HarperValleyBank: A Domain-Specific Spoken Dialog Corpus

Mike Wu[1]*, Jonathan Nafziger*[2]*, Anthony Scodary*[2]*, Andrew Maas*[1]

[1]Stanford University
[2]Gridspace, Inc.

{wumike,amaas}@cs.stanford.edu, {jonathan,anthony}@gridspace.com

Abstract

We introduce HARPERVALLEYBANK, a free, public domain spoken dialog corpus. The data simulate simple consumer banking interactions, containing about 23 hours of audio from 1,446 human-human conversations between 59 unique speakers. We selected intents and utterance templates to allow realistic variation while controlling overall task complexity and limiting vocabulary size to about 700 unique words. We provide audio data along with transcripts and annotations for speaker identity, caller intent, dialog actions, and emotional valence. The data size and domain specificity makes for quick transcription experiments with modern end-to-end neural approaches. Further, we provide baselines for representation learning, adapting recent work to embed waveforms for downstream prediction tasks. Our experiments show that tasks using our annotations are sensitive to both the model choice and corpus size.

1. Introduction

Recent innovations in deep learning approaches substantially improved spoken dialog systems in both academic research and industry applications. Speech recognition systems now regularly leverage neural network models to achieve near human performance [1, 2, 3, 4]. Modern systems increasingly use architecture themes including attention and end-to-end recurrent neural networks to encode few assumptions and rapidly adapt to new data [5, 6, 7, 8, 9, 10]. In parallel, approaches to spoken and text-based dialog systems increasingly leverage neural networks for dialog management and state representation [11, 12, 13].

We developed the HARPERVALLEYBANK corpus for homeworks and projects in Stanford's *Spoken Language Processing* course[1], as well as for research on speech recognition combined with other spoken language tasks. The goals of the corpus are to provide:

- Freely available data for education, research, or commercial development. We release the data using a Creative Commons license (CC-BY)[2].
- Sufficient size and variability to meaningfully evaluate end-to-end neural approaches for speech transcription.
- Manageable overall size and complexity to enable students to quickly iterate on experiments without requiring expensive compute hardware for training.
- Annotations for dialog-relevant tasks aside from speech transcription (e.g. intent, dialog action) to enable multi-task training and representation transfer.
- Realistic domain-specific, goal-oriented conversations to evaluate representation transfer approaches across domains in spoken dialog systems.

[1] cs224s.stanford.edu

[2] creativecommons.org/licenses/by/4.0

There is significant recent work in representation learning for domain and task transfer via embedding models [14, 15, 16, 17], which can be trained without any supervision and reused for many downstream tasks like predicting speaker identity [18, 19] and commands [20, 21]. Representation transfer is important to warm start deep learning based dialog systems on new task domains. Part of our motivation in designing this corpus is providing a realistic test case for representation learning approaches to adapt to speech tasks. See [22] for a review of availaable dialog datasets. We recorded two-sided phone conversations to simulate customer call center interactions in a financial services domain. The dataset is representative of human to human goal-oriented dialogs for consumer banking with a narrowly scoped set of intents.

In the next sections, we provide more details on the corpus and its collection, followed by experiments showcasing its applications to automatic speech recognition and unsupervised representation learning. In Sec. 2, we discuss basic corpus statistics, caller intents, and the data generation and annotation process. In Sec. 3, we explore end-to-end neural models with multi-task objective functions to simultaneously perform speech-to-text transcription and caller intent prediction. In Sec. 4, we explore using caller intents and sentiments as downstream objectives to evaluate representation transfer, and report unsupervised baselines. The full dataset with a PyTorch implementation reproducing the speech recognition and transfer experiments is publically available[3].

2. The HARPERVALLEYBANK Corpus

We compile a dataset of recorded audio conversations between an agent and a customer of a bank. Conversations are goal-oriented, such as ordering a new checkbook or checking the balance of an account. Fig. 1 shows an example conversation from the dataset. We collected data using the Gridspace Mixer platform, where crowd workers are randomly paired for short telephone conversations. Mixer membership includes hundreds of past and current professional call center agents who are trained to perform assorted Mixer tasks in domains including healthcare, telecommunications, and commerce.

