1SPU: 1-step Speech Processing Unit

Karan Singla, Shahab Jalavand, Andrej Ljolje, Antonio Moreno Daniel,
Srinivas Bangalore, **Yeon-Jun Kim**, **Ben Stern**
karan@whissle.ai, sjalalvand@interactions.com

Abstract

Recent studies have made some progress in refining end-to-end (E2E) speech recognition encoders by applying Connectionist Temporal Classification (CTC) loss to enhance named entity recognition within transcriptions. However, these methods have been constrained by their exclusive use of the ASCII character set, allowing only a limited array of semantic labels. We propose 1SPU, a 1-step Speech Processing Unit which can recognize speech events (e.g: speaker change) or an NL event (Intent, Emotion) while also transcribing vocal content. It extends the E2E automatic speech recognition (ASR) system's vocabulary by adding a set of unused placeholder symbols, conceptually akin to the <pad> tokens used in sequence modeling. These placeholders are then assigned to represent semantic events (in form of tags) and are integrated into the transcription process as distinct tokens.

We demonstrate notable improvements on the SLUE benchmark and yields results that are on par with those for the SLURP dataset. Additionally, we provide a visual analysis of the system's proficiency in accurately pinpointing meaningful tokens over time, illustrating the enhancement in transcription quality through the utilization of supplementary semantic tags.

1 Introduction

This paper extends the capabilities of End-to-End Automatic Speech Recognition (E2E ASR) systems from traditional transcription tasks to more nuanced speech understanding processes (Chan et al., 2015; Bahdanau et al., 2016). Presently, Deep Neural Networks (DNNs) are proficient in applying greedy or beam search decoding to predict vocabulary labels from audio inputs over time (Watanabe et al., 2017; He et al., 2019). These networks typically produce a sub-word text token for each segment of a continuous speech stream, focusing on transcription fidelity.

However, the transcription process has evolved with recent efforts to embed additional metadata, such as named entity boundaries and utterance classification (Serdyuk et al., 2018; Haghani et al., 2018). Multitask learning frameworks utilizing separate loss functions for transcription and classification have emerged (Wang et al., 2023; Ghannay et al., 2018), and there has been success in identifying sequences of intents using non-autoregressive models within a speech stream (Potdar et al., 2021). Although recognizing a limited set of named entities using CTC loss by marking them with specific begin and end tokens in transcriptions has been demonstrated (Ghannay et al., 2018). Methods like these are often use special ASCII characters for events, constrained by the size ASCII character set. Morever these special characters are originally defined to be useful for meaningful language transcription. This also happens a as ASR encoders are generally considered as intermediaries between speech and natural language (NL) models.

In response, our study proposes a single-step, integrated speech processing unit (1SPU) using E2E speech encoders. 1SPU not only transcribes spoken content but also annotates it with tags indicating various NLP and speech events, including intent and entity recognition as well as speaker changes. By incorporating $user\_defined\_symbols$ into the ASR vocabulary (Kudo and Richardson, 2018), akin to text-based entity tagging tokens, we enhance the encoder's lexicon. These symbols are introduced during the fine-tuning stage and are only used if present in the fine-tuning data, thereby maintaining computational efficiency. We fine-tune a pre-trained Conformer ASR model to perform transcription while also producing intent labels and entity tags, all under the CTC loss function. The model outputs transcriptions using standard greedy decoding and can function in both offline and streaming modes.

Our experimentation focuses on mono-channel

speech and yields promising results on the SLUE (Wang et al., 2021) and SLURP (Bastianelli et al., 2020) benchmarks. We posit that incorporating a Language Model (LM) could significantly elevate these results. Furthermore, we present visual evidence that utilizing context from previous utterances can enhance transcription and understanding in dual-channel human-human interactions, particularly in marking speaker changes. This ongoing work has the potential to improve prompting strategies and facilitate the development of multi-modal and multitasking E2E ASR applications in streaming contexts.

