TASK ORIENTED DIALOGUE AS A CATALYST FOR SELF-SUPERVISED AUTOMATIC SPEECH RECOGNITION

David M. Chan[⋆†] *Shalini Ghosh*[†] *Hitesh Tulsiani*[†] *Ariya Rastrow*[†] *Björn Hoffmeister*[†]

[⋆] University of California, Berkeley [†] Amazon Alexa AI

ABSTRACT

While word error rates of automatic speech recognition (ASR) systems have consistently fallen, natural language understanding (NLU) applications built on top of ASR systems still attribute significant numbers of failures to low-quality speech recognition results. Existing assistant systems collect large numbers of these unsuccessful interactions, but these systems usually fail to learn from these interactions, even in an offline fashion. In this work, we introduce CLC: Contrastive Learning for Conversations, a family of methods for contrastive fine-tuning of models in a self-supervised fashion, making use of easily detectable artifacts in unsuccessful conversations with assistants. We demonstrate that our CLC family of approaches can improve the performance of ASR models on OD3, a new public large-scale semi-synthetic meta-dataset of audio task-oriented dialogues, by up to 19.2%. These gains transfer to real-world systems as well, where we show that CLC can help to improve performance by up to 6.7% over baselines.[1]

Index Terms— Task Oriented Dialogue, Automatic Speech Recognition, Self-Supervised Learning

1. INTRODUCTION & BACKGROUND

When users interact with assistant systems in task oriented ways, they build rich conversational contexts, which contain information that may be relevant to future requests along with feedback on the performance of the system. When users are dissatisfied, they express that intent in many ways, from direct corrections of the system response, to repeating and rephrasing the original question [1]. This discourse provides a source of contextual user interaction signals that are relatively untapped in Automatic Speech Recognition (ASR).

Indeed, traditional systems for ASR have primarily focused on single-utterances [2, 3, 4, 5, 6, 7], which, although flexible, overlook the wealth of contextual cues available in task directed dialogues. While work has been done in natural language understanding (NLU) to exploit these cues [8], their potential in ASR has remained largely unexplored, primarily due to the limited availability of task-driven dialogue datasets in the audio domain [9, 10]. Current efforts to integrate context from non-dialogue sources into ASR often involve training models explicitly with external per-turn contextual inputs, often leveraging context attention mechanisms [6, 9, 11, 12, 13, 14, 15, 16, 17, 18]. While per-turn context is important for the ASR task, these methods do not draw from dialogue structures, nor do they account for interactive feedback present in labeled dialogues.

Instead of directly training on per-sample or per-turn context (e.g., contact names [13, 14], or external dictionaries [6]), we explore the potential of learning implicit contextual signals of user interactions, which remains relatively untapped in ASR-based dialog systems. Following work demonstrating the benefits of contrastive learning in ASR [19], closest to our work may be Chang *et al.* [9] who propose reducing ASR errors with contrastive learning between noisy and clean audio transcripts from task-oriented dialogues – however, their work focuses only on single turns of dialogues, not contextual dialogue cues. Our primary contributions are:

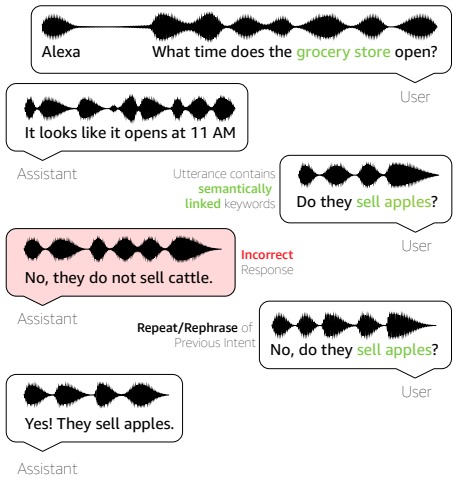

Fig. 1: Task oriented dialogues can contain a multitude of relevant information for performing automated speech recognition. In this work, we explore how we can learn from both semantically linked keywords within dialogues, and failed dialogue turns.

