Speech Recognition and Multi-Speaker Diarization of Long Conversations

Huanru Henry Mao, Shuyang Li, Julian McAuley, Garrison W. Cottrell

UC San Diego

{hhmao, shl008, jmcauley, gary}@eng.ucsd.edu

Abstract

Speech recognition (ASR) and speaker diarization (SD) models have traditionally been trained separately to produce rich conversation transcripts with speaker labels. Recent advances [1] have shown that joint ASR and SD models can learn to leverage audio-lexical inter-dependencies to improve word diarization performance. We introduce a new benchmark of hour-long podcasts collected from the weekly *This American Life* radio program to better compare these approaches when applied to extended multi-speaker conversations. We find that training separate ASR and SD models perform better when utterance boundaries are known but otherwise joint models can perform better. To handle long conversations with unknown utterance boundaries, we introduce a striding attention decoding algorithm and data augmentation techniques which, combined with model pre-training, improves ASR and SD.

Index Terms: speech recognition, speaker diarization, podcasts

1. Introduction

Automatic speech recognition (ASR) and speaker diarization (SD) of natural conversation are tasks of broad interest with applications including transcribing meetings, phone calls, and interviews, among others. Traditionally, ASR and SD systems each operate independently on acoustic information to generate transcript text and label speaker segments. These outputs are then reconciled into a final speaker-annotated transcript. A limitation of training independent ASR and SD systems is that the models are unable to leverage the inter-dependencies between these two predictive tasks. For example, lexical cues from transcripts can help improve speaker turn change prediction [2].

Recent work has shown promising results in learning sequence transduction models that jointly perform ASR and SD in a two-speaker clinical setting by simply adding a speaker change token to the model's vocabulary [1]. To further explore these types of end-to-end approaches, we expand the joint framework to encompass ASR and SD in an open-domain setting for extended multi-speaker conversations. We introduce a benchmark dataset for this setting, consisting of 663 podcast episodes and transcripts collected from the weekly *This American Life* (TAL) radio program.[1] TAL is unique in two ways: each episode is an hour-long conversation and contains an average of 18 unique speakers in three roles. We propose two tasks for joint ASR and diarization: TAL aligned and unaligned, to evaluate models under situations where utterance bounds are either provided or unknown respectively. To benchmark performance in each setting, we measure the transcription error via word error rate (WER) and introduce a new metric, multi-speaker word diarization error (MWDE), to evaluate word-level speaker alignment. MWDE generalizes the previously proposed two-speaker word diarization error rate [1] to multiple speakers.

[1]Data and code can be found on: https://github.com/calclavia/tal-asrd

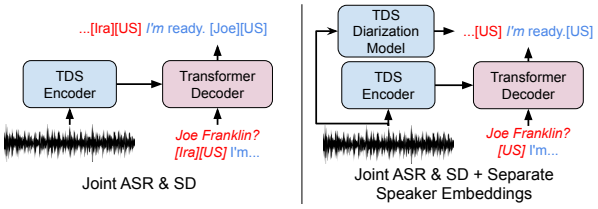

Figure 1: *The joint model concatenates utterances into a sequence delimited by a speaker token (e.g., [Ira]) and an utterance separator [US] token. Alternatively, separate speaker embeddings can be used instead of speaker tokens for diarization.*

We compare training separate ASR and SD models against the joint framework [1] for our multi-speaker setting and find that the separate framework is superior when known utterance boundaries are provided but worse otherwise, suggesting that joint models may be more appropriate as fully end-to-end rich transcription systems. To handle long conversations, we introduce a striding attention decoding algorithm to adapt a model trained on TAL aligned to the hour-long unaligned setting. We propose pre-training and data augmentation methods to complement the algorithm, achieving 16.1% and 15.8% absolute improvements to WER and MWDE on TAL unaligned.

