DEEP CONTEXTUALIZED ACOUSTIC REPRESENTATIONS FOR SEMI-SUPERVISED SPEECH RECOGNITION

Shaoshi Ling, Yuzong Liu, Julian Salazar, Katrin Kirchhoff

Amazon AWS AI

{shaosl,liuyuzon,julsal,katrinki}@amazon.com

ABSTRACT

We propose a novel approach to semi-supervised automatic speech recognition (ASR). We first exploit a large amount of unlabeled audio data via representation learning, where we reconstruct a temporal slice of filterbank features from past and future context frames. The resulting *deep contextualized acoustic representations* (DeCoAR) are then used to train a CTC-based end-to-end ASR system using a smaller amount of labeled audio data. In our experiments, we show that systems trained on DeCoAR consistently outperform ones trained on conventional filterbank features, giving 42% and 19% relative improvement over the baseline on WSJ *eval92* and LibriSpeech *test-clean*, respectively. Our approach can drastically reduce the amount of labeled data required; unsupervised training on LibriSpeech then supervision with 100 hours of labeled data achieves performance on par with training on all 960 hours directly.

Index Terms— speech recognition, acoustic representation learning, semi-supervised learning

1. INTRODUCTION

Current state-of-the-art models for speech recognition require vast amounts of transcribed audio data to attain good performance. In particular, end-to-end ASR models are more demanding in the amount of training data required when compared to traditional hybrid models. While obtaining a large amount of labeled data requires substantial effort and resources, it is much less costly to obtain abundant unlabeled data.

For this reason, *semi-supervised learning* (SSL) is often used when training ASR systems. The most commonly-used SSL approach in ASR is self-training [1, 2, 3, 4, 5]. In this approach, a smaller labeled set is used to train an initial seed model, which is applied to a larger amount of unlabeled data to generate hypotheses. The unlabeled data with the most reliable hypotheses are added to the training data for re-training. This process is repeated iteratively. However, self-training is sensitive to the quality of the hypotheses and requires careful calibration of the confidence measures. Other SSL approaches include: pre-training on a large amount of unlabeled data with restricted Boltzmann machines (RBMs) [6]; entropy minimization [7, 8, 9], where the uncertainty of the unlabeled data is incorporated as part of the training objective; and graph-based approaches [10], where the manifold smoothness assumption is exploited. Recently, transfer learning from large-scale pre-trained language models (LMs) [11, 12, 13] has shown great success and achieved state-of-the-art performance in many NLP tasks. The core idea of these approaches is to learn efficient word representations by pre-training on massive amounts of unlabeled text via word comple-

tion. These representations can then be used for downstream tasks with labeled data.

Inspired by this, we propose an SSL framework that learns efficient, context-aware acoustic representations using a large amount of unlabeled data, and then applies these representations to ASR tasks using a limited amount of labeled data. In our implementation, we perform acoustic representation learning using forward and backward LSTMs and a training objective that minimizes the reconstruction error of a temporal slice of filterbank features given previous and future context frames. After pre-training, we fix these parameters and add output layers with connectionist temporal classification (CTC) loss for the ASR task.

The paper is organized as follows: in Section 2, we give a brief overview of related work in acoustic representation learning and SSL. In Section 3, we describe an implementation of our SSL framework with DeCoAR learning. We describe the experimental setup in Section 4 and the results on WSJ and LibriSpeech in Section 5, followed by our conclusions in Section 6.

2. RELATED WORK

While semi-supervised learning has been exploited in a plethora of works in hybrid ASR system, there are very few work done in the end-to-end counterparts [4, 14, 15]. In [4], an intermediate representation of speech and text is learned via a shared encoder network. To train these representation, the encoder network was trained to optimize a combination of ASR loss, text-to-text autoencoder loss and inter-domain loss. The latter two loss functions did not require paired speech and text data. Learning efficient acoustic representation can be traced back to restricted Boltzmann machine [16, 17, 18], which allows pre-training on large amounts of unlabeled data before training the deep neural network acoustic models.

More recently, acoustic representation learning has drawn increasing attention [19, 20, 21, 22, 23, 24] in speech processing. For example, an autoregressive predictive coding model (APC) was proposed in [21] for unsupervised speech representation learning and was applied to phone classification and speaker verification. WaveNet auto-encoders [22] proposed contrastive predictive coding (CPC) to learn speech representations and was applied on unsupervised acoustic unit discovery task. Wav2vec [23] proposed a multi-layer convolutional neural network optimized via a noise contrastive binary classification and was applied to WSJ ASR tasks.