- We propose a new family of self-supervised fine-tuning losses, CLC, which incorporate self-supervised information from task oriented dialogues (TODs), and show that learning from TODs, even those with errors, provides benefits over fine-tuning.
- We introduce a new semi-synthetic benchmark meta-dataset, the Open Directed Dialogue Dataset (OD3), designed to enable further research in conversational interactions for ASR.

2. CONTRASTIVE LEARNING FOR CONVERSATIONS

In this work, we introduce two novel auxiliary losses, termed "Contrastive Learning for Conversations" (CLC), designed to enable learning from both successful and unsuccessful task-directed conversations with assistants (section 2), as well as a new synthetic dataset for the evaluation of contextual automated speech recognition models in task directed domains (subsection 2.1).

Learning from Past and Future Dialogues: As shown in Figure 1, utterances in a dialogue can contain important contextual hints useful

[1] Our Code/Data is publicly available at https://github.com/amazon-science/amazon-od3.

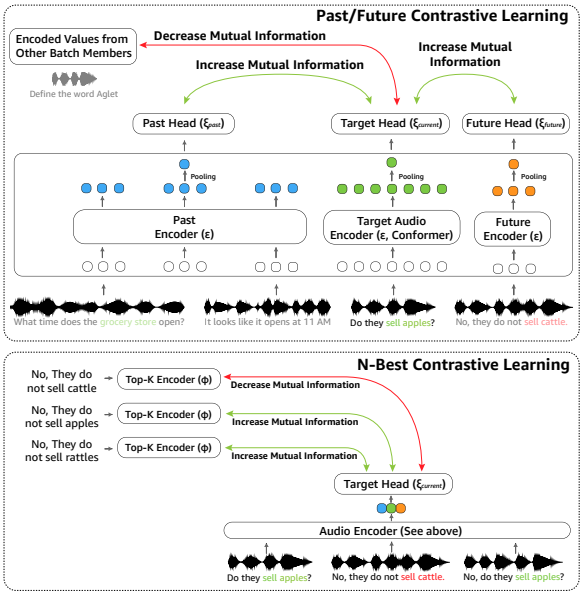

Fig. 2: Overview of CLC approaches. The Past-Future loss maximizes agreement between current, past, and future embeddings. The N-best loss minimizes agreement between current embeddings and top predictions of rephrases, while maximizing agreement otherwise.

for recognizing low-frequency words in the sentence. While we may not have access to past or future utterances at inference, we can often learn from these hints during training. The first auxiliary loss we introduce follows this key motivation; auditory information within a dialogue should share more semantic and representational overlap than auditory information from a second, unrelated dialogue.

This insight induces a natural contrastive loss: the speech encoder representations of audio within a session should be closer in the latent space (on average) than the representations between sessions. To implement a "Past-Future" contrastive loss, we consider the utterances $u_1,\ldots,u_N$ in a dialogue. Let the speech encoder be defined as $e_i = \epsilon(u_i) \in \mathbb{R}^{T\times k}$, where k is the dimension of the speech encoder embedding, and T is the number of frames of audio in the dialogue. We further introduce three "head" encoders, $\xi_{past}(e_i) \in \mathbb{R}^d, \xi_{current}(e_i) \in \mathbb{R}^d, \xi_{future}(e_i) \in \mathbb{R}^d$, which embed the sequential embeddings from the encoder ϵ of the current, past, and future frames into single vectors (of dimension d) representing the current, past, and future contexts. These head encoders take the form of global pooling followed by two layers of a shallow MLP with ReLU activations, LayerNorm, and Dropout. We can then compute the following contrastive loss terms (similar to [20]) for a batch of $1 \le i,j \le N$ samples (where embeddings are L2-normalized):

$$L^{i,j}_{future} = -\log\left[\frac{\exp(\xi_{current}(e_i)\cdot\xi_{future}(e_j)/\tau)}{\sum_{k=1}^{N}\exp(\xi_{current}(e_i)\cdot\xi_{future}(e_k)/\tau)}\right]$$

$$L^{i,j}_{past} = -\log\left[\frac{\exp(\xi_{current}(e_i)\cdot\xi_{past}(e_j)/\tau)}{\sum_{k=1}^{N}\exp(\xi_{current}(e_i)\cdot\xi_{past}(e_k)/\tau)}\right]$$