2. Dataset

We collected podcast episodes from the *This American Life* radio program from 1995 to 2020, comprising 701 episodes which we cleaned and processed for a total of 663 episodes (38 had alignment issues due to inserted ads) and 637.70 hours of audio. Each episode corresponds to a "single conversation", and TAL comprises hundreds of lengthy dialogs. On average, each conversation is 58 minutes long, consisting of 247 dialog turns between 18 different speakers. We additionally collected professionally transcribed, publicly-available transcripts for each episode, which are aligned at the utterance-level.[2] Each utterance comprises an average of 45 words and 3 sentences, for a total of 7,390,793 words and 520,676 sentences across 163,808 utterances. 90% of utterances fall just under 30 seconds in duration and 100 words in length.

Conversations are loosely organized around a theme (e.g., "Middle School") with several guests who tell stories and engage with the host. There are 6,608 unique speakers identified over all episodes, with each episode featuring an average of 18 unique interlocutors, posing a challenge for SD systems. These speakers have been annotated with role labels, with each speaker acting as a "host", "interviewer", or "subject". Hosts tend toward expository speech with long turns of dialog and are responsible for 42.4% of utterances. Interviewers pose questions to facilitate discussion, speaking 12.8% of the time with shorter utterances. Subjects make up the remaining 44.8% of ut-

[2]e.g., https://www.thisamericanlife.org/74/transcript

Table 1: *Comparison of benchmark datasets for ASR and SD. MPC: minutes/conversation; SPC: speakers/conversation.*

	Conv	Hours	MPC	SPC	Setting
LibriSpeech	–	1K	–	1	Book
CALLHOME	120	60	30.0	2	Phone
Switchboard	2.4K	260	6.5	2	Phone
Fisher	16.4K	2.7K	6.0	2	Phone
Clinical [1]	100K	15K	6.7	2	Medical
TAL (Ours)	663	637	**57.7**	**18**	Podcast

terances and generally speak at length. The subject matter and dialog structure also poses a challenge for transcription, with a vocabulary of 1,309,647 unique words and references to 53,792 unique named entities as identified via `spaCy`.[3]

We compare TAL to several benchmark datasets for ASR and SD in Table 1, alongside the clinical dataset used in [1]. A large body of research in ASR has focused on the 1,000-hour LibriSpeech dataset [3] of audiobook segments, while SD research focuses on telephone conversation transcripts from the Fisher [4], CALLHOME [5] and Switchboard [6] corpora. Of these datasets, only LibriSpeech and TAL are free and openly accessible. We note that RadioTalk [7] also collected a large corpus of radio program transcripts, but did so using a noisy automated system with no corresponding gold labels and did not release the audio. In contrast to other ASR datasets, TAL transcripts contain proper punctuation and casing. Professional transcribers for TAL may elect to ignore stutters and irrelevant repetitions, performing minor grammatical fixes to the spoken words. Thus, transcription models for this setting must capture higher-level semantics of the utterance. TAL also contains a diverse set of speaker accents, varying rates of speech, and background music, making it an acoustically challenging dataset.

We standardized TAL's raw audio by preprocessing all audio to 16KHz mono-channel wav format. Approximately 9% of the publicly released transcripts for TAL contain alignment errors, primarily stemming from advertising preceding each act. We manually checked the episodes, discarding 38 episodes with content-related errors and manually re-aligned and re-transcribed another 25 episodes. We split these cleaned conversations into disjoint training, validation, and test sets comprising 593, 34, and 36 episodes respectively.

3. Task

We present the TAL aligned and unaligned tasks to test a model's capability to diarize and transcribe text under bounded and unbounded conditions. The **TAL Aligned Task** measures ASR and SD performance when utterance bounds are provided. Given a single utterance of input speech $X = (x_1, \ldots, x_n)$ (in raw waveform or spectrogram format), a model must produce a sequence of vocabulary tokens $Y = (y_1, \ldots, y_m)$ and speaker labels $S = (s_1, \ldots, s_m)$ where $y_i \in V$ (vocabulary) and $s_i \in H$ (speaker IDs). We set the beginning and terminal tokens y_1, y_m as the special utterance separator `[US]` token. We only consider utterances between 3 to 30 seconds, comprising 6,774 utterances in the test set. The **TAL Unaligned Task** is similar to the above, but utterance bounds are not provided, forcing the model to conduct full-conversation ASR and SD. An hour-long podcast episode is provided as X, and the target outputs as Y and S, corresponding to the full episode transcript and gold speaker labels, respectively. To perform well under this setting, the model must learn to determine utterance alignments. This closely resembles real-world audio transcription without known segmentation.