The contributions of this paper are summarized as follows:

- We propose 1SPU, which incorporates the central idea of $user\_defined\_symbols$ into the ASR vocabulary to facilitate the tagging of speech and NL events within continuous speech transcriptions during fine-tuning.

- We demonstrate that our method not only improves performance on the SLUE benchmark but also shows promise for the SLURP dataset.

- We offer visual proof that the inclusion of context from prior events can improve speech transcription and understanding in scenarios requiring low-latency responses.

2   Related Work

A common practice is to transform the normalized token sequences produced by automatic speech recognition (ASR) systems into a *written form* that is more amenable for downstream dialog system components to process (Pusateri et al., 2017). This reformatted written output then serves as the basis for extracting *structured information*, such as intents and slot values, which are crucial for the continuation of a dialog (Radfar et al., 2020). Of late, there has been an emerging trend towards utilizing neural encoders that are optimized directly on speech inputs, a methodology often referred to as *end-to-end spoken language understanding (E2E SLU)* (Serdyuk et al., 2018; Haghani et al., 2018).

Text-based Sequence Tagging: The domain of natural language processing has seen extensive research in employing sequence-to-sequence (seq-2-seq) models to parse and extract structured information from textual data. This encompasses

advancements in identifying the positions and classifications of entities within a text (Ratinov and Roth, 2009), generating structured summaries of content (Gorinski et al., 2019), and constructing synctactic parse trees through timely tag predictions during the generation of language (Cabot and Navigli, 2021). Despite their efficacy, these approaches, which are primarily designed for monologic text, do not adequately address the dynamic nature of spoken dialogue.

E2E ASR: Modern Automatic Speech Recognition (ASR) systems excel at converting spoken language into a sequence of lexical tokens, effectively transcribing spoken words. This transcription process typically involves refining pre-trained self-supervised neural networks or initiating training from the ground up with non-autoregressive Connectionist Temporal Classification (CTC) loss or a sequential loss (Graves, 2012) to recognize speech tokens (Chan et al., 2015; Chorowski et al., 2015). Transcriptions are obtained either by greedy decoding or beam-search decoding using a language model.

E2E SLU: With the emergence of end-to-end ASR (Chorowski et al., 2015; Chan et al., 2016) and the successful pretraining of speech encoders, methods for SLU directly from the speech signal have recently shown comparable performance to the conventional approach of cascading ASR and text-based components in tasks such as named entity recognition (NER), translation, dialogue act prediction (DAP) (Vila et al., 2018; Dang et al., 2020), as well as inference tasks like emotion, intent or behavior understanding (Fayek et al., 2015; Price et al., 2020; Singla et al., 2020). Recent studies have attempted to identify the beginning and end of named entities in transcriptions by adding unique markers to the text (Ghannay et al., 2018; Tomashenko et al., 2019). However, the performance of these methods is often limited by the small selection of special characters available in the ASCII character set and fear of using special symbols which may conflict with making multilingual ASR systems or other desired symbols.

3   1SPU: 1-step Speech Processing Unit

For 1-step Speech Processing Unit (1SPU) we leverage a commercially available End-to-End Automatic Speech Recognition (E2E ASR) system, initializing it with a tokenizer that includes D user-defined dummy tokens. These labels are akin to
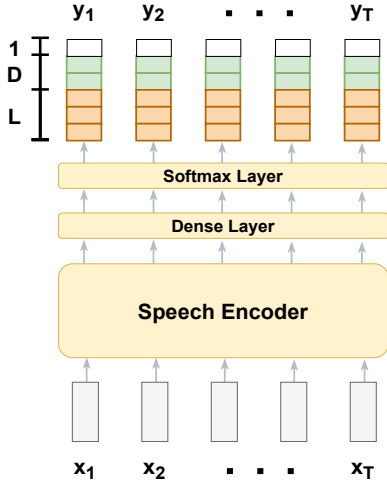
Figure 1: Overview of 1SPU. L refers to transcription tokens, D refers to un-used tokens learnt during fine-tuning to mark events. 1 is the *blank* token.

unique tokens, much like a '<pad>' token, and are not utilized during the ASR's initial training phase. Thus, the incorporation of these inactive tokens does not impact the duration of training or inference as they can be masked during the E2E ASR optimization process.