The "Past-Future" auxiliary loss is then a weighted sum:

$$L_{pf} = \frac{1}{N}\left[\alpha\sum_{i=1}^{N}L^{i,i}_{future} + \beta\sum_{i=1}^{N}L^{i,i}_{past}\right] \quad (1)$$

Here, we choose cosine distance (the dot-product) as the similarity function. The result of the loss function is that we aim to maximize the mutual information between the encoding of the current, future and past frames within a dialogue, while minimizing the mutual information between the current frames and frames from other dialogues. Note here that it's important that $e_i \neq e_j$, that is, the embedding of the past should not be identical to the embedding of the future (as they have different ASR content). Instead, we encourage high mutual information between the different segments, by leveraging contrastive learning on projection heads stemming from the shared representation. α and β are hyper-parameters which control the strength of the binding in the loss function, and τ is a temperature parameter. In our experiments, we found through grid hyper-parameter search of $\alpha,\beta \in [0.0001, 100]$ (logarithmic sweep) and $\tau \in [0.1, 1]$ (linear sweep) that $\alpha = 1.0, \beta = 0.7, \tau = 0.1$ is the most effective.

Learning from Failures: We can extract valuable semantic information from conversations, even those that don't proceed smoothly. It is often possible to detect dialogues where unsuccessful ASR has triggered repeats and rephrases of previous content by understanding when subsequent user turns have high semantic overlap, or tracking NLU failures in downstream systems. In these cases, we can further leverage contrastive learning to improve the performance of the model. Ideally, when there is a repeat or rephrase in a dialogue, we want to reduce the mutual information between the conformer encoder embedding of the initial turn triggering the repeat or rephrase, and the answer produced by the model in that dialogue. While we could use reinforcement learning to optimize for this signal (and it is interesting future work to do so), we often train models offline, and as the model trains, its decisions deviate from the original policy, leading to a breakdown in the learning process. Instead, as we know the "bad" solution, we can use supervised contrastive learning [20] to improve the model. When there is no rephrase, we want to increase the mutual information between the semantics of the top-1 prediction of the model and the current frames. When there is a rephrase, we want to decrease the mutual information between the semantics of the top-1 prediction, and instead encourage the model to produce a different output from the top-k. While it is possible that worse hypotheses with high similarity exist in the hypothesis set (leading to incorrect labels), we observe empirically that our models have high oracle WER, allowing this method to achieve a weak approximation to oracle re-ranking of the candidate set, which improves overall performance when smoothed over a large training set.

An overview of our n-best approach is given in Figure 2. For each sample u_i, let $\phi_1(u_i)\ldots\phi_K(u_i)$ be the semantic embeddings of the top-k predictions of the i'th utterance (using beam-search decoding) and $\xi_{current}(e_i)$ be an embedding of e_i for u_i. Using a similar set of heads to the network above, we compute positive and negative losses:

$$L^{i}_{pos} = -log\left[\frac{\exp(\xi_{current}(e_i)\cdot\phi_1(u_i)/\tau)}{\sum_{k=1}^{K}\exp(\xi_{current}(e_i)\cdot\phi_k(u_j)/\tau)}\right]$$

$$L^{i}_{neg} = -log\left[\frac{\max_{j\neq i}[\exp(\xi_{current}(e_i)\cdot\phi_j(u_i)/\tau)]}{\sum_{k=1}^{K}\exp(\xi_{current}(e_i)\cdot\phi_k(u_i)/\tau)}\right]$$