ASR is evaluated using word error rate (WER), comparing the model's output tokens $\hat{Y}$ against reference tokens Y. We retain casing and punctuation, calculating WER over model outputs tokenized via the Punkt tokenizer [8], with incorrectly cased words counted as errors. All generated outputs that do not terminate (with the `[US]` token) are treated as 100% WER.

Prior work in the joint ASR and SD setting evaluated diarization with WDER [1], defined as $\frac{S_w+C_w}{S+C}$ where S_w is the number of ASR substitutions with the wrong speaker label, C_w is the number of correct transcriptions with the wrong speaker label, and S and C are the total number of substituted and correct words, respectively. WDER was originally proposed for a doctor-patient setting where we care about absolute label correctness. In diarization settings, however, the goal is speaker disambiguation, and different speakers *within a single conversation* must have distinct labels (that need not match their ground truth identities). Thus, while WDER can measure role classification error, it is an inappropriate metric for multi-speaker diarization error. To measure the latter, we introduce a new metric: multi-speaker word diarization error (MWDE). For MWDE we first compute the optimal alignment between output and reference speakers m out of all possible alignments M—much like computing multi-speaker diarization error rate (DER) [9, 10]—and then calculate WDER with the new alignments:

$$\text{MWDE} = \min_{m \in M} \text{WDER}_m \tag{1}$$

Like WDER, MWDE does not account for ASR additions and deletions, due to ambiguous reference speaker labels. Combined with WER, MWDE allows us to evaluate joint diarization and transcription systems based on word-level alignments, which is appropriate for practical applications.

4. Approach

We compare three different frameworks to perform ASR and diarization. As it is computationally infeasible to directly train models on TAL unaligned, we train each of our models on TAL aligned and evaluate on both the aligned and unaligned test sets.

4.1. Separate ASR and Diarization

Typically, ASR and SD on multi-speaker audio are conducted through independent pipelines. The ASR model is trained to predict the spoken words Y in the audio. The SD model is trained to produce speaker embeddings which are then clustered during inference to determine who spoke when [11] and when the speakers change. A reconciliation step is then required to assign the SD model's time-position speaker labels to the ASR model's output Y to produce word-level speaker labels S.

For our baselines, we train a sequence transduction model to perform ASR. We seek to learn hidden representations $h_i = \text{dec}(\text{enc}(X), y_{<i})$ for each token output, where $\text{enc}(\ldots)$ and $\text{dec}(\ldots)$ refer to an encoder and decoder neural network respectively. This representation is used to predict probabilities for each token $P(y_i|X, y_{<i}) = \text{softmax}(Wh_i)$ of the output vocabulary, where W is a learned weight matrix. We minimize the cross entropy loss of $P(y_i|X, y_{<i})$ against true tokens in a conventional sequence-to-sequence manner. We train a diarization-only acoustic model to classify speakers [12] from the TAL training set. We learn features for each audio frame,

[3] https://spacy.io/, `en_core_web_sm` model

and calculate cross-entropy loss of predicted speaker against the true speaker for that frame. During evaluation, we compute the weighted average features for each word using the *attention focus* (Section 4.6) and then cluster with HDBScan [13] to assign word-level speaker labels.

4.2. Joint ASR and Diarization

To explore an end-to-end approach to simultaneous ASR and SD, we can formulate our task as a joint sequence transduction task to jointly predict the spoken words, speaker identity and turn changes [1]. Prior to each utterance termination token `[US]`, we insert the speaker ID token such that such that $Y_{\text{aug}} = (y_1, \ldots, h, y_m), y_{1 \ldots m} \in V, h \in H$. At test time, once the speaker token has been produced, all preceding tokens in the utterance are assigned to that speaker $s_i = h$. To handle unseen speakers, SD systems typically use some form of clustering [14, 15, 16]. Instead, we simply use the model's predicted speaker from the training set as our label for the unseen speaker.

