INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Speech Aware Dialog System Technology Challenge (DSTC11)

Hagen Soltau[1], *Izhak Shafran*[1], *Mingqiu Wang*[1], *Abhinav Rastogi*[1], *Jeffrey Zhao*[1], *Ye Jia*[2], *Wei Han*[1], *Yuan Cao*[1], *Aramys Miranda*[1]

[1]Google DeepMind, USA
[2]tomato.ai, USA

soltau,izhak,mingqiuwang,abhirast,jeffreyzhao,weihan,yuancao,aramys@google.com,
jiaye@tomato.ai

Abstract

Most research on task oriented dialog modeling is based on written text input. However, practical dialog systems often use spoken input. Typically, input speech is converted into text using an Automatic Speech Recognition (ASR) systems, which are error-prone. Furthermore, most systems don't address the differences in written and spoken language (e.g., disfluencies). The research on this topic is stymied by the lack of a public corpus. Motivated by these considerations, our goal in hosting the speech-aware dialog state tracking challenge was to create a public corpus or task which can be used to investigate the performance gap between the written and spoken forms of input, develop models that could alleviate this gap, and establish whether Text-to-Speech-based (TTS) systems is a reasonable surrogate to the more-labor intensive human data collection. We created and released three spoken versions of the popular written-domain MultiWoz task – (a) TTS-Verbatim: written user inputs were converted into speech waveforms using a TTS system, (b) Human-Verbatim: humans spoke the user inputs verbatim, and (c) Human-paraphrased: humans paraphrased the user inputs. Additionally, we provided different forms of ASR output to encourage wider participation from teams that may not have access to state-of-the-art ASR systems. These included ASR transcripts, word time stamps, and latent representations of the audio (audio encoder outputs). In this paper, we describe the corpus, report results from participating teams, provide preliminary analyses of their results, and summarize the current state-of-the-art in this domain.

1. Introduction

In recent years, Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) models have converged to utilize common components like Transformers and encoder/decoder modules. They increasingly rely on large amounts of data, large model sizes and large amounts of compute resources. This is a substantial departure from a previous era when ASR and NLP utilized different modeling architectures, chosen to inject domain-specific knowledge and constraints. This shift to a common paradigm has stimulated research in fusing audio and text modalities to substantially improve performance in tasks such as audio intent classification and speech-to-speech translation. The fusion of modalities in many cases allows the direct optimization of the end-task, overcoming the hurdles of the older cascaded approaches that often led to accumulation of errors.

Despite the general trend to develop end-to-end models for various tasks, spoken dialog systems stick out as a sore thumb. Most practical systems utilize a cascaded approach where the output of a general ASR system is fed into a dialog model trained separately on written domain. This mismatch between written and spoken inputs to the dialog models is not well-studied, largely due to the lack of a public task with spoken user inputs.

Research into combined audio-text models is limited by the lack of paired data. While the paired data requirement can be relaxed for training data via un- or self-supervised training techniques, test sets with paired data are crucial for model evaluation. In addition to an evaluation task, a training set with spoken input would also be helpful in quantifying the gains from supervised learning and recent advances in self-supervised learning.

The focus of our effort was to bring most benefit to the community with the limited resources available. While a Wizard-of-Oz style data collection in spoken domain would have been ideal to fully investigate all the phenomena of spoken domain, that would be extremely labor-intensive especially in annotating the dialog states and was beyond the scope of our effort. Instead, we chose to create a spoken version of a well-studied written-domain task, the MultiWoz task. One advantage of this approach was that we could directly compare the performance of the spoken version with the written domain.

In the Speech-Aware DSTC11 challenge, participants are asked to infer the dialog states from the sequence of agent (text-input) and user (audio-input) turns. We evaluated the performance on three versions of audio inputs – TTS-Verbatim, Human-Verbatim (humans speaking the written user inputs), and Human-Paraphrased (humans paraphrasing the written user inputs). Aside from the audio-inputs, we provided audio encodings and transcripts from a state-of-the-art ASR system trained on 33k hours of People's Speech corpus to encourage participation from teams that did not have an easy access to ASR systems.

In the course of developing this challenge, we developed a cascaded baseline system with data augmentation and report performances on a few variants of cascaded systems. In the process, we uncovered a bias in the MultiWoz evaluation task, the slot values in the evaluation set have a substantial overlap with those of the training set. To address this bias, we created a new version of the MultiWoz evaluation set, the dev-dsct11 and eval-dstc11. We observed that the new task surfaces many of the challenges in practical spoken dialog systems associated with mismatch in modalities, inability to recover from ASR errors, and more generally difficulty of extracting semantically relevant information from audio signals.