Let $\mathcal{R}$ be the set of utterances which trigger a repeat/rephrase, and $\mathcal{S}$ be the set of utterances which are considered successful. We can then combine the positive and negative losses as follows:

$$L_{nbest} = \frac{\gamma}{|\mathcal{R}|}\sum_{i\in\mathcal{R}}L^{i}_{neg} + \frac{\kappa}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}L^{i}_{pos}z \quad (2)$$

where γ and κ are hyper-parameters controlling the trade-off between

negative and positive reinforcement. Discovering the sets $\mathcal{S}$ and $\mathcal{R}$ can be challenging, however, we can detect repeats and rephrases with relatively high accuracy using semantic vector matching (such as matching BERT embeddings). Using grid search with $\gamma,\kappa \in [0.0001,100]$ (logarithmic sweep), we found $\gamma = 0.1, \kappa = 1.0$ was most effective.

2.1. Data

While a predominant portion of interactions with assistant systems revolves around task-directed dialogues, the availability of datasets (Table 4) encompassing task-directed audio interactions remains quite limited. Moreover, even within datasets that do incorporate such interactions, a conscious effort has been exerted to remove flawed turns (turns in which a dialogue assistant responds incorrectly, and must be corrected by the user). To evaluate our CLC methods for self-supervised fine-tuning, we use two datasets: a private collection of de-identified real-world conversations with a conversational assistant, and a new semi-synthetic meta-dataset, OD3, replicating flawed conversations often seen in real-world assistant interactions. The OD3 dataset is released as part of this work under the CC-BY-NC-SA (4.0) license.

2.1.1. Real-World (Internal) Data

To demonstrate the performance of our method, we train and evaluate our models on 130K hours of de-identified agent-centric task-directed dialogues constructed from independent interactions with a conversational assistant. These dialogues have a maximum of five utterances each (with an average of 1.2 turns per goal). Dialogues are constructed around a seed utterance by collecting interactions within $\rho = 90$ seconds on each size of the utterance. This process is repeated recursively until there are no more interactions. In the case that there are more than five utterances, we halve ρ, and repeat the process. This continues until either we have less than 5 utterances in the final set or we hit a minimum time gap of 15 seconds. During testing, only the past and current context is available to the model (the future remains hidden).

2.1.2. OD3: A new dataset for conversational learning

In addition to the results on real-world interactions in this paper, we further introduce a new semi-synthetic meta-dataset, OD3 (Open Directed Dialogue Dataset), which is designed to allow the community to explore further research into leveraging flawed conversational interactions to improve model performance. OD3 is a collection of 63K conversations (600K turns, 1,172 hours of audio) drawn from existing natural language task-oriented dialog datasets, and augmented with synthetic audio. OD3 is further augmented with turns containing *repeats* and *rephrases* of previous failed utterances. We compare our dataset with some others in the field in Table 1.

Constructing OD3: To construct OD3, we start with several seed datasets of natural language task oriented dialog data: KVRET [21], Multi-Woz [22], DSTC11 (Track 5) [23], NOESIS-II [24] and SIMMC-2.1 [25]. Here we focus on multi-turn dialogue, (as opposed to single-turn datasets such as those for question-answering like NMSQA [26]), as they contain the most relevant contextual information. This gives us a pool of $\approx$ 63K unique dialogues ($\approx$ 597K turns) containing no explicitly labeled errors or flaws. Because these datasets are not augmented with audio for each of the conversational turns, we leverage the NeMo Text Normalizer [27] and the YourTTS method [28] (voice cloning) to generate audio for each of the conversations. In all of the conversations, we hold the voice for the agent constant, and each voice used in voice cloning is randomly selected from the English subset of Common Voice [29] (which is CC0 licensed). We found that in some cases, the TTS induces errors in the generated speech, which we found correlated with a high number of deletions in the resulting ASR models. To clean the dataset, we filter out $\approx$4K utterances inducing a significant number of deletions in both our tested and third party ASR models. While we run our experiments in this paper on the clean data, we additionally release the noisy versions of the data as they could be useful for investigation into alternate directions of research.