4.3. Boosting with Separate Diarization Model

Recent work has shown the importance of lexical information in speaker change detection [2, 17], but the same has not been shown for speaker identity prediction. To investigate this, we explore an alternative approach where we use the above joint model to predict speaker change and a separate diarization model (from Section 4.1) to produce speaker embeddings. This setting differs from the separate framework in that the speaker change bounds are determined by the joint model and not the diarization model (Figure 1). We produce an utterance embedding by averaging all speaker embeddings that fall within an utterance, reducing the noise produced by individual embeddings. We then use HDBScan [13] to cluster speakers. We report diarization results via this "boosting" with a separately-trained SD model in the **SD+** column of Table 2.

4.4. Model Architecture

While Transformer acoustic models have shown promising performance in ASR [18], in preliminary experiments we have found that they scale poorly to long sequences due to high memory requirements. Instead, we use a Time-Depth Separable Convolution (TDS) [19] acoustic model, which has a better computation to performance trade-off. We followed the same configuration as TDS except the following. Instead of an RNN, our decoder is a 4-layer Transformer [20] decoder with 512 hidden units per layer and 64-dimensional factorized token embeddings [21]. We replace all LayerNorm [22] with ReZero initialization [23] and 2D convolutions with 1D convolutions [24]. These modifications accelerated training, and our model achieves comparable ASR performance to the original [19] on the LibriSpeech *clean* (6.18% vs. 5.58%) and *other* (15.62% vs. 15.30%) development sets. We use the same architecture for all our experiments. For training diarization-only models, we use only the TDS encoder without a decoder.

4.5. Pre-training

To boost our model's performance, we propose leveraging pre-training techniques, which have been shown to successfully improve natural language [25] and computer vision tasks [26]. We pre-train the encoder [27, 28] via ASR on the LibriSpeech corpus, discarding the decoder module due to vocabulary mismatch and transferring the encoder's learned weights.

4.6. Decoding Long Conversations

TAL aligned evaluation is straightforward, and we decode from our model following conventional approaches using beam search of size 5. TAL unaligned evaluation requires decoding over an hour of audio. We introduce an algorithm we call *striding window attention* to enable our model to efficiently decode full hour-long episodes. First, we run the full episode through the WebRTC[4] Voice Activity Detector (VAD) to remove non-speech segments and compute full-episode audio features using our TDS encoder. Due to memory limitations we can only attend to a window of features covering a 30-second receptive field when generating a particular output. The key challenge is to effectively stride this attention window to relevant audio segments without the aid of word boundaries during decoding.

We estimate the *attention focus* (AF) of our model, that is, the likely time-position from which the model decodes a given output token. We observe that attention patterns increase monotonically as tokens are decoded, so we heuristically define AF as the average attention weight position of all decoder layers and attention heads. When the AF shifts beyond a fixed proportion of our current attention window, we advance the window forward and proportionally truncate the decoder's context history. This operation is repeated until we have decoded the entire episode. A naive implementation of our algorithm often enters repetitive loops when encountering unintelligible speech or lyrical music not detected by our VAD, a known issue in attention models and neural text generation [29, 30]. We discover two patterns that arise when this happens: the number of n-gram repeats increases and the AF stops increasing. We use an n-gram repetition detector and track AF changes to recover from these errors by pruning out repeating n-grams. We use greedy search for unaligned evaluation due to memory limitations.

4.7. Data Augmentation

One issue with training speech models on well-aligned single utterances is that they cannot learn inter-utterance dependencies and adapt to imprecise utterance bounds. We propose **ShiftAug**, which augments the dataset with random 10 to 30 second audio segments from episodes and trains the model to output the text of all utterances whose bounds lie within the sampled span. We truncate the text of utterances that lie partially within these bounds proportional to the amount of intersecting audio. This method is noisy because a sampled audio segment may not contain all the words in the utterance. To address this issue, we introduce **AlignAug**, which uses heuristic forced-word alignments using the `Aeneas` tool[5]) to guide truncation.