Subsequently, we re-purpose the pre-trained speech encoder to generate transcriptions that embed semantic tokens indicative of event tags, optimizing with the CTC loss function. This strategy diverges from previous methods that relied on special characters for tagging within transcriptions using CTC loss. Our approach, which introduces dummy tags during the ASR's pre-training, obviates the need for alterations to the tokenizer or output layer before fine-tuning.

We detail a fine-tuning process for a pre-trained speech encoder that empowers it to recognize utterance intent, demarcate the beginning and end of entities, and carry out transcription, using pairs of speech files and tagged transcriptions. An example of a transcribed speech input is:

```
1  CALENDER_SET put EVENT_NAME
     meeting END with PERSON paul
     END for DATE tomorrow END TIME
     ten am END
```

Here, *CALENDAR_SET* signifies the intent of the utterance, while the entity tokens *PERSON*, *DATE*, and *TIME* terminate with a shared *END* token. These semantic tokens are mapped to the dummy vocabulary tokens, which are subsequently learned during the fine-tuning stage for the desig-

nated task. Although the tokens for utterance level semantics can be placed at the beginning or the end, our empirical evidence suggests a slight performance improvement when it is generated at the beginning.

In our study, we apply the Conformer end-to-end ASR framework(Gulati et al., 2020). We use tokenizer output vocabulary V of size 1024, comprised of 400 un-used tokens D for event labels, 624 transcription tokens L and 1 token representating a *Blank* label. These D tokens are provided to tokenizer as a list of *user_defined_symbols*

3.1 From network output to transcriptions with event labels

Adopting the computational approach of Connectionist Temporal Classification (CTC), our model is designed to classify sequences of unheard speech with the objective of minimizing the task-specific error metric. This involves correctly predicting output tokens that either correspond to transcription elements or signify an event. Consistent with established CTC methodologies, our network includes a softmax layer that features one additional unit beyond the number of tokens in V. This extra unit's activation represents the likelihood of a *blank* or no label. The activations of the first L units are interpreted as the probabilities of each transcription token occurring at specific time steps. The remaining D units are purposed to signal events such as intents, entities or speaker change when activated.

More Formally, for an input sequence x of length T(steps in a sample) define a speech DNN encoder with m inputs, n outputs and weight vector w as a continous map $N_w : (R^m)^T \mapsto (R^n)^T$. Let $y = N_w(x)$ be the sequence of network output, and denote by y_k^t the activation of output unit k at time t. y_k^t is interpreted as the probability of observing label k at time t, thus, defining a distribution over the set V'^T of length T sequences over the alphabet $L' = L \cup \{blank\}$:

$$p(\pi|x) = \prod_{t=1}^{T} y_{\pi_t}^t \qquad (1)$$

We refer to the elements of the V'^T as paths, and denote them as π.

Graves et al. (Graves and Graves, 2012) presuppose in their first equation that outputs of a network at various time steps are independent of each other. Yet, this assumption does not hold in the architecture of modern End-to-End (E2E) speech encoders

such as Citrinet, Conformer, and Wav2vec2. These models employ feedback loops that inherently link different temporal outputs, creating a conditional dependency between them. This characteristic interconnection may contribute significantly to the widespread adoption and effectiveness of E2E ASR systems based on CTC methodology.

Many$-$to$-$one map B is defined as $V' \mapsto V^{\leq T}$, where $V^{\leq T}$ refers to set of sequences of length less than or equal to T over the original label vocab V. To obtain read-able transcriptions we apply standard CTC rule of removing all blanks and repeated labels. Finally, B is used to define the conditional probability of an entity $l \in V^{\leq T}$ as sum of probabilities of all the paths corresponding to it:

$$\sum_{\pi \in B^{-1}(e)} p(\pi|x) \qquad (2)$$

The foundational mathematical framework of CTC loss, when combined with the contextualized representations generated by a speech encoder, enables us to refine outputs using previously unused dummy tokens to denote specific event labels in an input stream.