Table 1: Statistics for OD3. OD3 is much larger than existing TOD datasets, while including both audio and noisy conversations.

Dataset	Dialogues	Turns	Audio	Errors
DSTC-2 [31]	1,612	23,354	✓	
KVRET [21]	2,425	12,732		
MultiWOZ [22]	8,438	115,424		
DSTC-10 [32]	107	2,292		
SpokenWOZ [10]	5,700	203,074	✓	
OD3 (Ours)	**62,974**	**623,145**	✓	✓

We synthetically introduce errors and noisy conversations into the data. For that, we first compute ASR for each dialog turn using OpenAI's Whisper Large (v2) model [2]. We consider conversational turns with WER higher than 15% candidates for the injection of either a *repeat*, or a *rephrase* of the intent. We then insert repeats and rephrases into 20% of the possible candidate conversations. To insert a *repeat*, we introduce two conversational turns: a response for the agent which is a non-specific error response (such as "I'm sorry, I don't understand"), and a repeat of the phrase which triggered the ASR errors (re-sampled from the original TTS model). Inserting a *rephrase*, on the other hand, is much more complicated. Similar to the case of repeats, we first introduce a non-specific agent error message. We then generate a rephrase of the original triggering utterance using in-context learning with the MPT-30B language model [30], combined with the prompt:

```
Our automated speech recognition model found
"<input string>" hard to parse, so we rephrased
it to use easier to understand words as "...
```

We found that this prompt generated reasonable rephrases of the candidate sentences. For example, *"Are there noisy neighbors?"* was rephrased as *"Is the place quiet enough?"*. This gives us a total of $\approx$ 625K turns of dialogue in $\approx$ 62K sessions, and 1,172 hours of audio.

2.2. Models

For the speech encoder ϵ, we use a conformer architecture [33], with 17 layers, latent dimension of 1024, and two stride-two convolutional sub-sampling layers ($\approx$ 200M parameters). We use a 1-layer LSTM decoder with latent dimension of 320, and a 4K token vocabulary. The encoder/decoder are initialized with a model pre-trained on 120K hours of de-identified internal seed data. During training, we apply both kernel regularization and bias regularization with weight $1e^{-6}$, and dropout with weight 1.0. We optimize the overall loss:

$$L_{overall} = L_{asr} + \lambda L_{pf} + \delta L_{nbest} \quad (3)$$

The models are trained for at most 120 epochs with the Adam optimizer, following a linear increase, hold, exponential decay learning rate schedule starting at $1e^{-8}$, increasing linearly over 50K steps to hold at $4e^{-5}$ for 250K steps, and then decay back to $1e^{-6}$ over a further 300K steps. We use gradient clipping with limit 0.3, and a dynamic batch size (depending on input feature length) ranging between 128 and 1024. As contrastive learning cannot naively be

Table 2: Results on internal data, both overall and only on turns inducing repeats or rephrases. WERR (↑): Percent relative WER Improvement. SERR (↑): Percent relative SER improvement.

Model	Overall		Repeats/Rephrase	
	WERR	SERR	WERR	SERR
Zero-Shot (No Fine Tuning)	-23.02%	-17.46%	-4.65%	-5.75%
Baseline (Fine Tuned)	-	-	-	-
CLC ($\lambda=1,\delta=0$)	2.75%	2.88%	3.0%	3.39%
CLC ($\lambda=0,\delta=1$)	2.60%	2.39%	3.75%	3.87%
CLC ($\lambda=1,\delta=1$)	4.31%	3.88%	5.07%	5.31%

Table 3: Results on internal data for different values of α and β ($\tau=0.1$) in L_{pf}, as well as γ and κ in L_{nbest} for small scale (batch size 128) experiments. WERR (↑): Relative WER Improvement. SERR (↑): Relative SER improvement.