5. Results

5.1. Framework Comparison

We report results from all models in Table 2. We first compare the separate and joint frameworks for TAL aligned. We find that separately trained ASR and SD models (Separate), when reconciled, obtain modest performance. The same model trained jointly for ASR and SD (Joint) has a minor degradation in ASR performance—similar to findings in [1]—but the SD produced using speaker tokens is significantly worse than clustering embeddings from a separate SD model. As our decoder learns embeddings for each vocabulary token, including each

[4] https://webrtc.org/

[5] https://www.readbeyond.it/aeneas/

speaker token, we also tried clustering representations consisting of a weighted sum of speaker vocabulary embeddings for each utterance. This method, however, yielded a few percent worse MWDE than using predicted speaker IDs directly. We also trained an alternative joint setup where we used a separate speaker head in the decoder to classify speakers and treated speaker classification and ASR as a multitask loss, but this model was unable to converge.

We find that using clustering from an external diarization model confers significant benefits over merely using a joint model's speaker predictions in the aligned case (SD+ column). However, for the unaligned case, the joint model wins at diarization when the training is augmented by alignment training in the form of shifted augmentations. Using our joint framework to determine utterance bounds reduces the unaligned MWDE to 62.2%, a 29.1% absolute reduction. Our results suggest that ASR and speaker identification may be conflicting tasks better suited for separately trained models, or may require more sophisticated multi-task learning schemes.

The presence of casing and punctuation in TAL contributes to its difficulty—we simulated the evaluation methodology of other ASR datasets (unifying casing and stripping punctuation) and observed 13.9% and 38.2% WER for the aligned and unaligned settings of our ShiftAug model, respectively.

5.2. Pre-training

We find that pre-training the acoustic model on LibriSpeech provides a 3.5% absolute improvement to ASR performance in the aligned task, and a more pronounced 8.9% improvement to ASR in unaligned. This suggests that pre-training acoustic models on large audio corpora helps in learning useful features. Our augmentation models builds upon this pre-training method.

5.3. Unaligned Performance and Data Augmentation

Overall, unaligned performance lags behind aligned performance by a significant margin even with pre-training. In the unaligned setting, the separate framework relies on clustering speaker embeddings from the SD model (trained on speaker classification) for speaker change detection [11]. This method performs poorly in diarization, with the separate framework achieving 91% MWDE on unaligned. Empirically, we find the SD model is unable to determine relative speaker boundaries on TAL unaligned, primarily due to highly variable microphone quality, lyrical music, and intra-conversation speaker diversity. We find that using a jointly trained model to determine bounds, then averaging speaker embeddings with the utterances (Joint SD+) leads to much more stable predictions and more reasonable MWDE. We hypothesize that this is because speaker embeddings in TAL at fine-granularity are noisy and order-less, making it difficult to cluster properly [31].

A qualitative inspection suggests that accumulated VAD and utterance pruning errors contribute to the the disparity between aligned and unaligned performance. Without data augmentation, our models perform poorly on TAL unaligned, as they are unable to learn inter-speaker utterance boundaries. We find that ShiftAug is able to close some of the performance gap between unaligned and aligned results especially in diarization, and generally performs better in ASR than AlignAug likely due to the regularization from its noise. Manual inspection of heuristic word boundaries from AlignAug reveals that the heuristic is overly conservative, at many times pruning excess tokens. Augmentation improves joint prediction of speakers (SD), but we see no corresponding benefit when boosting with a separate SD model (SD+), suggesting that it improves speaker identification more than speaker boundary determination.