Unlike the speech representations described in [23, 21], our representations are optimized to use bi-directional contexts to autoregressively reconstruct unseen frames. Thus, they are deep contextualized representations that are functions of the entire input sentence. More importantly, our work is a general semi-supervised training framework that can be applied to different systems and
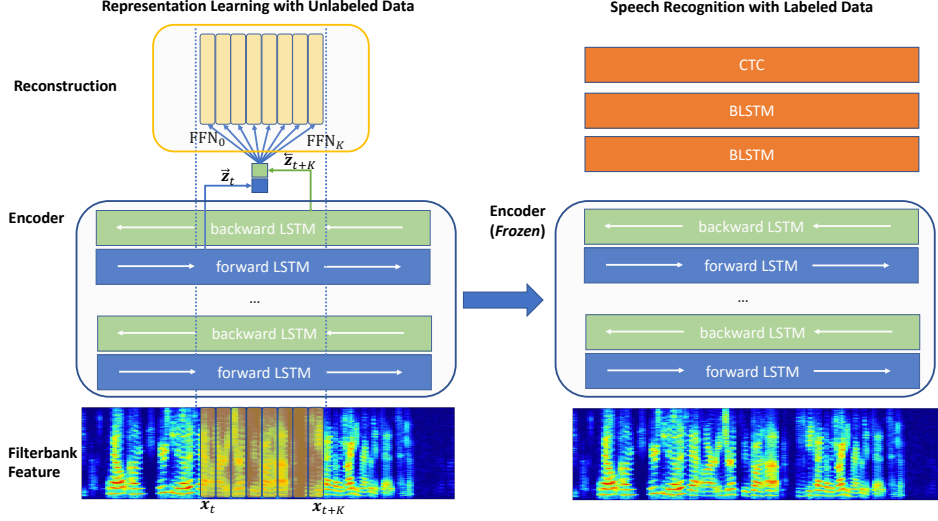
Fig. 1. Illustration of our semi-supervised speech recognition system.

requires no architecture change.

3. DEEP CONTEXTUALIZED ACOUSTIC REPRESENTATIONS

3.1. Representation learning from unlabeled data

Our approach is largely inspired by ELMo [11]. In ELMo, given a sequence of T tokens $(w_1, w_2, ..., w_T)$, a forward language model (implemented with an LSTM) computes its probability using the chain rule decomposition:

$$p(w_1, w_2, \cdots, w_T) = \prod_{t=1}^{T} p(w_k \mid w_1, w_2, \ldots, w_{t-1}).$$

Similarly, a backward language model computes the sequence probability by modeling the probability of token w_t given its future context $w_{t+1}, \ldots, w_T$ as follows:

$$p(w_1, w_2, \cdots, w_T) = \prod_{t=1}^{T} p(w_t \mid w_{t+1}, w_{t+2}, \ldots, w_T)$$

ELMo is trained by maximizing the joint log-likelihood of both forward and backward language model probabilities:

$$\sum_{t=1}^{T} [\log p(w_t \mid w_1, w_2, ..., w_{t-1}; \Theta_x, \overrightarrow{\Theta}_{\text{LSTM}}, \Theta_s)$$
$$+ \log p(w_t \mid w_{t+1}, w_{t+2}, ..., w_T; \Theta_x, \overleftarrow{\Theta}_{\text{LSTM}}, \Theta_s)] \quad (1)$$

where Θ_x is the parameter for the token representation layer, Θ_s is the parameter for the softmax layer, and $\overrightarrow{\Theta}_{\text{LSTM}}$, $\overleftarrow{\Theta}_{\text{LSTM}}$ are the parameters of forward and backward LSTM layers, respectively. As the word representations are learned with neural networks that use past and future information, they are referred to as deep contextualized word representations.

For speech processing, the input audio data is represented as a sequence of acoustic features $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_T)$ (e.g. filterbank or MFCC features). Predicting an acoustic feature $\mathbf{x}_t$ may be a trivial

task, as it could be solved by exploiting the temporal smoothness of the signal. In the APC model [21], the authors propose to predict a frame K steps ahead of the current one. Namely, the model aims to minimize the ℓ_1 loss between an acoustic feature vector $\mathbf{x}$ at time $t + K$ and a reconstruction $\mathbf{y}$ predicted at time t: $\sum_{t=1}^{T-K} |\mathbf{x}_{t+K} - \mathbf{y}_t|$. They conjectured this would induce the model to learn more global structure rather than simply leveraging local information within the signal.