2.1. Data Collection Procedure

Using the Mixer web platform, a person is randomly assigned the role of agent or customer and provided a script for the interaction along with a telephone number to call to start the conversation. Roles are randomly assigned for each call, so the same worker can appear as both customer and agent in different conversations. We created a set of conversation goals and scripts for each interaction using templates intended to capture variety in each intent while keeping workers' word choices and the

[3] https://github.com/cricketclub/gridspace-stanford-harper-valley

overall interactions fairly simple with limited vocabulary. We do not control the noise environment or microphones used by each worker, and there is natural variation across different types of phones and environments.

When a person calls in, they are paired with the next available conversation partner. The groups are large enough that many unique pairings occur over the course of one session. Once the Mixer task is live for the caller, the web application will change state, informing the caller whether they are acting in the role of the agent or caller. The instructions, data, and user interface adapt to the role and provide a rough script. The customer role initiates a call task by expressing an intent, and the agent role has an interactive web interface to simulate completing a task. We encouraged callers to use mobile phones or headsets to encourage a microphone transfer function that is acoustically representative of a real call center.

During each call, participants have the script for their side of the conversation in front of them in a web browser. The worker playing the customer role is given a single intent for the conversation, along with specific values for relevant slots (e.g. the amount of money to transfer and the source/target accounts). When playing the agent role, a worker is shown some simple buttons and menus they must click to perform the requested operation (e.g. "check account balance").

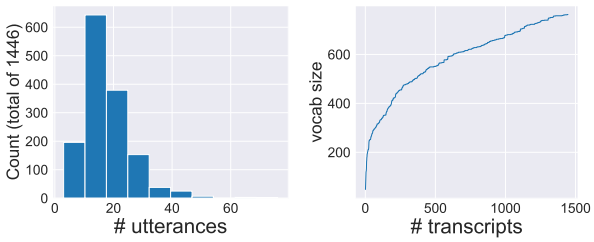
A conversation is deemed successful and considered for the dataset if the agent correctly executes the task provided to the customer caller. Names and slot values for different transactions are randomly generated, and we limit the number of possible names and proper nouns to reduce overall corpus vocabulary size. The Gridspace Mixer platform handles generating random templates from a high level specification, all associated telephony operations to pair callers, and recording audio along with metadata for each interaction.

AGENT: hello this is harper valley national bank my name is jay how can I help you today
CALLER: hi my name is mary davis
CALLER: [noise]
CALLER: i would like to schedule an appointment
AGENT: yeah sure what day what time
CALLER: thursday one thirty p m
AGENT: that's done anything else
CALLER: that's it
AGENT: have a good one.

Figure 1: *Example conversation from corpus.*

2.2. Data Labelling

Gridspace Mixer trains a subset of its community to perform a wide range of annotations. For this corpus, Mixers performed three primary labeling tasks: *text transcript, audio quality, and script adherence ratings*. Gridspace has provided the Mixer community with a highly specialized speech labeling tool called Scriber. Scriber is designed for rapid human transcription and data labeling. The tool also provides a wide array of convenience and ergonomic functions, designed to enable efficient labeling of large spoken language datasets. Every person trained to use Scriber must go through several training sessions, which requires them to watch training videos and perform well on a quiz. For dialog actions and emotional valence, labels were instead produced using a Gridspace API rather than human annotation. As a result, there may be some noise or bias, but our experiments indicate they are reasonable for a benchmark task.



(a) (b)

Figure 2: *(a) The empirical distribution over the number of utterances in a conversation transcript. (b) The number of unique words spoken as a function of the conversation count.*

The HARPERVALLEYBANK corpus was collected over three separate Mixer sessions and then filtered post-annotation, informed by the script adherence labels and audio quality labels, to ensure the data was simple and low variance. This filtering ensures the corpus provides conversational and task-oriented speech data while regulating for simplicity. The primary target of the cleaning were conversations where calls dropped or there were other technical issues which derailed the conversation. Specifically we removed conversations with script adherence ratings less than 4 and audio quality ratings less than 3. Furthermore we filtered out conversations which contained some words such as 'frozen', 'website', and 'refresh', which indicated conversation about technical issues with the task interface. In total we removed 375 conversations.