Table 2: *Model performance on TAL test set. ASR and SD/SD+ are evaluated via WER and MWDE in percentages respectively. SD is computed directly from the joint model's speaker tokens, whereas SD+ uses clustering from external diarization model.*

	Aligned			**Unaligned**		
	ASR	**SD**	**SD+**	**ASR**	**SD**	**SD+**
Separate	24.3	–	**15.4**	58.3	–	91.3
Joint	25.4	31.9	15.7	58.2	62.2	**54.0**
+ Pre-training	21.9	29.5	15.7	49.3	63.8	54.6
+ ShiftAug	**18.9**	29.1	15.6	**42.1**	38.2	55.8
+ AlignAug	19.1	**28.5**	15.7	51.0	**37.4**	55.2

6. Related Work

Joint ASR and SD: Joint ASR and SD within a single model was first explored by [32], who spliced together audiobook snippets as a proxy for conversation. [33] presented an alternating optimization strategy to jointly extract speaker embeddings and text in target-speaker transcription. [1] expanded the work into two-party conversations but did not release their dataset. Our work represents the first effort to jointly transcribe and diarize audio in a real-world setting with multiple speakers, punctuation, and casing, on a publicly available dataset.

Pre-training: Various approaches have been proposed for unsupervised pre-training of acoustic models to leverage large corpora of unlabeled audio, such as contrastive predictive coding [34, 35], pseudo-labeling [36], and masked audio modeling [37]. We conduct supervised pre-training using labeled examples from Librispeech as we were unable to scale some of these more complex approaches to the TDS architecture in our preliminary experiments.

Monotonic Attention: Traditional systems for transcribing long conversations rely on carefully-engineered pipelines and audio segmentation [38]. Our method is more similar to attention mechanisms that monotonically advance in time [39, 40, 41]. Our decoding algorithm is inspired by [42], who show that the learned attention peak position is a good heuristic for advancing monotonic attention.

7. Conclusion

We present a new benchmark for ASR and SD in an extended multi-speaker conversational setting and explore three frameworks for learning both tasks. We find that models that jointly learn ASR and SD perform best in the absence of known utterance bounds. When bounds are provided, boosting with an external SD model improves diarization. We introduce an algorithm that enables scaling ASR and SD to hour-long conversations and show significant performance improvement by incorporating pre-training and data augmentation methods. TAL unaligned presents a new challenge for rich transcription of extended conversation. We see opportunities for future work to investigate better pre-training and decoding algorithms.

Acknowledgements: This work was supported in part by NSF awards CNS-1730158, ACI-1540112, ACI-1541349, OAC-1826967, IIS-1750063, the UC Office of the President, the UCSD's California Institute for Telecommunications and Information Technology/Qualcomm Institute, CENIC for the 100Gpbs networks and the Alexa Prize Grand Challenge 3.

8. References

[1] L. E. Shafey, H. Soltau, and I. Shafran, "Joint speech recognition and speaker diarization via sequence transduction," in *Interspeech*, 2019.

[2] T. J. Park and P. G. Georgiou, "Multimodal speaker segmentation and diarization using lexical and acoustic cues via sequence to sequence neural networks," in *Interspeech*, 2018.

[3] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*. IEEE, 2015.

[4] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text," in *LREC*, 2004.

[5] A. Canavan, D. Graff, and G. Zipperlen, "Callhome american english speech," *Linguistic Data Consortium*, 1997.

[6] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *ICASSP*, vol. 1. IEEE, 1992.

[7] D. Beeferman, W. Brannon, and D. Roy, "Radiotalk: A large-scale corpus of talk radio transcripts," in *Interspeech*, 2019.

[8] T. Kiss and J. Strunk, "Unsupervised multilingual sentence boundary detection," *Computational Linguistics*, vol. 32, no. 4, pp. 485–525, 2006.

[9] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[10] O. Galibert, "Methodologies for the evaluation of speaker diarization and automatic speech recognition in the presence of overlapping speech," in *Interspeech*, 2013.

[11] R. Yin, H. Bredin, and C. Barras, "Neural speech turn segmentation and affinity propagation for speaker diarization," in *Interspeech*, 2018.

[12] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," in *Interspeech*, 2019.

[13] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *PAKDD*, 2013.

[14] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *ICASSP*, 2017.

[15] G. Sell and D. Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in *SLT*, 2014.

[16] A. Zhai and H. Wu, "Classification is a strong baseline for deep metric learning," in *BMVC*, 2019.