We propose combining the bidirectionality of ELMo and the reconstruction objective of APC to give *deep contextualized acoustic representations* (DeCoAR). We train the model to predict a *slice* of acoustic feature vectors, given past and future acoustic vectors. As depicted on the left side of Figure 1, a stack of forward and backward LSTMs are applied to the entire unlabeled input sequence $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_T)$, where $\mathbf{x}_t$ is 40-dimensional log fiterbank feature. The network computes a hidden representation that encodes information from both previous and future frames (i.e. $\overrightarrow{\mathbf{z}}_t$, $\overleftarrow{\mathbf{z}}_t$) for each frame $\mathbf{x}_t$. Given a sequence of acoustic feature inputs $(\mathbf{x}_1, ..., \mathbf{x}_T) \in \mathbb{R}^d$, for each slice $(\mathbf{x}_t, \mathbf{x}_{t+1}, ..., \mathbf{x}_{t+K})$ starting at time step t, our objective is defined as follows:

$$\mathcal{L}_t = \sum_{i=0}^{K} |\mathbf{x}_{t+i} - \text{FFN}_i([\overrightarrow{\mathbf{z}}_t; \overleftarrow{\mathbf{z}}_{t+K}])| \quad (2)$$

where $[\overrightarrow{\mathbf{z}}_t; \overleftarrow{\mathbf{z}}_t] \in \mathbb{R}^{2h}$ are the concatenated forward and backward states from the last LSTM layer, and

$$\text{FFN}_i(\mathbf{v}) = \mathbf{W}_{i,2}\text{ReLU}(\mathbf{W}_{i,1}\mathbf{v} + \mathbf{b}_{i,1}) + \mathbf{b}_{i,2} \quad (3)$$

is a position-dependent feed-forward network with 512 hidden dimensions. The final loss $\mathcal{L}$ is summed over all possible slices in the entire sequence:

$$\mathcal{L} = \sum_{t=1}^{T-K} \mathcal{L}_t \quad (4)$$

Note this can be implemented efficiently as a layer which predicts these $(K+1)$ frames at each position t, all at once. We compare with the use of unidirectional LSTMs and various slice sizes in Section 5.

Representation	100 hours		360 hours		460 hours		960 hours	
	test-clean	test-other	test-clean	test-other	test-clean	test-other	test-clean	test-other
filterbank	9.36	30.20	7.57	25.28	7.11	24.31	5.82	14.50
wav2vec [23]	6.92	20.00	6.26	18.17	6.01	17.00	5.12	13.07
DeCoAR	6.10	17.43	5.23	14.67	5.12	14.10	4.74	12.20

Table 1. Semi-supervised LibriSpeech results.

3.2. End-to-end ASR training with labeled data

After we have pre-trained the DeCoAR on unlabeled data, we freeze the parameters in the architecture. To train an end-to-end ASR system using labeled data, we remove the reconstruction layer and add two BLSTM layers with CTC loss [25], as illustrated on the right side of Figure 1. The DeCoAR vectors induced by the labeled data in the forward and backward layers are concatenated. We fine-tune the parameters of this ASR-specific new layer on the labeled data.

While we use LSTMs and CTC loss in our implementation, our SSL approach should work for other layer choices (e.g. TDNN, CNN, self-attention) and other downstream ASR models (e.g. hybrid, seq2seq, RNN transducers) as well.

4. EXPERIMENTAL SETUP

4.1. Data

We conducted our experiments on the WSJ and LibriSpeech datasets, pre-training by using one of the two training sets as unlabeled data. To simulate the SSL setting in WSJ, we used 30%, 50% as well as 100% of labeled data for ASR training, consisting of 25 hours, 40 hours, and 81 hours, respectively. We used *dev93* for validation and *eval92* and evaluation. For LibriSpeech, the amount of training data used varied from 100 hours to the entire 960 hours. We used *dev-clean* for validation and *test-clean*, *test-other* for evaluation.

4.2. ASR systems

Our experiments consisted of three different setups: 1) a fully-supervised system using all labeled data; 2) an SSL system using wav2vec features; 3) an SSL system using our proposed DeCoAR features. All models used were based on deep BLSTMs with the CTC loss criterion.