2.3. Corpus Statistics

HARPERVALLEYBANK Statistics	
Hours of audio	23.7
# of conversations/transcripts	1,446
# of utterances	25,730
# of unique words	735
Mean # of lines per conversation	17.8
Median # of lines per conversation	16
Mean # of words per utterance	4.1
Median # of words per utterance	4.5
# of unique speakers	59
# of task classes	8
# of dialog action classes	16
# of sentiment classes	3

Table 1: *Basic corpus statistics* .

Table 1 shows basic statistics of the HARPERVALLEYBANK corpus. The corpus contains about 23 hours of audio in total, across 1,446 conversations. Conversations range from 2 to 60 utterances, with an average of 18. Each utterance roughly corresponds to a single turn in the conversation. Due to automatic segmentation of utterances, there can be multiple utterances in a row from a single speaker's turn. Notably, the corpus has a small vocabulary of approximately 700 unique words. Many of the most common words in the vocabulary are domain-specific to customer service e.g. "help", "thank", or "please". Fig. 2b depicts how vocabulary size scales with dataset size.

The Gridspace platform records each side of the conversation separately, and we release the audio in speaker-separated

files encoded as 8kHz per the original telephony data. We transcribed the utterances via crowd workers with basic speech transcription training again using the Gridspace Mixer platform. Workers are not instructed to carefully transcribe word fragments or non-speech noises. Leveraging crowd workers and transcribing without precise fragments and non-speech tags has been shown to be a viable approach for training speech recognition systems [23]. In addition to human transcriptions of each conversation, we include four additional labels:

Intent. Each conversation has a single intent representing the customer's goal in the conversation. An intent can be one of eight categories: *order checks, check balance, replace card, reset password, get branch hours, pay bill, schedule appointment, transfer money*. Fig. 3a shows the distribution of intents to be roughly balanced. Conversation intents are derived automatically from the tasks assigned to callers during collection.

Emotional Valence. Utterances are automatically labeled with three sentiment categories, *negative, neutral, and positive*. There is a probability estimate label for each category, generated by the Gridspace Speech API that is trained on a large corpus of proprietary data from multiple domains. Fig. 3b shows the distribution of each sentiments across utterances.

Speaker ID. Utterances have a unique ID out of 59 speakers. The number of utterances per speaker are imbalanced, with most speakers responsible for less than 50 utterances.

Dialog Action. Every utterance is accompanies by one or more labels representing a "conversational mode". The distribution over actions is imbalanced with "greeting" being the most frequent and many infrequent actions combined into the "other" category. The 16 possible actions are: *"yes" response, greeting, response, data confirmation, procedure explanation, data question, closing, data communication, "bear with me" response, acknowledgement, data response, filler disfluency, thanks, open question, problem description*, and *other*. Fig. 3d shows the distribution of dialog actions for utterances.

3. Spoken Language Understanding

Using the HARPERVALLEYBANK corpus, we evaluate three common approaches for automatic speech recognition: connectionist temporal classification or CTC [24], Listen-Attend-Spell or LAS [6], and finally, a "multi-task" objective combining the two previous losses [25], MTL.

