Improving Textless Spoken Language Understanding with Discrete Units as Intermediate Target

Guan-Wei Wu[1†], *Guan-Ting Lin*[1†], *Shang-Wen Li*[2], *Hung-yi Lee*[1]

[1]National Taiwan University, Taiwan
[2]Meta AI, USA

{b08901019, f10942104, hungyilee}@ntu.edu.tw

Abstract

Spoken Language Understanding (SLU) is a task that aims to extract semantic information from spoken utterances. Previous research has made progress in end-to-end SLU by using paired speech-text data, such as pre-trained Automatic Speech Recognition (ASR) models or paired text as intermediate targets. However, acquiring paired transcripts is expensive and impractical for unwritten languages. On the other hand, Textless SLU extracts semantic information from speech without utilizing paired transcripts. However, the absence of intermediate targets and training guidance for textless SLU often leads to suboptimal performance. In this work, inspired by the content-disentangled discrete units from self-supervised speech models, we proposed to use discrete units as intermediate guidance to improve textless SLU performance. Our method surpasses the baseline method on five SLU benchmark corpora. Additionally, we find that unit guidance facilitates few-shot learning and enhances noise robustness.

Index Terms: Spoken Language Understanding, Textless NLP, Self-Supervised Learning

1. Introduction

Spoken Language Understanding (SLU) is a task that aims to capture semantic information from spoken utterances. Recent advancements in SLU have led to successful applications in various industries, such as voice assistants and voice-controlled smart devices. The traditional approach to SLU involves a cascaded system comprising an Automatic Speech Recognition (ASR) model followed by a text-based Natural Language Understanding (NLU) model. This approach has shown promising results, but training the ASR model requires a large amount of labeled data, which can be challenging and costly to obtain. Additionally, there are many languages that are primarily spoken and lack written sources, making it impossible to build an ASR system for these languages. Another research direction is *end-to-end SLU* [1], which intends to predict semantic labels directly from speech features. However, previous studies rely on using a pre-trained ASR model as model initialization or jointly training ASR/NLU and SLU simultaneously [2, 3, 4, 5, 6, 7] with paired transcripts guidance.

To alleviate the reliance on paired transcripts, we step forward to *Textless SLU* [8], which extracts the semantic information without the need for paired transcripts[1]. Our approach draws inspiration from recent advancements in Self-Supervised Learning (SSL) of speech [9, 10, 11] and Textless

---

[†]Equal contribution.

[1]Noting that if the task requires to predict entity names, such as slot filing, the entity names are still utilized for training, while all other parts of transcripts are discarded.
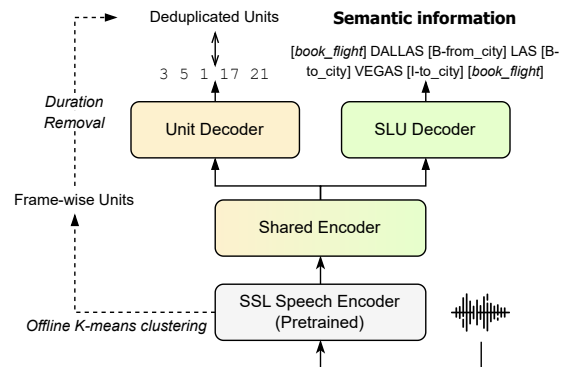

Figure 1: *The proposed textless SLU framework with discrete units as the intermediate target. The transformer encoder is shared between the unit decoder and SLU decoder. Noting that we train the model to predict only semantic information without using full transcripts of the utterances.*

NLP [12, 13, 14, 15, 16]. It has been observed that the self-supervised discovered units are highly related to content information [13]. Based on this observation, we hypothesize that discrete units can effectively serve as unsupervised targets for capturing content information in SLU.

In this work, we propose to leverage self-supervised discovered speech units as the intermediate target for end-to-end textless SLU. Our approach involves incorporating a transformer-based encoder-decoder model for the main SLU task, along with a separate decoder that shares the same encoder to predict discrete units. The underlying rationale is that discrete units primarily encode content-related information. By using them as guidance, we encourage the intermediate representation to prioritize content information, which is crucial for SLU tasks.

To evaluate our approach, we conduct experiments on five publicly available SLU benchmarks: ATIS [17], SLUE-SNER [18], SLURP [19], SNIPS [20], and STOP [21]. The experimental results demonstrate that our proposed method generally enhances textless SLU performance compared to the baseline and achieves competitive performance comparable to methods utilizing paired transcripts. For instance, on the ATIS dataset, we observe a 5.8% improvement in F1-score and a 3.6% improvement in intent accuracy compared to the approach by Morais *et al.* [3] under a similar architecture when not using text-speech paired data.