In the supervised ASR setup, we used conventional log-mel filterbank features, which were extracted with a 25ms sliding window at a 10ms frame rate. The features were normalized via mean subtraction and variance normalization on a per-speaker basis. The model had 6 BLSTM layers, with 512 cells in each direction. We found that increasing the number of cells to a larger number did not further improve the performance and thus used it as our best supervised ASR baseline. The output CTC labels were 71 phonemes[1] plus one blank symbol.

In the SSL ASR setup, we pre-trained a 4-layer BLSTM (1024 cells per sub-layer) to learn DeCoAR features according to the loss defined in Equation 2 and use a slice size of 18. We optimized the network with SGD and use a Noam learning rate schedule, where we started with a learning rate of 0.001, gradually warm up for 500 updates, and then perform inverse square-root decay. We grouped the input sequences by length with a batch size of 64, and trained the models on 8 GPUs. After the representation network was trained,

we froze the parameters, and added a projection layer, followed by 2-layer BLSTM with CTC loss on top it. We fed the labeled data to the network. For comparison, we obtained 512-dimensional wav2vec representations [23] from the wav2vec-large model[2]. Their model was pre-trained on 960-hour LibriSpeech data with constrastive loss and had 12 convolutional layers with skip connections.

For evaluation purposes, we applied WFST-based decoding using EESEN [26]. We composed the CTC labels, lexicons and language models (unpruned trigram LM for WSJ, 4-gram for LibriSpeech[3]) into a decoding graph. The acoustic model score was set to 0.8 and 1.0 for WSJ and LibriSpeech, respectively, and the blank symbol prior scale was set to 0.3 for both tasks. We report the performance in word error rate (WER).

Representation	Unlabeled	Labeled	dev93	eval92
filterbank	-	81h	8.21	5.44
wav2vec [23]	960h Libri.	81h	6.84	3.97
DeCoAR	960h Libri.	81h	6.30	3.17
filterbank	-	25h	18.16	11.04
filterbank	-	40h	13.50	9.20
DeCoAR	81h WSJ	25h	10.38	5.81
DeCoAR	81h WSJ	40h	9.41	5.09
DeCoAR	81h WSJ	81h	8.34	4.64

Table 2. Semi-supervised WSJ results. *Unlabeled* indicates the amount of unlabeled data used for acoustic representation learning, and *Labeled* indicates the amount of labeled data in ASR training.

5. RESULTS

5.1. Semi-supervised WSJ results

Table 2 shows our results on semi-supervised WSJ. We demonstrate that DeCoAR feature outperforms filterbank and wav2vec features, with a relative improvement of 42% and 20%, respectively. The lower part of the table shows that with smaller amounts of labeled data, the DeCoAR features are significantly better than the filterbank features: Compared to the system trained on 100% labeled data with filterbank features, we achieve comparable results on *eval92* using 30% of the labeled data and better performance on *eval92* using 50% of the labeled data.
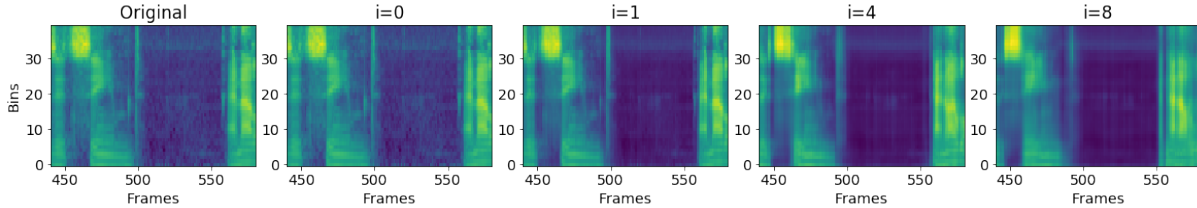
5.2. Semi-supervised LibriSpeech results

Table 1 shows the results on semi-supervised LibriSpeech. Both our representations and wav2vec[23] are trained on 960h LibriSpeech data. We conduct our semi-supervised experiments using 100h

---

Fig. 2. The spectrograms for a portion of LibriSpeech *dev-clean* utterance 2428-83699-0034 reconstructed by generated by taking the i-th frame prediction (slice size 18) at each time step. The reconstruction becomes less noisy but more simplistic when predicting further into the masked slice.

