SpeechStew + Fine-tune (100M)	33.1	40.6
SpeechStew + Fine-tune (1B)	**31.9**	**38.9**

Table 2: *We apply transfer learning and fine-tune SpeechStew on CHiME-6.*

(June 22 2020) snapshot with approximately 1500 hours. The data was collected at 48 KHz, and we resampled it to 16 KHz.

3. English Broadcast News (LDC97S44, LDC97T22, LDC98S71, LDC98T28). English Broadcast News is approximately 50 hours of television news.
4. LibriSpeech [35]. LibriSpeech is approximately 960 hours of speech from audiobooks.
5. Switchboard/Fisher (LDC2004T19, LDC2005T19, LDC2004S13, LDC2005S13, LDC97S62). Switchboard/Fisher is approximately 2000 hours of telephone conversations. The data was collected at 8 KHz, and we upsampled it to 16 KHz.
6. TED-LIUM v3 [36, 37]. TED-LIUM is approximately 450 hours of TED talks.
7. Wall Street Journal (LDC93S6B, LDC94S13B). WSJ is approximately 80 hours of clean speech.

2.3. Transfer Learning

We demonstrate the transfer learning capabilities of SpeechStew. Once we have a general purpose SpeechStew model (trained on the datasets mentioned in Section 2.2), we can fine-tune and adapt SpeechStew onto a new task. CHiME-6 [7] is a noisy low resource dataset set, which contains approximately 40 hours of distant microphone conversational speech recognition in everyday home environments. CHiME-6 is difficult for end-to-end speech recognition models to train directly due to over-fitting issues [8]. We fine-tune SpeechStew on CHiME-6 to demonstrate the transfer learning capabilities.

The transfer learning capabilities of SpeechStew are extremely practical. It implies we can train a general purpose model once, then fine-tune to specific low resource tasks. This can be done at a very low cost, since fine-tuning typically requires only a few thousand steps, compared to ≈100k steps needed to train a model from scratch.

3. Experiments

We build single task mode baselines, where the models are trained only on their respective domains. We use the Conformer 100M architecture [31] for these baselines; we found the 1B model to overfit and perform dramatically worse. We perform model selection via the development sets per baseline task.

Our SpeechStew model uses the 100M parameter and 1B parameter Conformer architecture [31, 6]. We used the default experimental settings of these references to train the models. We train all our SpeechStew models for exactly 100k steps, without any model selection.

Table 1 summarizes our results. We emphasize that the reported performance of SpeechStew is obtained without the use of an external language model. SpeechStew outperforms almost all prior work, including those using strong language models.

3.1. Transfer Learning and CHiME-6

CHiME-6 [7] is a low resource noisy speech dataset. We fine-tune our 100M and 1B SpeechStew models with the CHiME-6 data, and compare these results against baselines obtained by training a 100M parameter Conformer model and a 1B-parameter Conformer model (with LibriLight-only pre-training) with CHiME-6. Table 2 summarizes our results.

4. Conclusion

We presented SpeechStew. We achieve SoTA or near SoTA results across a variety of speech tasks. Our approach simply mixes all available training data to train one large neural network. We can apply transfer learning and finetune SpeechStew to new tasks.

5. References

[1] A. Graves and N. Jaitly, "Towards End-to-End Speech Recognition with Recurrent Neural Networks," in *ICML*, 2014.

[2] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition," in *ICASSP*, 2016.

[3] W. Chan, C. Saharia, G. Hinton, M. Norouzi, and N. Jaitly, "Imputer: Sequence Modelling via Imputation and Dynamic Programming," in *ICML*, 2020.

[4] D. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. Cubuk, and Q. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *INTERSPEECH*, 2019.

[5] C.-C. Chiu, T. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, "State-of-the-art Speech Recognition With Sequence-to-Sequence Models," in *ICASSP*, 2018.

[6] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition," in *arXiv:2010.10504*, 2020.

[7] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. S. Subramanian, J. Trmal, B. B. Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, T. Yoshioka, and N. Ryant, "CHiME-6 Challenge:Tackling Multispeaker Speech Recognition for Unsegmented Recordings," in *arXiv*, 2020.

[8] A. Andrusenko, A. Laptev, and I. Medennikov, "Towards a Competitive End-to-End Speech Recognition for CHiME-6 Dinner Party Transcription," in *arXiv*, 2020.

[9] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya1, I. Soroki, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, and A. Romanenko, "The STC System for the CHiME-6 Challenge," in *CHiME Workshop*, 2020.