2. Related Work

Motivated by similar consideration as ours, [1] organized a DST Challenge in 2021 where they created a task using spoken human-human dialogs on tourist information in San Francisco for three

10.21437/Interspeech.2023-1037

target domains: hotel, restaurant, and attraction. One of the serious limitations of the challenge was that the audio data was not released, only the ASR transcripts. The transcripts had an error rate of about 26.25% which is significantly higher than the average performance of most state-of-the-art ASR systems. The larger focus of their effort was on evaluating correctness of detecting knowledge-seeking turns, identifying the knowledge snippets and knowledge supplied in the generated responses. Prior to this effort, there have been much smaller efforts with fewer domains and dialogs in DSTC2 and DSTC3 [2, 3]. Here again the organizers only provided ASR transcripts with error rates in high 20s and low 30s, which limited the utility of the corpus as the ASR systems improved over time.

Meanwhile considerable progress has been achieved in improving the naturalness of dialog systems, for example, with chat-bots like Meena [4], raising expectations of interacting with dialog systems using spoken language. Similarly, the convergence of model architectures for ASR and NLP has stimulated research in creating joint audio-text encoders that could potentially compensate for ASR errors and other spoken language phenomena [5, 6, 7, 8, 9, 10, 11], where they propose different approaches to align the speech input (frames, phones, utterances) to the corresponding text units. However, none of them have been evaluated on dialog models due to the lack of a dialog task and corpus with audio input. We hope the task and corpus released in this work will bridge this gap along with other recent speech understanding tasks such as superb and slurp [12, 13]

3. Data

We chose to create a spoken version of MultiWoz [14] (2.1 version) so we could directly compare the written and spoken versions and reuse the annotations of dialog state labels, avoiding the labor-intensive process of annotating reference labels. Our corpus is freely available at `https://dstc11.dstc.community`.

3.1. Redesigned DSTC11 Evaluation Sets

Before launching into the data collection and in the process of developing baseline systems, we noticed that there is a substantial overlap in slot values between the training and evaluation sets, leading to overestimation of performance of the models that memorize the slot values, as reported elsewhere [15]. To illustrate the issue, we probed an existing DST model with input whose slot values were replaced with other viable values (e.g., *cambridge* to *new york*, *ely* to *dublin*), *17:43* to *17:41* and *19:00* to *19:12*). We chose a state-of-the-art model on the task with JGA of 55.4% [16]. The model ignores the new slot values and regurgitates the original slot values memorized from the training data, as shown in the Table 1.

Text Input	Text Output
i want to go to new york	dest=cambridge
i want to go to dublin	dest=ely
train leaving at 17:41	time=17:43
train leaving at 19:12	time=19:00

Table 1: *Examples of model outputting memorized values.*

This overlap in slot values between the training and evaluation sets of MultiWoz will mask the effect of misrecognition of the slot values by ASR systems in a practical spoken dialog system. For a fair evaluation of the models, we redesigned the MultiWoz evaluation sets to replace the slot values in the evaluation sets with new slot values as described below. The replacements were performed at the dialog level to maintain consistency across turns.

1. *Location Names*: The destination and departure cities for trains and buses were replaced with randomly sampled city names from 12655 cities in the United States.
2. *Hotel Names*: Hotel names were replaced with random names sampled from 1562 hotels in the United States.
3. *Restaurant Names*: Restaurant names were replaced by sampling from 214 restaurants in New York City.
4. *Time Slots*: Timestamps were offset by a random value across all the times mentioned in the dialog.

We measured the impact of the redesigned evaluation sets, henceforth referred to as dev-dstc11 and test-dstc11, using two dialog models – a seq-to-seq model [16] and a model that utilizes the power of large language model by fine-tuning the prefix encoding with dialog-specific instructions [17]. The second model has three variants – D3ST-base, D3ST-large and D3ST-XXL – related to the size of the underlying T5x large language model [18]. The results, shown in the Table 2, confirms our overestimation of the original MultiWoz evaluation sets. The actual performance is almost 50% worse than reported on the original evaluation sets for most models with the exception of D3ST-XXL.