Moreover, our findings suggest that the proposed method exhibits robustness in few-shot and noisy scenarios. The model trained with unit guidance consistently outperforms the baseline approach in situations with limited data or domain mis-

- **Transcript** : we cannot influence tariff reductions that would benefit the people of the united kingdom
- **SNER** : united **B-PLACE** kingdom **I-PLACE**

- **Transcript** : i want to fly from baltimore to dallas round trip
- **ICSF** : **atis_flight** baltimore **B-fromloc.city_name** dallas **B-toloc.city_name** round **B-round_trip** trip **I-round_trip atis_flight**

- **Transcript** : stop the work timer
- **Parse Tree** : **[IN:PAUSE_TIMER** stop the work **[SL:METHOD_TIMER** timer **]]**
- **SSP** : **[IN:PAUSE_TIMER [SL:METHOD_TIMER]]**

Figure 2: *Transcriptions and sequence configurations for three SLU tasks, including SNER, ICSF, and SSP.*

match. This indicates that our work can enhance the performance and generalizability of the model when faced with data scarcity or variations in domain. As discrete units predominantly contain content-related information, the shared encoder can prioritize content information by learning to predict these units. This focus on content information makes our method more robust to irrelevant noise. Furthermore, since SLU tasks aim to capture semantics from spoken content, unit prediction acts as a content-related regularization technique, thereby improving few-shot training.

The primary contributions of this paper are as follows: 1) Proposing self-supervised units as guidance for textless SLU. 2) Demonstrating consistent improvement on different SLU benchmark corpora. 3) Showing the proposed method's capacity for few-shot and noise robustness.

2. Method

2.1. Problem Formulation

Spoken Language Understanding (SLU) tasks intend to capture semantic information in an utterance. We consider three types of SLU tasks: 1) **Spoken Name Entity Recognition (SNER)**: Extract entity names and their corresponding labels. 2) **Joint Intent Classification and Slot Filling (ICSF)**: Similar to SNER, but also needs to recognize the utterance-wise intent label. 3) **Spoken Semantic Parsing (SSP)**: Creating a semantic parsing tree based on the utterance.

We formulate these different textless SLU tasks as sequence generation tasks. For the SNER task, the targeted ground truth output sequence is constructed in the BI format, which means adding "B-" as a prefix to the entity class name for the first word, and "I-" for the rest of words in the entity. For the ICSF task, sequences are designed in a similar BI format as SNER, but with added intent class labels at the beginning and end of the sequence. Finally, for the SSP task, the sparse tree is transformed into a sequence format where different semantic labels are separated by brackets, along with the corresponding transcripts. This formulation helps maintain information on tree structures. However, due to the textless premise, we slightly modify the sequence from carrying the transcripts, which means that we only focus on predicting the semantic labels and the tree structure. Examples of each formulation are shown in Figure 2.

2.2. Self-Supervised Speech Units

Recent studies in Textless NLP [12, 13, 14, 15, 16] have utilized self-supervised clustered discrete units to model speech and

Table 1: *Evaluation metrics for different datasets.*

Dataset	Evaluation Metric
ATIS	F1 / ST-F1 / INT-Acc
SLUE-SNER	F1 / ST-F1 / SV-CER
SLURP	SLU-F1 / INT-Acc
SNIPS	ST-F1 / SV-CER / INT-Acc
STOP	EM-Tree

have shown that discrete units mostly preserve spoken content. In this work, we propose to leverage these self-supervised units as intermediate guidance for content-related tasks like SLU. We follow a similar pipeline as Lin *et al.* [15] to generate the discrete units. Specifically, we employ HuBERT-Large [22], which is pre-trained on LibriLight [23] 60k-hour dataset, to produce 1024-dimension features. Each semantic frame representing 20 ms of speech. The resulting features are then passed through a K-means clustering model ($K = 500$), which is pre-trained on LibriSpeech [24] 100-hour dataset, to transform them into discrete tokens. Tokens are deduplicated to remove redundant information, resulting in shortened and simplified unit sequences (e.g., `17 17 5 8 8` to `17 5 8`) as the intermediate target for SLU.