(*train-clean-100*), 360h (*train-clean-360*), 460h, and 960h of training data. Our approach outperforms both the baseline and wav2vec model in each SSL scenario. One notable observation is that using only 100 hours of transcribed data achieves very similar performance to the system trained on the full 960-hour data with filterbank features. On the more challenging *test-other* dataset, we also achieve performance on par with the filterbank baseline using a 360h subset. Furthermore, training with with our DeCoAR features approach improves the baseline even when using the exact same training data (960h). Note that while [27] introduced SpecAugment to significantly improve LibriSpeech performance via data augmentation, and [28] achieved state-of-the-art results using both hybrid and end-to-end models, our approach focuses on the SSL case with less labeled training data via our DeCoAR features.

5.3. Ablation Study and Analysis

5.3.1. Context window size

We study the effect of the context window size during pre-training. Table 3 shows that masking and predicting a larger slice of frames can actually degrade performance, while increasing training time. A similar effect was found in SpanBERT [29], another deep contextual word representation which found that masking a mean span of 3.8 consecutive words was ideal for their word reconstruction objective.

Slice size	dev93	eval92
12	6.58	3.50
18	6.30	3.17
22	6.62	3.62

Table 3. Comparison of WERs on WSJ after pre-training with different slice window sizes ($K + 1$) on LibriSpeech.

5.3.2. Unidirectional versus bidirectional context

Next, we study the importance of bidirectional context by training a unidirectional LSTM, which corresponds to only using $\overrightarrow{\mathbf{z}}_t$ to predict $\mathbf{x}_t, \ldots, \mathbf{x}_{t+K}$. Table 4 shows that this unidirectional model achieves comparable performance to the wav2vec model [23], suggesting that bidirectionality is the largest contributor to DeCoAR's improved performance.

5.3.3. DeCoAR as denoiser

Since our model is trained by predicting masked frames, DeCoAR has the side effect of learning decoder feed-forward networks FFN_i

Representation	Context	dev93	eval92
filterbank	–	8.21	5.44
wav2vec	unidirectional	6.84	3.97
DeCoAR	unidirectional	6.87	3.62
DeCoAR	bidirectional	6.30	3.17

Table 4. Comparison of WERs on WSJ after pre-training using unidirectional or bidirectional context on LibriSpeech.

which reconstruct the $(t + i)$-th filterbank frame from contexts $\overrightarrow{\mathbf{z}}_t$ and $\overleftarrow{\mathbf{z}}_{t+K}$. In this section, we consider the spectrogram reconstructed by taking the output of FFN_i at all times t.

The qualitative result is depicted in Figure 2 where the slice size is 18. We see that when $i = 0$ (i.e., when reconstructing the t-th frame from $[\overrightarrow{\mathbf{z}}_t; \overleftarrow{\mathbf{z}}_{t+K}]$), the reconstruction is almost perfect. However, as soon as one predicts unseen frames $i = 1, 4, 8$ (of 16), the reconstruction becomes more simplistic, but not by much. Background energy in the silent frames 510-550 is zeroed out. By $i = 8$ artifacts begin to occur, such as an erroneous sharp band of energy being predicted around frame 555. This behavior is compatible with recent NLP works that interpret contextual word representations as denoising autoencoders [13].

The surprising ability of DeCoAR to broadly reconstruct a frame $\overrightarrow{\mathbf{x}}_{t+K/2}$ in the middle of a missing 16-frame slice suggests that its representations $[\overrightarrow{\mathbf{z}}_t; \overleftarrow{\mathbf{z}}_{t+K}]$ capture longer-term phonetic structure during unsupervised pre-training, as with APC [21]. This motivates its success in the semi-supervised ASR task with only two additional layers, as it suggests DeCoAR learns phonetic representations similar to those likely learned by the first 4 layers of a corresponding end-to-end ASR model.

6. CONCLUSION

In this paper, we introduce a novel semi-supervised learning approach for automatic speech recognition. We first propose a novel objective for a deep bidirectional LSTM network, where large amounts of unlabeled data are used to learn deep contextualized acoustic representations (DeCoAR). These DeCoAR features are used as the representations of labeled data to train a CTC-based end-to-end ASR model. In our experiments, we show a 42% relative improvement on WSJ compared to a baseline trained on log-mel filterbank features. On LibriSpeech, we achieve similar performance to training on 960 hours of labeled by pretraining then using only 100 hours of labeled data. While we use BLSTM-CTC as our ASR model, our approach can be applied to other end-to-end ASR models.

7. REFERENCES

[1] Samuel Thomas, Michael L Seltzer, Kenneth Church, and Hynek Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *ICASSP*. IEEE, 2013, pp. 6704–6708.