Model	org. Dev	Dev-DSTC11
seq2seq	55.4	20.8
D3ST-base	54.2	22.0
D3ST-large	54.5	25.2
D3ST-xxl	57.8	43.1

Table 2: *Performance gap, in JGA, between biased (dev) and unbiased (dev-dstc11) with the same model.*

The memorization of the slot values completely distorts the effect of ASR errors on the downstream dialog models and paints an overly rosy picture. We illustrate this with a simple cascaded baseline where an ASR system transcribes the input speech into written form which is fed into a seq-to-seq dialog model [16]. In the Table 3 we compare the performance degradation from switching from written to spoken input and then retraining the seq-to-seq model on the ASR transcripts. According to the performance on the original dev set, the degradation from written to ASR transcripts is 58.1% to 47.2%. Furthermore, most of the degradation is recovered by retraining on the ASR transcripts. Both these results are incorrect and misleading as will show in the Section 5.3, training on ASR output is not nearly as effective as the results in the table suggests.

Train Data	Test Data	JGA
Written	Written	58.1
Written	Spoken/ASR	47.2
Spoken/ASR	Spoken/ASR	56.0

Table 3: *Misleading performance of seq-to-seq model on written and ASR transcripts.*

3.2. Text-to-Speech Version

One of the questions we were interested in understanding was whether TTS is a good substitute for speech collected from humans. For answering this question, we generated *TTS-Verbatim*, a TTS version of the evaluation test and training sets using the

system described in [19], a system that represents phonemes and graphemes to represent input text. Additionally, for training data, we generated four versions of the training data, each corresponding to a different TTS speaker. There was no overlap between the speakers in the training and evaluation sets.

3.3. Human Data Collection

We focused our data collection on the user turns since the agent turns are already available to any practical dialog system. The data collection was performed via Amazon Mechanical Turk and consisted of two versions – *Human-Verbatim* and *Human-Paraphrased*. Crowd-workers were presented a full dialog in text including both user and agent turns. In the verbatim version, the workers were instructed to utter the user turns verbatim as naturally as possible, one at a time till the end of the dialog. In the paraphrased version, the workers were instructed to paraphrase the user turn preserving the semantic meaning and the entities (e.g., names, times). After they finished the dialog, the workers transcribed their own paraphrased utterances.

The quality of the recordings were measured and only those above certain quality were retained. The pertinent factors include: 1) missing audio, 2) incomplete dialog, 3) unintelligible speech, 4) high background noise, 5) speakers uttering verbatim when they should have paraphrased, and 6) speakers not transcribing their paraphrased utterance.

We developed objective measures to perform quality control in bulk. The missing audio and incomplete dialogs were detected programmatically. To filter out utterances with unintelligible speech and high background noise, we transcribed the collected user utterances with an ASR system (for details of the system see Section 5) and measured the accuracy with respect to the corresponding transcripts. Since insertions may occur due to disfluencies, we used deletions as a stronger quality indicator and used a threshold of 25%. For the paraphrased version we used additional criteria: 1) paraphrased utterances had at least 60% as many words as verbatim, 2) the paraphrased version differed from the verbatim version by at least 45% as indicated by the WER between recognized words and the written user prompt, and 3) less than 30% WER with respect to the crowd-worker generated transcripts.

4. Challenge Task and Evaluation

The main task in the DSTC11 Challenge is to infer the dialog states correctly from the audio inputs that were provided corresponding to the redesigned MultiWoz test set, redesigned as described in Section 3.1. For ablation studies, the participants were required to submit their results for three conditions: *TTS-Verbatim*, *Human-Verbatim* and *Human-paraphrased*, which were collected as described in Section 3.2 and 3.3. The participants were free to use their own ASR system or the outputs provided from our baseline ASR system, described in Section 5.1.

The performance in the challenge is measured using the standard Joint Goal Accuracy (JGA) as a primary metric and Slot Error Rate (SER) as a secondary metric for fine-grained comparisons. In the literature, results are often difficult to compare since research groups apply their own output normalization before scoring the results. To sidestep the resulting confusion, this challenge is evaluated using standard evaluation tool published on GitHub [20]. The participants were allowed to utilize any publicly available data or checkpoints, but no private data to allow fair comparisons. They were allowed to augment data and no restrictions were imposed on model sizes or computational costs.