Model	WERR	SERR
Baseline (CLC, $\lambda=0,\delta=0$)	-	-
CLC ($\alpha=1,\beta=0,\gamma=0,\kappa=0$)	3.28%	2.26%
CLC ($\alpha=0,\beta=1,\gamma=0,\kappa=0$)	2.74%	3.68%
CLC ($\alpha=1,\beta=1,\gamma=0,\kappa=0$)	4.50%	5.34%
CLC ($\alpha=1,\beta=0.7,\gamma=0,\kappa=0$)	5.17%	4.67%
CLC ($\alpha=0,\beta=0,\gamma=1,\kappa=0$)	-11.81%	-10.43%
CLC ($\alpha=0,\beta=0,\gamma=1,\kappa=1$)	-1.88%	-2.21%
CLC $\alpha=0,\beta=0,\gamma=0,\kappa=1$)	6.23%	5.59%
CLC ($\alpha=0,\beta=0,\gamma=0.1,\kappa=1.0$)	6.77%	6.25%

scaled across GPUs, we leverage techniques similar to BASIC [34] and perform memory efficient contrastive mini-batching.

3. RESULTS & DISCUSSION

We first demonstrate the performance of our method on our internal session data. From the results in Table 2, we can see that all three settings of CLC improve the overall WER/SER of the model, particularly over zero-shot models. We notice that setting $\lambda=1$ is the most effective at reducing overall WER, as in most situations, contextual information from previous (and future) turns can provide more powerful hints to the content of an utterance. While δ is helpful as well, it is less important to overall WER.

Table 3 shows the performance of CLC across different values of α and β for L_{pf}. We can see that taking into account both past and future information is important. Unsurprisingly, past information is a more powerful indicator of the current ASR context; however it's important to note that pre-training with the information from the future allows the model to improve the predictive ability of the audio representations, leading to improvements (particularly in SER). Table 3 also shows the performance for values of γ and κ in the L_{nbest} loss. We can see here that placing too much weight on the γ term leads to a destabilization of the loss, however small magnitude γ values can help with overall performance. We believe that this destabilization is caused by the high variance of the $\max_{j\neq i}[\exp(\xi_{current}(e_i)\cdot\phi_j(u_i)/\tau)]$ term, and it is future work to explore how functional implementations such as a softmax could reduce the gradient variance stemming from this loss term.

Table 2 also shows the performance of our method when restricted to only defective utterances: utterances triggering repeats and rephrases in the dataset. We can see that setting $\delta=1$ is helpful, since the additional losses nudge the model away from high-confidence decisions in detected repeats/rephrases and makes an impact on the model's ability to correctly recognize challenging samples. Note that

Table 4: Results on the OD3 dataset (overall and repeat/rephrase inducing). WER (↓): Word Error Rate, BERT-S (↑): Bert Score.

Model	Overall		Repeat/Rephrases	
	WER	BERT-S	WER	BERT-S
Baseline (206M)	11.13	0.9762	16.17	0.9690
CLC ($\lambda=1,\delta=0$)	9.57	0.9801	14.12	0.9702
CLC ($\lambda=0,\delta=1$)	9.38	0.9803	13.94	0.9721
CLC ($\lambda=1,\delta=1$)	8.99	0.9812	13.81	0.9737

Table 5: Zero-shot results on OD3 for several open-source models. Models in this table are not directly comparable (trained on differing data, setups, hyperparameters, optimizers etc.), but serve as a benchmark for performance on OD3 under several varying setups. WER (↓): Word Error Rate, BERT-S (↑): Bert Score.

Model	Overall		Repeat/Rephrases	
	WER	BERT-S	WER	BERT-S
CLC best model	8.99	0.9812	13.81	0.9737
Whisper S (200M) [2]	11.24	0.9775	14.17	0.9727
Whisper L (1.3B) [2]	8.51	0.9852	12.37	0.9792
Conformer (100M, Librispeech) [33]	19.26	0.9612	22.19	0.9571
Wav2Vec 2 (433M, Librispeech) [3]	19.41	0.9582	22.03	0.9544
Streaming Conformer (45M) [35]	14.38	0.9701	16.70	0.9665

WERR/SERR gains are statistically significant over the large-scale test set ($\approx 1K$ hours of test audio).