3.2 Pre-trained checkpoint and fine-tuning details

We utilize readily available Conformer models(Gulati et al., 2020) accessible from the NeMo library[1], which have been pre-trained on a corpus comprising 7,000 hours of transcribed public domain audio. These models employ a SentencePiece tokenizer(Kudo and Richardson, 2018) configured with a vocabulary size of 1024 tokens, denoted as L. For the integration of dummy event labels, we conduct an additional phase of fine-tuning focused on transcription tasks. This fine-tuning leverages a dataset amalgamating Fisher (Cieri et al., 2004), GigaSpeech (Chen et al., 2021) and SwitchBoard (Godfrey et al., 1992) corpora, which were also part of the initial training set utilized by the pre-trained model checkpoints. In our experiments L event tokens are learnt directly on the event annotated fine-tuning data. We use a batch-size of 16 with an initial learning rate of .001. We use a weight decay of .001 and update the model with 8 accumulated batches with adam back-propagation algorithm. We use 3 A100 GPUs for pre-training and 1 A100 GPU for fine-tuning purposes.The final revision will include more details about the experiment configuration and models.

4 Dataset

We evaluate our method on the standardized *SLUE-voxpopuli* (Wang et al., 2021) and *SLURP* corpus (Bastianelli et al., 2020).

4.1 SLUE

The SLUE-voxpopuli corpus represents a substantial multilingual collection of spontaneous speech recordings from the European Parliament proceedings. Named Entity (NE) annotations for a 15-hour segment of its training subset and the entire standard development set were recently made available by Shon et al.(Shon et al., 2022). While the annotations for the test set are not publicly accessible, we are able to evaluate our models on this set by adhering to the submission guidelines provided by the SLUE project[2]. Details pertaining to the distribution of these data subsets are outlined in Table 1. For the scope of this study, we focus on using combined entity labels, encompassing eight distinct types, rather than employing the more detailed intent labels.

Label		Count of phrases	
Raw	Combined	Fine-tune	Dev
GPE, LOC,	PLACE	2012	642
CARDINAL, ORDINAL QUANTITY, MONEY, PERCENT	QUANT	923	327
ORG	ORG	864	259
DATE, TIME	WHEN	762	260
NORP	NORP	647	220
PERSON	PERSON	272	51
LAW	LAW	250	60
FAC, EVENT, WORK_OF_ART, PRODUCT, LANGUAGE	DISCARD		
TOTAL Entities		5820	1862
Total Duration		5k (15hrs)	1.7k (5hr)

Table 1: Data Statistics of SLUE-voxpopuli corpus with Named Entity annotations

The evaluation methodology adopted by the SLUE framework (Shon et al., 2022) entails the transformation of entity occurrences within an utterance into a structured format that includes the entity's *Type*, its *Phrase*, and *Frequency* for both the reference text and the system's output. The

---

[1] https://tinyurl.com/bdf38tmz

selection of the phrase is determined by identifying the beginning and end markers of an entity. In this context, each entity type is signified by a unique beginning token, whereas the end token is standardized across all entity types for the purpose of evaluation. The scores are then computed as follows:

- $Recall = Total - correct/Total - reference$

- $Precision = Total - correct/Total - system$

- $F1score = 2 * (Precision/Recall)/(Precision + Recall)$

This evaluation metric does not report the transcription accuracy. Therefore, we report standard WER (after removing semantic event labels) for our experiments.

4.2 SLURP Corpus

The Spoken Language Understanding Resource Package (SLURP) dataset comprises a rich collection of 72,000 audio clips, summing up to 58 hours of speech. These recordings simulate single-turn interactions with a virtual home assistant and are annotated across three semantic dimensions: Scenario, Action, and Entities. The dataset is diverse, featuring 56 types of entities, 46 distinct actions, and 18 different scenarios. For automation purposes, action and scenario labels are typically amalgamated to denote the user's intent, resulting in 93 unique intent labels.