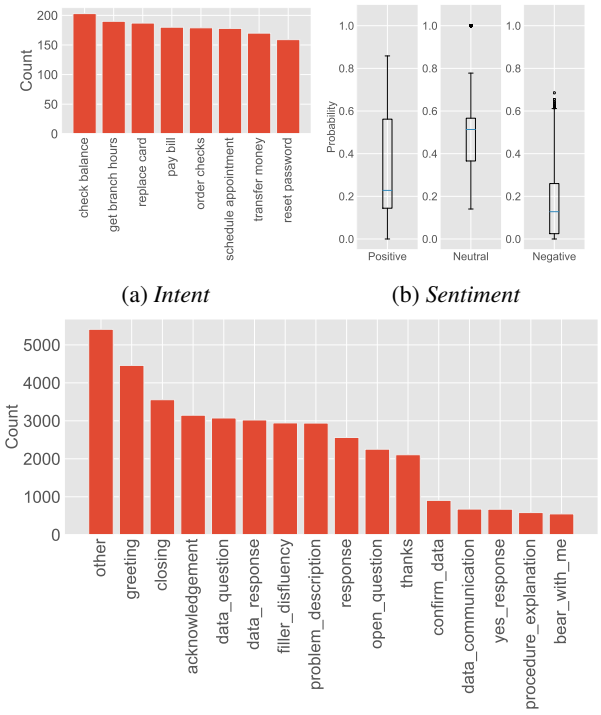
In addition to optimizing the speech recognition objective, denoted $\mathcal{L}_{\text{asr}}$, we fit four linear layers mapping the encoding of the audio signal to a prediction of the intent, dialog action, and sentiment labels. These three auxiliary objectives are optimized jointly with the speech recognition objective:

$$\beta\mathcal{L}_{\text{asr}} + (1-\beta)\left(\mathcal{L}_{\text{task}} + \mathcal{L}_{\text{action}} + \mathcal{L}_{\text{sent}}\right)$$

where $\mathcal{L}_{\text{task}}$, and $\mathcal{L}_{\text{sent}}$ are cross entropy losses, whereas $\mathcal{L}_{\text{action}}$ comprises a sum of 16 binary cross entropy losses. The scalar $\beta \in [0, 1]$ weights the recognition and auxiliary objectives.

3.1. Training Details

Audio is preprocessed to 128 Mel-frequency spectrogram features with a sampling rate of 8kHz, a hop length of 128, and a window size of 256. For CTC, we encode log-Mel features using a bi-directional LSTM with two layers and 128 hidden dimensions. CTC decoding is done greedily with no language



(a) *Intent* (b) *Sentiment*

(c) *Dialog Actions*

Figure 3: *Distribution over auxiliary labels. Subfigures (a) & (c) show the counts for intent and dialog action. Subfigure (d) show boxplots for each of the three sentiments.*

model. In LAS, the listener network is composed of three stacked pyramid bi-directional LSTMs with 128 hidden dimensions whereas the speller network is an uni-directional LSTM with 256 hidden dimensions and a single-headed attention layer. For MTL, we use the same encoder as LAS but include both the speller and CTC decoder. In MTL, we can interpret CTC as a secondary objective whose main role is to regularize the LAS encoder to respect CTC alignments:

$$\beta(\alpha\mathcal{L}_{\text{ctc}} + (1-\alpha)\mathcal{L}_{\text{las}}) + (1-\beta)\left(\mathcal{L}_{\text{task}} + \mathcal{L}_{\text{action}} + \mathcal{L}_{\text{sent}}\right)$$

where α is chosen to be 0.7 by grid search. The LAS and MTL objectives are optimized with teacher forcing with probability 0.5. We train each model for 200 epochs with Adam [26] using a learning rate of 1e-3, batch size 128, and gradient clipping.

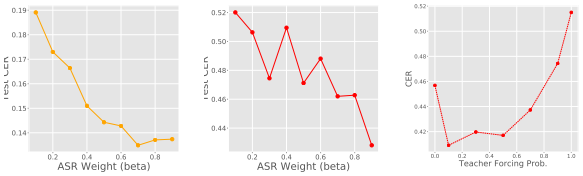
3.2. Results and Analysis

Table 2 shows the performance of CTC, LAS, and MTL on a speaker split test set. We report CER to measure transcription quality, accuracy for sentiment and intent prediction, and F1 for dialog action prediction (due to class imbalance). The highest performing model in each metric is bolded.

Model	CER	Action (F1)	Sentiment	Intent
CTC	**14.43**	**0.3864**	**84.46**	**45.47**
LAS	47.45	0.2931	72.09	34.96
MTL	38.59	0.3222	76.12	42.28

Table 2: *Speech recognition and auxiliary task performance using three modern end-to-end neural approaches.*

Interestingly, we find that CTC outperforms LAS and MTL. We attribute this to the small size of the HARPERVALLEYBANK corpus that encourages expressive autoregressive models to heavily overfit. In this case, the structure of the CTC loss is beneficial compared with a model that learns attention. We see further evidence of this by MTL outperforming LAS, where CTC acts as a regularizer. LAS can improve a bit by tuning the probability of teacher forcing (see Fig 4).