On OD3, our approach produces even more defined results, demonstrated in Table 4, where our model produces a 19.22% improvement over baselines, clearly showing how learning from additional contextual clues can benefit ASR models. Interestingly, despite a high word error rate, the semantic similarity, as indicated by the BERTScore [36] remains high — this suggests that ASR errors, while numerous, do not significantly impact the semantic meaning. Several major questions remain unanswered for future work, for example, it remains an open question how the approaches scale with model parameters, as well as understanding to what extent different mixes of pre-training data alter the performance of the model.

Even for models with strong language models, large vocabularies, and training data focused on open-domain conversational language, Table 5 shows that OD3 is challenging. Models demonstrated increased insertions and substitutions, as there are a large number of challenging low-frequency words that must be recognized accurately. It's interesting to see that the streaming conformer [35] (trained on Gigaspeech) outperforms some of the larger models. This is likely due to the training data mix: training smaller models on more robust datasets is more effective than training larger models on sparse or biased data.

4. CONCLUSION

This work introduces CLC, a self-supervised fine-tuning approach for enhancing contextual automated speech recognition (ASR) in task-oriented dialog systems. We also introduced OD3, the largest-ever dataset for task-oriented automated speech recognition. By leveraging both successful and unsuccessful conversational interactions, our method enhances the underlying ASR model's ability to handle challenging and contextually rich utterances. In real-world data, we demonstrate as much as 6.77% improvement over baselines. Further, for OD3 we show up to a 19.22% improvement over baselines. We hope that our approaches and datasets will help address ASR challenges within intricate and error-prone dialog settings, elevating user experiences and enabling more effective interactions between humans and AI agents.

5. REFERENCES

[1] W.-C. Kwan *et al.*, "A survey on recent advances and challenges in reinforcement learning methods for task-oriented dialogue policy learning," *Machine Intel. Res.*, vol. 20, no. 3, 2023.

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023.

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS*, 2020.

[4] W.-N. Hsu, Y.-H. H. Tsai, B. Bolte, R. Salakhutdinov, and A. Mohamed, "Hubert: How much can a bad teacher benefit asr pre-training?" in *ICASSP*, 2021.

[5] D. M. Chan, S. Ghosh, D. Chakrabarty, and B. Hoffmeister, "Multi-modal pre-training for automated speech recognition," in *ICASSP*, 2022.

[6] D. M. Chan, S. Ghosh, A. Rastrow, and B. Hoffmeister, "Domain adaptation with external off-policy acoustic catalogs for scalable contextual end-to-end automated speech recognition," in *ICASSP*, 2023.

[7] S. Mitra, S. N. Ray, B. Padi, R. Bilgi, H. Arsikere, S. Ghosh, A. Srinivasamurthy, and S. Garimella, "Unified modeling of multi-domain multi-device ASR systems," in *TSD*, 2023.

[8] B. Min *et al.*, "Recent advances in natural language processing via large pre-trained language models: A survey," *ACM Computing Surveys*, 2021.

[9] S.-Y. Chang *et al.*, "Context-aware end-to-end asr using self-attentive embedding and tensor fusion," in *ICASSP*, 2023.

[10] S. Si, W. Ma, H. Gao, Y. Wu, T.-E. Lin, Y. Dai, H. Li, R. Yan, F. Huang, and Y. Li, "Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents," *arXiv:2305.13040*, 2023.

[11] S. Kim and F. Metze, "Dialog-context aware end-to-end speech recognition," in *SLT*, 2018.

[12] F.-J. Chang, J. Liu, M. Radfar, A. Mouchtaris, M. Omologo, A. Rastrow, and S. Kunzmann, "Context-aware transformer transducer for speech recognition," in *ASRU*, 2021.

[13] Z. Chen, M. Jain, Y. Wang, M. L. Seltzer, and C. Fuegen, "Joint grapheme and phoneme embeddings for contextual end-to-end asr." in *Interspeech*, 2019.