2.3. Framework

The proposed framework is constructed based on the S3PRL toolkit[2] to connect the pre-trained self-supervised model to the downstream model. For the upstream model, we selected HuBERT-Base [22], and its last layer outputs are used as input for the downstream model. The downstream model is a transformer-based encoder-decoder network designed for the main SLU task. In addition to predicting SLU tasks, we incorporate unit predictions as auxiliary guidance to direct the intermediate representations toward content-related information. The auxiliary unit prediction task shares the same upstream model and encoder as the main task but has a separate transformer decoder for the unit prediction. The overview of the proposed framework is illustrated in Figure 1.

During the training phase, the overall loss function is constructed as a weighted sum of the main and auxiliary tasks, and the upstream model is fine-tuned accordingly. The loss function is expressed by the following equation:

$$\mathcal{L} = (1 - \lambda) \times \mathcal{L}_{slu} + \lambda \times \mathcal{L}_{aux} \quad (1)$$

where $\mathcal{L}_{slu}$ is a cross-entropy loss for sequence generation, and $\mathcal{L}_{aux}$ is also a cross-entropy loss but for unit sequence prediction.

3. Experimental Setup

3.1. Dataset

We evaluate the proposed method on five well-known English SLU datasets: ATIS [17], SLUE-SNER [18], SLURP [19], SNIPS [20], and STOP [21]. ATIS is a classical corpus that measures the progress of SLU systems with the speech and transcriptions in the aviation domain. SLUE-SNER is a benchmark task created for spoken name entity recognition. We follow the official training set and evaluate using the development set, as the test set is not publicly accessible. SLURP is another widely used SLU package that covers 18 domains and contains a large amount of data. For SNIPS, we use the Slot Filling (SF)

---

[2]https://github.com/s3prl/s3prl

Table 2: *The performance of baseline, unit-guiding, and text-guiding approaches measured on test sets of ATIS, SLURP, SNIPS, STOP, and the development set of SLUE-SNER. The "Previous" results are sourced from [3] for ATIS, [18] for SLUE-SNER, [5] for SLURP, [10] for SNIPS, and [21] for STOP. The notation "*" with gray color indicates that the training uses ASR transcripts, which are not directly comparable to our setup. "N/A" means that we cannot find performance reported by prior works.*

Dataset	ATIS			SLUE-SNER			SLURP		SNIPS			STOP
Metric	F1↑	ST-F1↑	INT-Acc↑	F1↑	ST-F1↑	SV-CER↓	SLU-F1↑	INT-Acc↑	ST-F1↑	SV-CER↓	INT-Acc↑	EM-Tree↑
Previous	76.6	N/A	93.2	70.3*	N/A	N/A	71.9*	77.0	89.8*	21.8*	N/A	82.9*
Baseline	79.1	84.3	96.5	64.8	74.1	35.2	63.2	78.7	77.6	42.9	96.7	80.0
Unit	**82.4**	**86.0**	**96.8**	**68.6**	**78.1**	**29.4**	**67.9**	**80.9**	**82.7**	**31.9**	**97.0**	**84.4**
Text	84.5	87.7	97.4	69.2	78.2	29.0	69.9	82.5	83.2	30.7	97.0	84.5

Table 3: *The performance of few-shot training with 10% of the training set on SLURP and SNIPS datasets. Values denoted in the δ column indicate the performance drop from 100% training data to 10% training data.*

Dataset	SLURP						SNIPS					
Metric	SLU-F1↑			INT-Acc↑			ST-F1↑			SV-CER↓		
Portion	100%	10%	δ	100%	10%	δ	100%	10%	δ	100%	10%	δ
Baseline	63.2	45.2	18.0	78.7	54.7	24.0	77.6	64.2	13.4	42.9	62.2	19.3
Unit	**67.9**	**53.7**	**14.2**	**80.9**	**69.5**	**11.4**	**82.7**	**78.0**	**4.7**	**31.9**	**40.3**	**8.4**

task from the SUPERB [10] benchmark, which adopts Audio SNIPS [20] and follows the standard split in SNIPS [25]. Lastly, STOP is a recent SLU semantic parsing dataset that includes a huge amount of audio files and speakers. Overall, we evaluate the SNER task using SLUE-SNER; the ICSF task using ATIS, SLURP, and SNIPS, and the SSP task using STOP.