The fine-tuning and assessment of the speech encoder leverage the designated official splits for training, validation, and testing from SLURP. We present our findings using the SLURP-specific evaluation metrics, accessible via their official GitHub repository. Unlike the standard SLUE evaluation framework, our entity recognition accuracy is gauged using the provided human-annotated entity boundaries within the audio files. We evaluate system performance by comparing the tuple (Type, Phrase, Frequency) against the reference for each time segment. The calculation of recall, precision, and F-1 score follows the methodology outlined in the preceding section, focusing on the correct identification of entity tags and their corresponding phrases. However, it is important to note that these metrics do not consider the temporal accuracy of

the entity tags' start and end points, which is crucial for real-time applications like speech redaction as discussed in Gouvea et al. (Gouvêa et al., 2023).

5 Experiments & Results

5.1 SLUE-Voxpopuli

In the two-step methodology, a text tagger, when applied to human-transcribed speech, achieves the highest accuracy for tagging entities. Past research by Pasad et al. (Pasad et al., 2021) indicates that when the Wav2Vec2 (W2V2) model is fine-tuned for both transcription and tagging tasks, there is a noticeable decline in performance. Although Pasad et al. provide results using the W2V2 model, we conducted our experiments using Conformer models. We discovered that employing unused vocabulary labels during the fine-tuning process mitigates the performance degradation. Our proposed method not only improves transcription results but also enhances the accuracy of entity tagging. Our findings also reveal that assigning separate tags for the beginning of an entity is less effective compared to other methods. It is observed that encoders fine-tuned with Connectionist Temporal Classification (CTC) outperform those using the sequence-to-sequence RNNT optimization loss (Ghodsi et al., 2020). The baseline results for these observations are documented in the work by Shon et al. (Shon et al., 2022).

Speech Model	Text Model	WER	NER-F1 Comb
Text NER			
N/A	DeBERTa-B	N/A	86.0
2-step Pipeline			
W2V2-B-LS960 (165m)	DeBERTa-B	18.4	78.6
Conformer-CTC (115m)	TPT	9.8	80.3
Conformer-RNNT (115m)	TPT	9.2	82.1
1-step E2E			
Baseline			
W2V2-B-LS960	N/A	18.4	49.0
Conformer − CTC		11.0	65.2
Conformer − RNNT		12.1	67.2
Ours - 1SPU			
Conformer − CTC	N/A	**9.0**	**74.4**
Conformer − RNNT		10.5	71.1

Table 2: Results for SLUE benchmark.

5.2 SLURP

Table 3 presents the outcomes for our one-step model and benchmarks it against other E2E approaches. Stutructed prediction Model from NVIDIA[3] (no citation available) , employing an