[17] Z. Meng, L. Mou, and Z. Jin, "Hierarchical RNN with static sentence-level attention for text-based speaker change detection," in *CIKM*, 2017.

[18] Y. Wang, A. Mohamed, D. Le, C. Liu, A. Xiao, J. Mahadeokar, H. Huang, A. Tjandra, X. Zhang, F. Zhang, C. Fuegen, G. Zweig, and M. L. Seltzer, "Transformer-based acoustic modeling for hybrid speech recognition," *CoRR*, vol. abs/1910.09799, 2019.

[19] A. Hannun, A. Lee, Q. Xu, and R. Collobert, "Sequence-to-sequence speech recognition with time-depth separable convolutions," in *Interspeech*. ISCA, 2019.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[21] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," *CoRR*, vol. abs/1909.11942, 2019.

[22] L. J. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *CoRR*, vol. abs/1607.06450, 2016.

[23] T. Bachlechner, B. P. Majumder, H. H. Mao, G. W. Cottrell, and J. J. McAuley, "Rezero is all you need: Fast convergence at large depth," *CoRR*, vol. abs/2003.04887, 2020.

[24] V. Pratap, Q. Xu, J. Kahn, G. Avidov, T. Likhomanenko, A. Hannun, V. Liptchinsky, G. Synnaeve, and R. Collobert, "Scaling up online speech recognition using convnets," 2020.

[25] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*. Association for Computational Linguistics, 2019.

[26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[27] S. Toshniwal, A. Kannan, C. Chiu, Y. Wu, T. N. Sainath, and K. Livescu, "A comparison of techniques for language model integration in encoder-decoder speech recognition," in *SLT*, 2018.

[28] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Interspeech*, 2019.

[29] P. Keung, W. Niu, Y. Lu, J. Salazar, and V. Bhardwaj, "Attentional speech recognition models misbehave on out-of-domain utterances," *CoRR*, vol. abs/2002.05150, 2020.

[30] S. Welleck, I. Kulikov, S. Roller, E. Dinan, K. Cho, and J. Weston, "Neural text generation with unlikelihood training," in *ICLR*, 2020.

[31] A. Zhang, Q. Wang, Z. Zhu, J. W. Paisley, and C. Wang, "Fully supervised speaker diarization," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*. IEEE, 2019, pp. 6301–6305.

[32] A. Sarkar, S. Dasgupta, S. K. Naskar, and S. Bandyopadhyay, "Says who? deep learning models for joint speech recognition, segmentation and diarization," in *ICASSP*, 2018.

[33] N. Kanda, S. Horiguchi, Y. Fujita, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," 2019.

[34] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *CoRR*, vol. abs/1904.05862, 2019.

[35] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.

[36] G. Synnaeve, Q. Xu, J. Kahn, E. Grave, T. Likhomanenko, V. Pratap, A. Sriram, V. Liptchinsky, and R. Collobert, "End-to-end ASR: from supervised to semi-supervised learning with modern architectures," *CoRR*, vol. abs/1911.08460, 2019.

[37] A. T. Liu, S. Yang, P. Chi, P. Hsu, and H. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," *CoRR*, vol. abs/1910.12638, 2019.

[38] T. Hain, V. Wan, L. Burget, M. Karafiát, J. Dines, J. Vepa, G. Garau, and M. Lincoln, "The AMI system for the transcription of speech in meetings," in *ICASSP*, 2007.

[39] C. Chiu and C. Raffel, "Monotonic chunkwise attention," *CoRR*, vol. abs/1712.05382, 2017.

[40] X. Ma, J. Pino, J. Cross, L. Puzon, and J. Gu, "Monotonic multihead attention," *CoRR*, vol. abs/1909.12406, 2019.

[41] R. Fan, P. Zhou, W. Chen, J. Jia, and G. Liu, "An online attention-based model for speech recognition," *CoRR*, vol. abs/1811.05247, 2018.

[42] A. Merboldt, A. Zeyer, R. Schlüter, and H. Ney, "An analysis of local monotonic attention variants," in *Interspeech 2019*, 2019.