We measure the performance of the predicted outputs in the main task with several evaluation metrics. **F1** requires both the slot label and value to be correctly predicted, **Slot-Type-F1** (ST-F1) measures the F1-score for slot-types, and **Slot-Value-CER** (SV-CER) calculates the Character Error Rate (CER) for slot-values. **SLU-F1** is specific to the SLURP dataset and combines F1 evaluation with text-based distance measurements, including CER on the character level and the Word Error Rate (WER). **INT-Acc** measures the accuracy of predicting intent classes in the ICSF task, and **EM-Tree** represents the accuracy of producing the semantic parse labels in the SSP task. The corresponding metrics for each dataset are listed in Table 1.

3.2. Implementation Details

The experiments are implemented under a 3-layer encoder followed by a 6-layer SLU decoder with a size of 512 for ATIS and 1024 for the rest four datasets. The attention dimension is 512 and there are 4 heads. The unit decoder contains 2 layers and the unit size is 500. For each dataset, there are three ways to train the model. **Baseline**: only do the main task (i.e., $\lambda = 0$). **Unit**: use discrete units as the target of the auxiliary task with $\lambda = 0.5$. **Text**: use transcripts as the target instead of units. For every experiment, we search the learning rate from the range $[2e-5, 6e-5, 2e-4, 6e-4, 2e-3]$, batch size from [96, 192, 384], and report the best results. The total training steps are 10k. Furthermore, we augment the waveforms of every data with a speed perturbation by a factor of 0.9 and 1.1.

4. Results

Table 2 shows the results for ATIS, SLUE-SNER, SLURP, SNIPS, and STOP datasets to compare the baseline approach with our proposed method "Unit". In Table2, the row labeled "Previous" is prior end-to-end SLU works that do not utilize

ASR transcripts[3].

Regarding the **SNER** problem, the results on the SLUE-SNER dataset demonstrate a 3.8% improvement in F1-score, 4.0% in ST-F1, and 5.8% in SV-CER when unit guidance is utilized. Additionally, the proposed method achieves a comparable F1-score to [18], which is a topline approach leveraging paired transcripts and an additional language model for this task.

For the **ICSF** tasks, we observed superior performance on the ATIS and SLURP datasets when comparing the "Baseline" with the "Unit" method. For instance, in the case of ATIS, the F1-score improved by 3.3% and the ST-F1 increased by 1.7% when unit guidance was added. Furthermore, our method outperforms previous work by 5.8% and 3.6% in F1-score and intent accuracy, respectively. Similar results were observed in the SLURP dataset. On the other hand, in the SNIPS dataset, although the previous work has improved the ST-F1 by 7.1% and the SV-CER by 10.1% compared to our "Unit" results, it used additional ASR transcriptions. More importantly, the improvement from "Baseline" to "Unit" is about 5.1% in ST-F1 and 11.0% in SV-CER, which is a significant increase compared to other datasets.

In the case of the STOP dataset used in the **SSP** task, we discovered a 4.4% improvement in the EM-Tree metric when using discrete units, and this improvement is 1.5% higher than the previous work. It should be noted that while the performance of the previous work is higher than the baseline in our method, we do not use any transcription in the STOP dataset to generate semantic labels. This is a premise that makes the training stage of the SSP task harder than the previous work, and therefore, a direct comparison is not feasible.

To verify the effectiveness of units compared to text, we included the results trained with text guidance (i.e., utilizing the transcripts as targets) in the fourth row labeled "Text". It is worth noting that most of the gaps between "Text" and "Unit" are within the range of 0.1% to 1.7%, which is a relatively small performance loss compared to the baseline. These results indicate that our method yields competitive performance to models that leverage paired transcripts.

---

[3]Not every dataset has the results of end-to-end SLU without ASR transcripts, so if some prior works leverage transcripts, we use the notation "*" to mark the difference.

Table 4: *Results of the performance drop on the test set of SLURP dataset under different noisy configurations. G is Gaussian noise with the followed amplitude value. M is MUSAN background noise, with the dB value representing signal-to-noise ratio. Reverb represents the reverberation effect.*

Metric	SLU-F1↑						INT-Acc↑					
Noise	w/o	G-0.01	G-0.02	M-20dB	M-10dB	Reverb	w/o	G-0.01	G-0.02	M-20dB	M-10dB	Reverb
Baseline	63.2	-3.9	-7.7	-1.7	-5.4	-3.9	78.7	-4.8	-9.8	-2.2	-6.4	-4.2
Unit	**67.9**	**-3.2**	**-6.8**	**-1.4**	**-4.8**	**-2.6**	**80.9**	**-3.7**	**-5.3**	**-1.5**	**-5.2**	**-2.7**