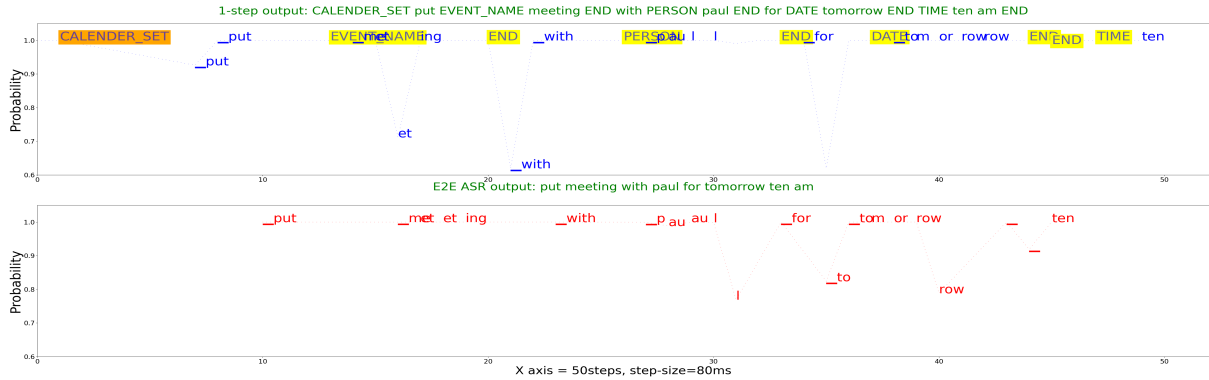---

[3]https://tinyurl.com/nemo-rnnt-structured

Figure 2: Output of softmax probability for token predicted at every time-step. Steps with no visualized token represent "empty" token being predicted. Blue tokens mark our system's output which red tokens represent tokens from pretrained E2E ASR. Our system learns to insert tags by generating less empty tokens and use them to mark event labels instead.

Speech Model	WER	Intent-Acc	Slurp-F1
Previous			
E2E Encoder-Decoder			
W2V2-B-LS960	22.3	83.0	65.2
Conformer-RNNT	14.2	90.1	77.22
E2E Encoder only			
W2V2-B-LS960	19.2	82.0	59.2
Conformer-CTC	15.6	83.1	66
Proposed 1SPU			
Conformer-CTC	14.5	85.5	69.3

Table 3: Results for SLURP corpus.

encoder-decoder framework that produces information following a predetermined template, achieves the best performance. In comparison, our encoder-only model, which leverages CTC loss combined with greedy decoding, outperforms prior models with a similar encoder-only structure. It is important to note, however, that despite the superior performance of the encoder-decoder approach, it yields a structured output, which is not conducive to streamed inference and thus restricts its application to offline processing.

6 Observations

Figure 2 illustrates the token output from both a two-step pipeline system and a single-step entity recognition system at each time-step. Tokens from the end-to-end automatic speech recognition (E2E ASR) are marked in red, whereas tokens pertinent to entities identified by the single-step system are highlighted in blue.

In addition to these experiments, we also experiment with propriety human-human dual channel conversations between a caller and an agent. Assuming negligible overlap in the audio from both parties, we combine the two channels to create a single-channel audio track. To denote transitions between speakers, we introduce a unique label for "Speaker-change." A video demonstration, accessible via the provided link[4], showcases a sample conversation displayed in a format akin to Figure 2. From this, we notice an improvement in both the accuracy of entity prediction and overall transcription, which we attribute to the model's capacity to predict semantic events alongside the speech content.

7 Conclusions

In our study, we demonstrate that simultaneous high-quality event tagging (encompassing intent, entities, and speaker changes) and transcription is feasible with our proposed method. This technique is adaptable across a broad spectrum of events, making it suitable for extensive applications in speech understanding and transcription tasks.

We are confident that this methodology can be expanded to accommodate additional languages and various domains due to its multi-tasking nature, facilitated by a single, unified loss function. This makes it an ideal candidate for comprehensive, one-step, automated conversational systems. Currently, we are developing a multi-lingual version of this system, which has the capability to perform language translation while concurrently tagging events. Additionally, we aim to refine end-to-end automatic speech recognition (E2E ASR) systems to respond to event recognition prompts based on user-specified natural language queries.

---

[4]https://www.youtube.com/watch?v=10yte4YUqW0

References

Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. 2016. End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4945–4949. IEEE.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. Slurp: A spoken language understanding resource package. *arXiv preprint arXiv:2011.13205*.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. Rebel: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964. IEEE.

William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. 2015. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. *Advances in neural information processing systems*, 28.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: A resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71.

Viet-Trung Dang, Tianyu Zhao, Sei Ueno, Hirofumi Inaguma, and Tatsuya Kawahara. 2020. End-to-end speech-to-dialog-act recognition. *arXiv preprint arXiv:2004.11419*.

Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. 2015. Towards real-time speech emotion recognition using deep neural networks. In *2015 9th international conference on signal processing and communication systems (ICSPCS)*, pages 1–5. IEEE.