[2] Yan Huang, Yongqiang Wang, and Yifan Gong, "Semi-supervised training in deep learning acoustic model," in *Interspeech*, 2016, pp. 3848–3852.

[3] Vimal Manohar, Hossein Hadian, Daniel Povey, and Sanjeev Khudanpur, "Semi-supervised training of acoustic models using lattice-free mmi," in *ICASSP*. IEEE, 2018, pp. 4844–4848.

[4] Shigeki Karita, Shinji Watanabe, Tomoharu Iwata, Atsunori Ogawa, and Marc Delcroix, "Semi-supervised end-to-end speech recognition.," in *Interspeech*, 2018, pp. 2–6.

[5] Sree Hari Krishnan Parthasarathi and Nikko Strom, "Lessons from building acoustic models with a million hours of speech," in *ICASSP*. IEEE, 2019, pp. 6670–6674.

[6] Karel Veselỳ, Mirko Hannemann, and Lukáš Burget, "Semi-supervised training of deep neural networks," in *ASRU*. IEEE, 2013, pp. 267–272.

[7] Yves Grandvalet and Yoshua Bengio, "Semi-supervised learning by entropy minimization," in *Advances in neural information processing systems*, 2005, pp. 529–536.

[8] Jui-Ting Huang and Mark Hasegawa-Johnson, "Semi-supervised training of gaussian mixture models by conditional entropy minimization," in *Interspeech*, 2010, pp. 1353–1356.

[9] Dong Yu, Balakrishnan Varadarajan, Li Deng, and Alex Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," *Computer Speech & Language*, vol. 24, no. 3, pp. 433–444, 2010.

[10] Yuzong Liu and Katrin Kirchhoff, "Graph-based semisupervised learning for acoustic modeling in automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 11, pp. 1946–1956, 2016.

[11] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171–4186.

[13] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le, "XLNet: Generalized autoregressive pretraining for language understanding," *arXiv preprint arXiv:1906.08237*, 2019.

[14] Murali Karthick Baskar, Shinji Watanabe, Ramon Astudillo, Takaaki Hori, Lukáš Burget, and Jan Černocký, "Semi-Supervised Sequence-to-Sequence ASR Using Unpaired Speech and Text," in *Interspeech*, 2019, pp. 3790–3794.

[15] Subhadeep Dey, Petr Motlicek, Trung Bui, and Franck Dernoncourt, "Exploiting Semi-Supervised Training Through a Dropout Regularization in End-to-End Speech Recognition," in *Interspeech*, 2019, pp. 734–738.

[16] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[17] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdelrahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, et al., "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal processing magazine*, vol. 29, 2012.

[18] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle, "Greedy layer-wise training of deep networks," in *Advances in neural information processing systems*, 2007, pp. 153–160.

[19] Wei-Ning Hsu and James Glass, "Extracting domain invariant features by unsupervised learning for robust automatic speech recognition," in *ICASSP*. IEEE, 2018, pp. 5614–5618.

[20] Yu-An Chung and James Glass, "Speech2Vec: A sequence-to-sequence framework for learning word embeddings from speech," in *Interspeech*, 2018, pp. 811–815.

[21] Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass, "An unsupervised autoregressive model for speech representation learning," in *Interspeech*, 2019, pp. 146–150.

[22] Jan Chorowski, Ron J Weiss, Samy Bengio, and Aäron van den Oord, "Unsupervised speech representation learning using wavenet autoencoders," *arXiv preprint arXiv:1901.08810*, 2019.

[23] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Interspeech*, 2019, pp. 3465–3469.

[24] Alexei Baevski, Steffen Schneider, and Michael Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," *arXiv preprint arXiv:1910.05453*, 2019.

[25] Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006, pp. 369–376.

[26] Yajie Miao, Mohammad Gowayyed, and Florian Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *ASRU*. IEEE, 2015, pp. 167–174.

[27] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Interspeech*, 2019, pp. 2613–2617.

[28] Christoph Lüscher, Eugen Beck, Kazuki Irie, Markus Kitza, Wilfried Michel, Albert Zeyer, Ralf Schlüter, and Hermann Ney, "RWTH ASR systems for LibriSpeech: Hybrid vs attention," in *Interspeech*, 2019, pp. 231–235.

[29] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy, "SpanBERT: Improving pretraining by representing and predicting spans," *arXiv preprint arXiv:1907.10529*, 2019.