Table 5: *Results of the performance drop on the test set of SNIPS dataset under different noisy configurations. The notations for noises (G, M, Reverb) are the same as Table 4.*

Metric	ST-F1↑						SV-CER↓					
Noise	w/o	G-0.005	G-0.01	M-20dB	M-10dB	Reverb	w/o	G-0.005	G-0.01	M-20dB	M-10dB	Reverb
Baseline	77.6	-6.9	-15.4	-3.1	-13.4	-2.3	42.9	+9.1	+19.0	+4.1	+16.2	-3.3
Unit	**82.7**	**-2.9**	**- 9.4**	**-1.9**	**-11.2**	**-2.1**	**31.9**	**+4.9**	**+14.4**	**+2.8**	**+15.5**	**-3.1**

5. Discussion

In the following section, we further explore the advantages of our method, including the few-shot capability, noise robustness, and applicability to other SSL models.

5.1. Few-Shot Capability

Table 3 displays the training performances of SLURP and SNIPS under few-shot scenarios. Concretely, we randomly sample 10% of the data from the training and validation sets, while the original test set is utilized for evaluation. All experimental settings are the same as the training process with the full data, except for the number of training epochs, which is reduced to 5k steps during few-shot training. Results demonstrate that utilizing discrete units as guidance leads to a smaller performance drop when the size of the training dataset decreases. Moreover, the performance achieved with units on 10% SNIPS data is even better than that without units on 100% SNIPS data, highlighting the few-shot ability of our approach.

We hypothesize that the use of unit guidance for the intermediate representation enables the encoder to capture mostly content information. The SLU decoder benefits from the representative input feature, allowing it to achieve better performance with a limited amount of data.

5.2. Noise Robustness

We discover that unit prediction not only guides the encoder to extract content-related information from speech, but also serves as the regularization to help the model focus on useful content instead of irrelevant information (such as additive noises and environmental conditions). In order to discover how unit guidance can help when the target domains contain irrelevant information, we artificially create mismatched test sets with various types of noises. Specifically, noises include additive Gaussian noise (G), background noise MUSAN [26] (M), and reverberation effect (Reverb).

In Table 4 and 5, we utilize models trained with the setting of "Baseline" and "Unit" for the SNIPS and SLURP datasets mentioned in Table 2. For Gaussian noise, we set the amplitude to 0.005, 0.01, and 0.02. For MUSAN noise, we set the signal-to-noise ratio (SNR) to 20 dB and 10 dB.

As shown in Table 4 and 5, under the same level of domain mismatch, models trained with discrete units on the SLURP and SNIPS have better capacities of noisy robustness and smaller

Table 6: *The performance of unit-guiding and text-guiding approaches using wav2vec 2.0 Base as the upstream SSL model. The dashed line means that the training fails to converge.*

Dataset	SLURP		SNIPS	
Metric	SLU-F1↑	INT-Acc↑	ST-F1↑	SV-CER↓
Baseline	–	–	72.2	53.0
Unit	**65.4**	**79.5**	**78.6**	**38.1**
Text	66.7	80.9	80.7	35.0

performance degradation in all three mismatched conditions. For instance, when the amplitude of Gaussian noise is set to 0.01, the SNIPS dataset shows a 15.4% decrease in "Baseline" ST-F1 and only a 9.4% decrease in "Unit" ST-F1. This reveals a 6.0% difference in performance drop between the two cases, showing better robustness during inference time.

5.3. Different SSL Models

In order to explore the applicability of our method to other pre-trained SSL models, we conducted experiments using the wav2vec 2.0 [27] model while keeping other components consistent. Table 6 presents the results obtained using the wav2vec 2.0 Base model. Our findings from these models, trained on SLURP and SNIPS datasets, show the same trend as observed in Table 2. They suggest that discrete units can serve as effective guidance for different self-supervised models, leading to performance improvements and better training convergence.

6. Conclusion

In this work, we propose a textless SLU framework that can achieve comparable performance without paired transcripts. We utilize self-supervised discrete units as the intermediate target to guide the model in learning content-related representations. The results show significant improvements accross five SLU benchmarks when compared to the baseline method. Additionally, we demonstrate that our approach provides the few-shot capability (10% of data) and alleviates the performance degradation in mismatched and noisy environments. Future work includes input augmentation with clean unit guidance to enhance noise robustness and exploring the intermediate target using unsupervised ASR [28, 29].

7. References

[1] D. Serdyuk, Y. Wang, C. Fuegen, A. Kumar, B. Liu, and Y. Bengio, "Towards end-to-end spoken language understanding," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5754–5758.