Sahar Ghannay, Antoine Caubrière, Yannick Estève, Nathalie Camelin, Edwin Simonnet, Antoine Laurent, and Emmanuel Morin. 2018. End-to-end named entity and semantic concept extraction from speech. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 692–699. IEEE.

Mohammadreza Ghodsi, Xiaofeng Liu, James Apfel, Rodrigo Cabrera, and Eugene Weinstein. 2020. Rnn-transducer with stateless prediction network. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7049–7053. IEEE.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Philip John Gorinski, Honghan Wu, Claire Grover, Richard Tobin, Conn Talbot, Heather Whalley, Cathie Sudlow, William Whiteley, and Beatrice Alex. 2019. Named entity recognition for electronic health records: a comparison of rule-based and machine learning approaches. *arXiv preprint arXiv:1903.03985*.

Evandro Gouvêa, Ali Dadgar, Shahab Jalalvand, Rathi Chengalvarayan, Badrinath Jayakumar, Ryan Price, Nicholas Ruiz, Jennifer McGovern, Srinivas Bangalore, and Ben Stern. 2023. Trustera: A live conversation redaction system. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Alex Graves and Alex Graves. 2012. Connectionist temporal classification. *Supervised sequence labelling with recurrent neural networks*, pages 61–93.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Parisa Haghani, Arun Narayanan, Michiel Bacchiani, Galen Chuang, Neeraj Gaur, Pedro Moreno, Rohit Prabhavalkar, Zhongdi Qu, and Austin Waters. 2018. From audio to semantics: Approaches to end-to-end spoken language understanding. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 720–726. IEEE.

Yanzhang He, Tara N Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, et al. 2019. Streaming end-to-end speech recognition for mobile devices. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6381–6385. IEEE.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Ankita Pasad, Felix Wu, Suwon Shon, Karen Livescu, and Kyu J Han. 2021. On the use of external data for spoken named entity recognition. *arXiv preprint arXiv:2112.07648*.

Nihal Potdar, Anderson R Avila, Chao Xing, Dong Wang, Yiran Cao, and Xiao Chen. 2021. A streaming end-to-end framework for spoken language understanding. *arXiv preprint arXiv:2105.10042*.

Ryan Price, Mahnoosh Mehrabani, and Srinivas Bangalore. 2020. Improved end-to-end spoken utterance classification with a self-attention acoustic classifier. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8504–8508. IEEE.

Ernest Pusateri, Bharat Ram Ambati, Elizabeth Brooks, Ondrej Platek, Donald McAllaster, and Venki Nagesha. 2017. A mostly data-driven approach to inverse text normalization. In *INTERSPEECH*, pages 2784–2788. Stockholm.

Martin Radfar, Athanasios Mouchtaris, and Siegfried Kunzmann. 2020. End-to-end neural transformer based spoken language understanding. *arXiv preprint arXiv:2008.10984*.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning (CoNLL-2009)*, pages 147–155.

Dmitriy Serdyuk, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua Bengio. 2018. Towards end-to-end spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5754–5758. IEEE.

Suwon Shon, Ankita Pasad, Felix Wu, Pablo Brusco, Yoav Artzi, Karen Livescu, and Kyu J Han. 2022. 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)*, pages 7927–7931. IEEE.

Karan Singla, Zhuohao Chen, David Atkins, and Shrikanth Narayanan. 2020. Towards end-2-end learning for predicting behavior codes from spoken utterances in psychotherapy conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3797–3803.

Natalia Tomashenko, Antoine Caubrière, and Yannick Estève. 2019. Investigating adaptation and transfer learning for end-to-end spoken language understanding from speech. In *Interspeech 2019*, pages 824–828. ISCA.

Laura Cross Vila, Carlos Escolano, José AR Fonollosa, and Marta R Costa-Jussa. 2018. End-to-end speech translation with the transformer. In *IberSPEECH*, pages 60–63.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*.

Jixuan Wang, Martin Radfar, Kai Wei, and Clement Chung. 2023. End-to-end spoken language understanding using joint ctc loss and self-supervised, pretrained acoustic encoders. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. 2017. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.