(a) *CTC CER* (b) *LAS CER* (c) *Teacher Forcing*

Figure 4: *Subfigure (a) and (b) show the effect of increased weight on the test CER error for CTC and LAS models. Subfigure (c) shows the effect of teacher forcing in training LAS.*

4. Unsupervised Speech Representations

Unsupervised representation learning seeks to derive useful representations of speech waveforms without any human annotations. The learned representations are reused for downstream tasks, such as predicting caller intent or dialog action. As HARPERVALLEYBANK is a small dataset, it is a suitable candidate to measure the effectiveness of speech representations. In our experiments, we pretrain representations on the 100 hour split or 960 hour split of LibriSpeech [18].

Ou baselines use recent ideas from contrastive learning [27, 28, 29, 30, 31, 32, 33, 34] where representations are learned by discriminating between specific instances in a dataset. In particular, we adapt four algorithms from computer vision to speech: Instance Discrimination or IR [27], Local Aggregation or LA [28], Momentum Contrast or MoCo [31], and SimCLR [32]. One of our contributions is to establish comparable baselines for audio representation learning. As a close relative, we also evaluate representations learned using Wav2Vec-1.0 [16] and Wav2Vec-2.0 [17].

4.1. Training Details

Given a waveform, we truncate to 150k frames, and compute the log-Mel spectrogram. Spectrograms are z-scored using training statistics, which we found to be important for generalization. By default, we augment waveforms by selecting contiguous crops with a minimum and maximum ratio of 0.08 to 1.0, along with Gaussian noise with a scale of 1.0. We separately explore first computing the spectrogram, then applying a time and frequency mask as augmentation, denoted by the (*) superscript in Table 3. Regardless, we fit a ResNet50 [35] to map the spectrogram to a 2048 dimensional embedding.

After representation learning, we measure the quality of an embedding by linear evaluation [27, 28, 31, 32]. The HARPERVALLEYBANK corpus is split into train (80%) and test (20%) sets *by class* to ensure both sets have instances of each class. Thus, each transfer task has its own train test split. Refer to the public repository for training hyperparameters.

Model	Spk.	Intent	Action	Sent.
Wav2Vec 1.0 (960hr)	18.2	17.1	0.0	53.7
Wav2Vec 2.0 (100hr)	22.3	19.7	0.0	54.3
Wav2Vec 2.0 (960hr)	27.3	20.5	0.0	55.5
IR (100hr)	99.5	99.1	0.0	51.3
LA (100hr)	99.5	98.8	0.0	50.5
MoCo (100hr)	99.6	98.9	0.0	53.2
SimCLR (100hr)	99.8	99.3	0.0	53.9
IR* (100hr)	99.5	84.5	0.0	51.4
LA* (100hr)	97.5	75.8	1.4	55.1
MoCo* (100hr)	99.1	82.6	0.0	54.0
SimCLR* (100hr)	99.2	81.4	0.0	54.6
IR (960hr)	99.9	99.9	17.4	66.3
LA (960hr)	99.9	99.9	18.4	64.6
MoCo (960hr)	99.9	99.9	17.3	65.5
SimCLR (960hr)	99.9	99.9	17.4	65.9
IR* (960hr)	99.9	86.7	17.6	64.8
LA* (960hr)	99.9	79.8	18.0	64.6
MoCo* (960hr)	99.5	86.1	16.2	64.3
SimCLR* (960hr)	98.6	82.6	16.1	65.6

Table 3: *Performance on speaker identity, intent, dialog action, and emotional valence. We report F1 score for performance on predicting dialog action. The superscript (*) represents using spectral augmentations rather than wavform augmentations.*

4.2. Results and Analysis

Table 3 compares the different unsupervised models: IR, LA, MoCo, and SimCLR surpass supervised methods (e.g. CTC, LAS, and MTL) in intent prediction, falling short in dialog action and sentiment prediction. In particular, dialog action prediction is a surprisingly difficult task for our unsupervised representations. Whereas supervised methods achieve upwards of 0.3 F1, the best models in Table 3 achieve only half the score, despite seeing 960 hours of speech. That being said, adding 860 hours of speech improved the representations, as shown by overall better performance. Finally, compared to Wav2Vec, the contrastive objectives find large gains of up to 70%.