5. Baseline

5.1. ASR and Related Outputs

We provided the output of a baseline ASR system to reduce the barrier for participation by teams which didn't have ASR systems available to them. We trained an RNN-T [21, 22] ASR model on the PeopleSpeech public corpus [23] of approx. 32,000 hours of audio data. The encoder consists of 16 Transformer-XL [24] layers, and the language model (decoder) is a single LSTM layer, altogether a 220m parameter model. The performance of the model on the MultiWoz evaluation sets are reported in Table 4.

	TTS-Verbatim	Human-Verbatim
dev-dstc11	8.1(5.7/0.3/2.1)	11.9(7.7/2.7/1.5)
test-dstc11	8.2(5.7/0.3/2.1)	13.0(8.2/3.4/1.5)

Table 4: *Baseline ASR performance WER(Sub/Del/Ins)*

The following outputs are available as part of the DSTC11 challenge corpus.

1. *Audio Waveforms (16Khz/16-bit PCM)*: This allows researchers to investigate DST models that directly operate on audio such as end-to-end systems.
2. *Audio Encodings (75ms frame rate, 1024-dim)*: These are the output activation from the last Transformer layer in the audio encoder. The audio encoder reduces the logmel frame rate of 25ms to 75ms. These encodings are provided to support research in loosely-coupled ASR-DST cascaded systems.
3. *ASR Hypotheses with Time Alignment*: The ASR recognition outputs are provided to support research in different cascaded ASR-DST systems. Time alignments allow teams to pinpoint the location of recognized tokens in the audio encodings. Note, due the nature of RNN-T, multiple output tokens (words) can correspond to the same audio frame.

Typical recognition errors include time formatting issues, spoken single-digit numbers, split words, and more general misspelled content words, where the last category is very relevant for DST purposes.

5.2. Data Augmentation

Data augmentation is a common technique to improve accuracy and robustness. Since our preliminary results in Table 2 unearthed problems related to memorization, we felt the need to incorporate data augmentation into our baseline systems. We created new versions of user responses for training data by replacing slot values with randomly picked city names, time offsets and restaurant names, as described in Section 3.1. The new slot names were drawn from a different list than the ones used for generating the redesigned evaluation sets. In all, we generated about 100x training data.

5.3. Cascaded ASR-DST System

For a baseline, we created a cascaded ASR-DST system where the transcripts from the ASR model in Section 5.1 were fed as input to the D3ST model [17], a seq-to-seq model [25, 26], where the input to this model contains a prompt which describes slot names in short natural language descriptions along with

potential values. Results are shown in table 5.

Test Data	Training Data			
	Text (Ref)		ASR Hyp	
	1x	100x	1x	100x
Text	43.1	52.8	33.9	43.0
TTS-Verbatim	27.3	32.1	26.8	38.4
H-Verbatim	23.6	27.9	23.7	31.8
H-Paraphrased	21.8	26.1	21.9	30.9

Table 5: *ASR-D3ST* **xxl** *cascaded model (JGA)*

We want to outline some observations from these results:

1. TTS (TTS-Verbatim) shows a degradation with respect to written version. However, the degradation is larger with Human-Verbatim (41.0% vs. 33.5% JGA), confirming our suspicion that we cannot rely on TTS as a surrogate for human speech.
2. Surprisingly, the drop in performance from Human-Verbatim to Human-Paraphrased is not very large. This is not because the two versions are similar since quality checks described in Section 3.3 assures us that the paraphrased version is sufficiently different from verbatim version.
3. Scaling the D3ST model from D3ST-Large (3B params) to D3ST-XXL (11B params) has substantial impact on performance (24.8% vs 33.5% JGA). This raises a question of fairness since teams may not have resources to work with larger model sizes. Nonetheless, we did find examples in results provided by participants were smaller models were able to outperform larger models, see Section 6 for details.
4. Training DST models with ASR hypotheses is simple and improves the performance substantially from 28.2% to 33.5% even when the training data is based on TTS.
5. Data augmentation gives consistent gains across all conditions.

6. Results

We received 11 submissions from 6 teams and the results are shown in figure 1. The full set of numbers are also available in the link under the DSTC11 Challenge website `https://dstc11.dstc.community`. We provide a high level overview of the results, but refer to the team's system descriptions for details. While we provided data to build direct audio-to-dst and tightly coupled models, all teams chose a cascaded approach with separately trained ASR and DST models. Many teams employed an explicit ASR error correction model and re-trained their DST models on ASR hypotheses together with various forms of TTS-based data augmentation. Within this general approach teams experimented with several variations and as a result the performance across submissions vary substantially. The highest performing submission obtains a JGA of 37.9% while the lowest performance is at 18.2%.