[2] M. Saxon, S. Choudhary, J. P. McKenna, and A. Mouchtaris, "End-to-End Spoken Language Understanding for Generalized Voice Assistants," in *Proc. Interspeech 2021*, 2021, pp. 4738–4742.

[3] E. Morais, H.-K. J. Kuo, S. Thomas, Z. Tüske, and B. Kingsbury, "End-to-end spoken language understanding using transformer networks and self-supervised pre-trained features," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7483–7487.

[4] S. Thomas, H.-K. J. Kuo, G. Saon, Z. Tüske, B. Kingsbury, G. Kurata, Z. Kons, and R. Hoory, "Rnn transducer models for spoken language understanding," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7493–7497.

[5] S. Arora, S. Dalmia, P. Denisov, X. Chang, Y. Ueda, Y. Peng, Y. Zhang, S. Kumar, K. Ganesan, B. Yan *et al.*, "Espnet-slu: Advancing spoken language understanding through espnet," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7167–7171.

[6] P. Wang, X. Ye, X. Zhou, J. Xie, and H. Wang, "Speech2slot: An end-to-end knowledge-based slot filling from speech," *arXiv preprint arXiv:2105.04719*, 2021.

[7] S. Thomas, H.-K. J. Kuo, B. Kingsbury, and G. Saon, "Towards reducing the need for speech training data to build spoken language understanding systems," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7932–7936.

[8] H.-K. J. Kuo, Z. Tüske, S. Thomas, Y. Huang, K. Audhkhasi, B. Kingsbury, G. Kurata, Z. Kons, R. Hoory, and L. Lastras, "End-to-end spoken language understanding without full transcripts," *Proc. Interspeech 2020*, pp. 906–910, 2020.

[9] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath, and S. Watanabe, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[10] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[11] G.-T. Lin, C.-L. Feng, W.-P. Huang, Y. Tseng, T.-H. Lin, C.-A. Li, H.-y. Lee, and N. G. Ward, "On the utility of self-supervised models for prosody-related tasks," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 1104–1111.

[12] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed *et al.*, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[13] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," *arXiv preprint arXiv:2104.00355*, 2021.

[14] E. Kharitonov, A. Lee, A. Polyak, Y. Adi, J. Copet, K. Lakhotia, T.-A. Nguyen, M. Rivière, A. Mohamed, E. Dupoux *et al.*, "Text-free prosody-aware generative spoken language modeling," *arXiv preprint arXiv:2109.03264*, 2021.

[15] 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 *Proc. Interspeech 2022*, 2022, pp. 5165–5169.

[16] E. Kharitonov, J. Copet, K. Lakhotia, T. A. Nguyen, P. Tomasello, A. Lee, A. Elkahky, W.-N. Hsu, A. Mohamed, E. Dupoux, and Y. Adi, "textless-lib: a library for textless spoken language processing," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*. Hybrid: Seattle, Washington + Online: Association for Computational Linguistics, Jul. 2022, pp. 1–9. [Online]. Available: https://aclanthology.org/2022.naacl-demo.1

[17] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, "The atis spoken language systems pilot corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.

[18] S. Shon, A. Pasad, F. Wu, P. Brusco, Y. Artzi, K. Livescu, and K. J. Han, "Slue: New benchmark tasks for spoken language understanding evaluation on natural speech," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7927–7931.

[19] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, "Slurp: A spoken language understanding resource package," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7252–7262.

[20] C.-I. Lai, Y.-S. Chuang, H.-Y. Lee, S.-W. Li, and J. Glass, "Semi-supervised spoken language understanding via self-supervised speech and language model pretraining," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7468–7472.

[21] P. Tomasello, A. Shrivastava, D. Lazar, P.-C. Hsu, D. Le, A. Sagar, A. Elkahky, J. Copet, W.-N. Hsu, Y. Mordechay *et al.*, "Stop: A dataset for spoken task oriented semantic parsing," *arXiv preprint arXiv:2207.10643*, 2022.

[22] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[23] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7669–7673.

[24] 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.

[25] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril *et al.*, "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces," *arXiv preprint arXiv:1805.10190*, 2018.

[26] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[27] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[28] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27 826–27 839, 2021.

[29] G.-T. Lin, C.-J. Hsu, D.-R. Liu, H.-Y. Lee, and Y. Tsao, "Analyzing the robustness of unsupervised speech recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8202–8206.