5. Conclusion

We introduced HARPERVALLEYBANK, a new speech corpus of transcribed conversations between employees and customers in a bank transcaction. The corpus includes additional labels, including speaker identity, caller intent, dialog actions, and sentiment. In our experiments, we established baseline models that showed this corpus to be an interesting challenge for future algorithms, and a useful educational tool for modern deep learning approaches to spoken dialog. Our experiments analyzed utterances independently, future work can explore using the HARPERVALLEYBANK corpus in conversation modelling and its related downstream optimization.

6. Acknowledgments

We thank the Gridspace team for their support in data collection and analysis. Experiments used PyTorch, PyTorch Lightning, and Weights & Biases [36]. We thank Samuel Kwong and Dan Jurafsky for helpful discussion and feedback.

7. References

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim *et al.*, "English conversational telephone speech recognition by humans and machines," *arXiv preprint arXiv:1703.02136*, 2017.

[3] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The microsoft 2017 conversational speech recognition system," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5934–5938.

[4] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition."

[5] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[6] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[7] A. Maas, Z. Xie, D. Jurafsky, and A. Y. Ng, "Lexicon-free conversational speech recognition with neural networks," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 345–354.

[8] T. Likhomanenko, G. Synnaeve, and R. Collobert, "Who needs words? lexicon-free speech recognition," *arXiv preprint arXiv:1904.04479*, 2019.

[9] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*, 2014, pp. 1764–1772.

[10] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4945–4949.

[11] C. Khatri, B. Hedayatnia, A. Venkatesh, J. Nunn, Y. Pan, Q. Liu, H. Song, A. Gottardi, S. Kwatra, S. Pancholi *et al.*, "Advancing the state of the art in open domain dialog systems through the alexa prize," *arXiv preprint arXiv:1812.10757*, 2018.

[12] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Learning to compose neural networks for question answering," *arXiv preprint arXiv:1601.01705*, 2016.

[13] B. Liu and I. Lane, "Iterative policy learning in end-to-end trainable task-oriented neural dialog models," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 482–489.

[14] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[15] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 12, pp. 2041–2053, 2019.

[16] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[17] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[19] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[20] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[21] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," *arXiv preprint arXiv:1904.03670*, 2019.

[22] I. V. Serban, R. Lowe, P. Henderson, L. Charlin, and J. Pineau, "A survey of available corpora for building data-driven dialogue systems," *arXiv preprint arXiv:1512.05742*, 2015.

[23] S. Novotney and C. Callison-Burch, "Cheap, fast and good enough: Automatic speech recognition with non-expert transcription," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 207–215.

[24] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[25] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] Z. Wu, Y. Xiong, S. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance-level discrimination," *arXiv preprint arXiv:1805.01978*, 2018.

[28] C. Zhuang, A. L. Zhai, and D. Yamins, "Local aggregation for unsupervised learning of visual embeddings," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6002–6012.

[29] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in *Advances in Neural Information Processing Systems*, 2019, pp. 15 535–15 545.

[30] I. Misra and L. v. d. Maaten, "Self-supervised learning of pretext-invariant representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6707–6717.

[31] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[32] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[33] M. Wu, C. Zhuang, M. Mosse, D. Yamins, and N. Goodman, "On mutual information in contrastive learning for visual representations," *arXiv preprint arXiv:2005.13149*, 2020.

[34] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "Spanbert: Improving pre-training by representing and predicting spans," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 64–77, 2020.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[36] L. Biewald, "Experiment tracking with weights and biases," *URL https://www. wandb. com/. Software available from wandb. com*, 2020.