6.1. Alternative ASR models

While we provided ASR output with our baseline ASR system, described in Section 5.1 based on *PeopleSpeech* corpus, three submissions used *Whisper* instead [27]. We compared the two ASR models to tease apart the differences. As reported in Table 6, not surprisingly, we found that *Whisper* transcribes the evaluation sets more accurately than our baseline model (see Table 4) since it is trained on a magnitude order more data. We evaluated *Whisper* in the cascaded ASR-DST system with two models, one trained on 100x augmented written text and one

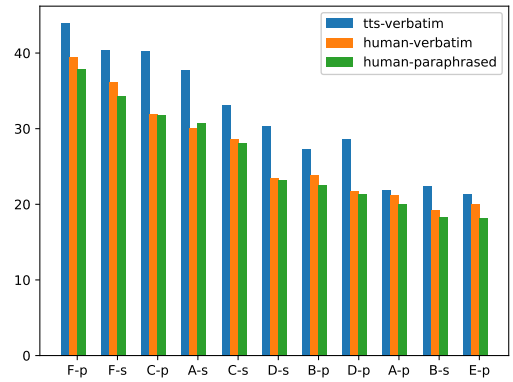

Figure 1: *Joint Goal Accuracy (JGA) of team's submissions*

trained on *Whisper* transcripts of the TTS version of the same training data. The results shown in Table 7 clearly demonstrate that the improvements in transcription accuracy translates to improvement in DST accuracy.

	TTS-Verbatim	Human-Verbatim
dev-dstc11	4.8(3.8/0.6/0.4)	8.5(5.8/1.4/1.4)
test-dstc11	4.6(3.7/0.6/0.3)	8.9(6.1/1.5/1.3)

Table 6: *ASR Performance of Whisper on evaluation sets.*

Test Data	100x Training Data	
	Text (Ref)	ASR Hyp
Text	52.7	45.3
TTS-Verbatim	35.6	41.3
H-Verbatim	32.3	35.5
H-Paraphrased	30.9	34.3

Table 7: *JGA of cascaded Whisper trained D3ST-XXL.*

7. Conclusions

In this paper, we describe a new corpus for stimulating research in modeling spoken dialogs that builds on popular written dialog corpus, MultiWoz. Releasing a spoken version of the same evaluation set allows researchers to study and bridge the performance gap between written and spoken dialog models. We observed that there is substantial overlap between the slot values in the training and evaluation sets of the original MultiWoz corpus and redesigned the evaluation sets by sampling new non-overlapping slot values and show that the new sets captures the weakness of the written dialog models better. We released three versions of the task – TTS-Verbatim, Human-Verbatim and Human-Paraphrased.

While the performance improves with model size and data augmentation, even the best models show substantial drop in performance when switching from written version to the spoken version (53.8% to 26.1% JGA). Retraining the D3ST-XXL model on the ASR hypotheses, improves the performance to 30.9% JGA but still leaves substantial ground to be covered. We report the results of 11 submissions from the DSTC11 Challenge. The dominant paradigm across teams was to rely on large language models for DST. Several submissions inserted ASR error correction modules of different complexities.

One area of research that is not well explored is better utilization of the latent representations of the audio encoder in the ASR in the downstream DST models. Similarly, we hope the release of the audio and the audio encoders outputs will allow researchers to evaluate the power of joint audio-text encoders on dialog tasks.

8. References

[1] S. Kim, Y. Liu, D. Jin, A. Papangelis, B. Hedayatnia, K. Gopalakrishnan, and D. Hakkani-Tur, "Knowledge-grounded task-oriented dialogue modeling on spoken conversations track at dstc10," in *AAAI 2022 Workshop on Dialog System Technology Challenge*, 2022.

[2] M. Henderson, B. Thomson, and J. D. Williams, "The second dialog state tracking challenge," in *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Philadelphia, PA, U.S.A.: Association for Computational Linguistics, Jun. 2014, pp. 263–272. [Online]. Available: https://aclanthology.org/W14-4337

[3] M. Henderson, B. Thomson, and J. Williams, "The third dialog state tracking challenge," *2014 IEEE Spoken Language Technology Workshop (SLT)*, pp. 324–329, 2014.

[4] D. Adiwardana, M. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, and Q. V. Le, "Towards a human-like open-domain chatbot," *CoRR*, vol. abs/2001.09977, 2020.