[10] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, D. Povey, A. Rastrowand, R. C. Rose, and S. Thomas, "Multilingual Acoustic Modeling for Speech Recognition based on Subspace Gaussian Mixture Models," in *ICASSP*, 2010.

[11] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual Acoustic Models Using Distributed Deep Neural Networks," in *ICASSP*, 2013.

[12] https://github.com/kaldi-asr/kaldi/commits/master/egs/multi_en.

[13] A. Narayanan, A. Misra, K. C. Sim, G. Pundak, A. Tripathi, M. Elfeky, P. Haghani, T. Strohman, and M. Bacchiani, "Toward domain-invariant speech recognition via large scale training," in *SLT*, 2018.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *NAACL*, 2019.

[15] A. M. M. A. Alexei Baevski, Henry Zhou, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *arXiv:2006.11477*, 2020.

[16] G. Synnaeve, Q. Xu, J. Kahn, E. Grave, T. Likhomanenko, V. Pratap, A. Sriram, V. Liptchinsky, and R. Collobert, "End-to-end ASR: from Supervised to Semi-Supervised Learning with Modern Architectures," in *ICML: Workshop on Self-supervision in Audio and Speech*, 2020.

[17] D. S. Park, Y. Zhang, Y. Jia, W. Han, C.-C. Chiu, B. Li, Y. Wu, and Q. V. Le, "Improved Noisy Student Training for Automatic Speech Recognition," in *INTERSPEECH*, 2020.

[18] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition," in *ICML*, 2014.

[19] S. Kornblith, J. Shlens, and Q. Le, "Do Better ImageNet Models Transfer Better?" in *CVPR*, 2019.

[20] B. Li, Y. Zhang, T. Sainath, Y. Wu, and W. Chan, "Bytes are All You Need: End-to-End Multilingual Speech Recognition and Synthesis with Bytes," in *ICASSP*, 2019.

[21] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-training for Speech Recognition," in *arXiv*, 2019.

[22] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big Transfer (BiT): General Visual Representation Learning," in *arXiv*, 2019.

[23] T. Likhomanenko, Q. Xu, V. Pratap, P. Tomasello, J. Kahn, G. Avidov, R. Collobert, and G. Synnaeve, "Rethinking Evaluation in ASR: Are Our Models Robust Enough?" *arXiv preprint arXiv:2010.11745*, 2020.

[24] S. Sabour, W. Chan, and M. Norouzi, "Optimal Completion Distillation for Sequence Learning," in *ICLR*, 2019.

[25] N. Kanda, G. Ye, Y. Wu, Y. Gaur, X. Wang, Z. Meng, Z. Chen, and T. Yoshioka, "Large-Scale Pre-Training of End-to-End Multi-Talker ASR for Meeting Transcription with Single Distant Microphone," in *arXiv*, 2021.

[26] G. Sun, C. Zhang, and P. C. Woodland, "Transformer Language Models with LSTM-based Cross-utterance Information Representation," in *arXiv*, 2021.

[27] N. Kanda, Y. Fujita, S. Horiguchi, R. Ikeshita, K. Nagamatsu, and S. Watanabe, "Acoustic Modeling for Distant Multi-talker Speech Recognition with Single- and Multi-channel Branches," in *ICASSP*, 2019.

[28] W. Wang, G. Wang, A. Bhatnagar, Y. Zhou, C. Xiong, and R. Socher, "An investigation of phone-based subword units for end-to-end speech recognition," in *INTERSPEECH*, 2020.

[29] W. Zhou, W. Michel, K. Irie, M. Kitza, R. Schlüter, and H. Ney, "The RWTH ASR System for TED-LIUM Release 2: Improving Hybrid HMM with SpecAugment," in *ICASSP*, 2020.

[30] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "End-to-end speech recognition using lattice-free MMI," in *INTERSPEECH*, 2018.

[31] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *INTERSPEECH*, 2020.

[32] A. Graves, "Sequence Transduction with Recurrent Neural Networks," in *ICML Representation Learning Workshop*, 2012.

[33] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI Meeting Corpus: A Pre-Announcement," in *International Workshop on Machine Learning for Multimodal Interaction*, 2005.

[34] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common Voice: A Massively-Multilingual Speech Corpus," in *LREC*, 2020.

[35] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*, 2015.

[36] A. Rousseau, P. Deléglise, and Y. Estève, "TED-LIUM: an Automatic Speech Recognition dedicated corpus," in *LREC*, 2012.

[37] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Estève, "TED-LIUM 3: Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation," in *SPECOM*, 2018.