[14] K. M. Sathyendra, T. Muniyappa, F.-J. Chang, J. Liu, J. Su, G. P. Strimel, A. Mouchtaris, and S. Kunzmann, "Contextual adapters for personalized speech recognition in neural transducers," in *ICASSP*, 2022.

[15] K. Wei *et al.*, "Attentive contextual carryover for multi-turn end-to-end spoken language understanding," in *2021 ASRU*. IEEE, 2021, pp. 837–844.

[16] C.-H. H. Yang, Y.-L. Gu, Y.-C. Liu, S. Ghosh, I. Bulyko, and A. Stolcke, "Generative asr error correction with large language models," in *ASRU*, 2023.

[17] D. M. Chan, S. Ghosh, A. Rastrow, and B. Hoffmeister., "Using external off-policy speech-to-text mappings in contextual end-to-end automated speech recognition," in *ICASSP*, 2023.

[18] S. Mahadevan, B. Mishra, and S. Ghosh, "A unified framework for domain adaptation using metric learning on manifolds," in *ECML PKDD*, 2019.

[19] D. M. Chan and S. Ghosh, "Content-context factorized representations for automated speech recognition," in *Interspeech*, 2022.

[20] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *arXiv:2004.11362*, 2021.

[21] M. Eric, L. Krishnan, F. Charette, and C. D. Manning, "Key-value retrieval networks for task-oriented dialogue," in *SIGDIAL*, 2017.

[22] P. Budzianowski *et al.*, "MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling," pp. 5016–5026, Oct.-Nov. 2018. [Online]. Available: https://aclanthology.org/D18-1547

[23] C. Zhao, S. Gella, S. Kim, D. Jin, D. Hazarika, A. Papangelis, B. Hedayatnia, M. Namazifar, Y. Liu, and D. Hakkani-Tur, ""what do others think?": Task-oriented conversational modeling with subjective knowledge," in *arxiv:2305.12091*, 2023.

[24] C. Gunasekara *et al.*, "Noesis ii: Predicting responses, identifying success, and managing complexity in task-oriented dialogue," in *AAAI: Workshop on Dialog System Tech Challenges*, 2020.

[25] S. Kottur, S. Moon, A. Geramifard, and B. Damavandi, "SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations," in *EMNLP*, 2021.

[26] G.-T. Lin, Y.-S. Chuang, H.-L. Chung, S. wen Yang, H.-J. Chen, S. A. Dong, S.-W. Li, A. Mohamed, H. yi Lee, and L. shan Lee, "DUAL: Discrete Spoken Unit Adaptive Learning for Textless Spoken Question Answering," in *Interspeech*, 2022.

[27] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook *et al.*, "Nemo: a toolkit for building ai applications using neural modules," *arXiv:1909.09577*, 2019.

[28] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "YourTTS: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *ICML*, 2022.

[29] R. Ardila *et al.*, "Common voice: A massively-multilingual speech corpus," pp. 4218–4222, May 2020. [Online]. Available: https://aclanthology.org/2020.lrec-1.520

[30] M. N. Team. (2023) Introducing mpt-30b: Raising the bar for open-source foundation models. Accessed: 2023-06-22. [Online]. Available: www.mosaicml.com/blog/mpt-30b

[31] M. Henderson, B. Thomson, and J. D. Williams, "The second dialog state tracking challenge," in *SIGDIAL*, 2014.

[32] S. Kim *et al.*, ""how robust ru?": Evaluating task-oriented dialogue systems on spoken conversations," in *ASRU*, 2021.

[33] 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," *arXiv:2005.08100*, 2020.

[34] H. Pham *et al.*, "Combined scaling for zero-shot transfer learning," *Neurocomputing*, 2023.

[35] E. Tsunoo, Y. Kashiwagi, and S. Watanabe, "Streaming transformer asr with blockwise synchronous beam search," in *SLT*, 2021.

[36] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *ICLR*, 2019.