[5] J. Drexler and J. R. Glass, "Explicit alignment of text and speech encodings for attention-based end-to-end speech recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019*. IEEE, 2019, pp. 913–919.

[6] J. Chen, X. Tan, Y. Leng, J. Xu, G. Wen, T. Qin, and T.-Y. Liu, "Speech-t: Transducer for text to speech and beyond," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, 2021, pp. 6621–6633.

[7] Y. Chung, C. Zhu, and M. Zeng, "Semi-supervised speech-language joint pre-training for spoken language understanding," *CoRR*, vol. abs/2010.02295, 2020. [Online]. Available: https://arxiv.org/abs/2010.02295

[8] Y. Tang, H. Gong, N. Dong, C. Wang, W.-N. Hsu, J. Gu, A. Baevski, X. Li, A. Mohamed, M. Auli, and J. Pino, "Unified speech-text pre-training for speech translation and recognition," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1488–1499. [Online]. Available: https://aclanthology.org/2022.acl-long.105

[9] A. Bapna, Y. Chung, N. Wu, A. Gulati, Y. Jia, J. H. Clark, M. Johnson, J. Riesa, A. Conneau, and Y. Zhang, "SLAM: A unified encoder for speech and language modeling via speech-text joint pre-training," *CoRR*, 2021.

[10] A. Bapna, C. Cherry, Y. Zhang, Y. Jia, M. Johnson, Y. Cheng, S. Khanuja, J. Riesa, and A. Conneau, "mslam: Massively multilingual joint pre-training for speech and text," *CoRR*, vol. abs/2202.01374, 2022.

[11] Z. Chen, Y. Zhang, A. Rosenberg, B. Ramabhadran, P. J. Moreno, A. Bapna, and H. Zen, "MAESTRO: matched speech text representations through modality matching," *CoRR*, vol. abs/2204.03409, 2022.

[12] S. Yang, P. Chi, Y. Chuang, C. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G. Lin, T. Huang, W. Tseng, K. Lee, D. Liu, Z. Huang, S. Dong, S. Li, S. Watanabe, A. Mohamed, and H. Lee, "SUPERB: speech processing universal performance benchmark," *CoRR*, vol. abs/2105.01051, 2021.

[13] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, "SLURP: A spoken language understanding resource package," *CoRR*, vol. abs/2011.13205, 2020.

[14] M. Eric, R. Goel, S. Paul, A. Kumar, A. Sethi, P. Ku, A. K. Goyal, S. Agarwal, S. Gao, and D. Hakkani-Tur, "Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines," 2019. [Online]. Available: https://arxiv.org/abs/1907.01669

[15] X. Song, L. Zang, Y. Su, X. Wu, J. Han, and S. Hu, "Data augmentation for copy-mechanism in dialogue state tracking," *CoRR*, vol. abs/2002.09634, 2020.

[16] J. Zhao, M. Mahdieh, Y. Zhang, Y. Cao, and Y. Wu, "Effective sequence-to-sequence dialogue state tracking," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2021.

[17] J. Zhao, R. Gupta, Y. Cao, D. Yu, M. Wang, H. Lee, A. Rastogi, I. Shafran, and Y. Wu, "Description-driven task-oriented dialog modeling," *CoRR*, vol. abs/2201.08904, 2022.

[18] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 140:1–140:67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[19] Y. Jia, H. Zen, J. Shen, Y. Zhang, and Y. Wu, "Png bert: Augmented bert on phonemes and graphemes for neural tts," in *Proc. Interspeech*. ISCA, 2021.

[20] X. Zang, A. Rastogi, S. Sunkara, R. Gupta, J. Zhang, and J. Chen, "MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines," in *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*. Online: Association for Computational Linguistics, Jul. 2020.

[21] A. Graves, "Sequence transduction with recurrent neural networks," *CoRR*, 2012.

[22] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proceeedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 6645–6649.

[23] D. Galvez, G. Diamos, J. Ciro, J. F. Cerón, K. Achorn, A. Gopi, D. Kanter, M. Lam, M. Mazumder, and V. J. Reddi, "The people's speech: A large-scale diverse english speech recognition dataset for commercial usage," *CoRR*, vol. abs/2111.09344, 2021.

[24] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2978–2988. [Online]. Available: https://aclanthology.org/P19-1285

[25] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf

[26] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.0473

[27] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *OpenAI Blog*, 2022.