SCaLa: Supervised Contrastive Learning for End-to-End Speech Recognition

Li Fu, Xiaoxiao Li, Runyu Wang, Lu Fan, Zhengchen Zhang,
Meng Chen, Youzheng Wu, Xiaodong He

JD AI Research, Beijing, China

{fuli3,lixiaoxiao10,wangrunyu3,fanlu,zhangzhengchen1,
chenmeng20,wuyouzheng1,hexiaodong}@jd.com

Abstract

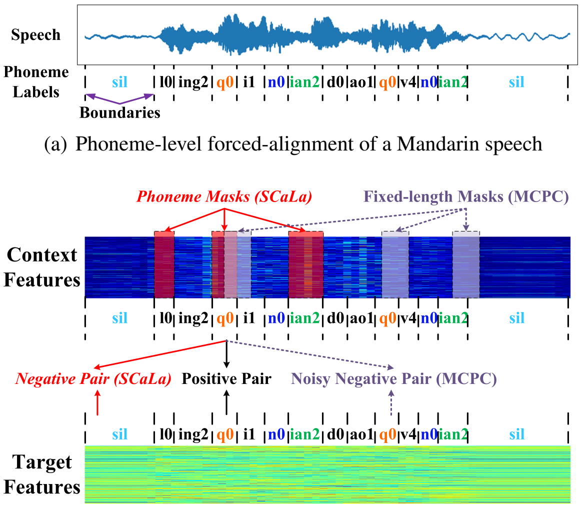
End-to-end Automatic Speech Recognition (ASR) models are usually trained to optimize the loss of the whole token sequence, while neglecting explicit phonemic-granularity supervision. This could result in recognition errors due to similar-phoneme confusion or phoneme reduction. To alleviate this problem, we propose a novel framework based on Supervised Contrastive Learning (SCaLa) to enhance phonemic representation learning for end-to-end ASR systems. Specifically, we extend the self-supervised Masked Contrastive Predictive Coding (MCPC) to a fully-supervised setting, where the supervision is applied in the following way. First, SCaLa masks variable-length encoder features according to phoneme boundaries given phoneme forced-alignment extracted from a pre-trained acoustic model; it then predicts the masked features via contrastive learning. The forced-alignment can provide phoneme labels to mitigate the noise introduced by positive-negative pairs in self-supervised MCPC. Experiments on reading and spontaneous speech datasets show that our proposed approach achieves 2.8 and 1.4 points Character Error Rate (CER) absolute reductions compared to the baseline, respectively.

Index Terms: supervised contrastive learning, masked contrastive predictive coding, automatic speech recognition

1. Introduction

In recent years, the accuracy of end-to-end Automatic Speech Recognition (ASR) systems has been significantly improved for various datasets [1–5]. Typically, the models are optimized to improve the average performance over the entire sequence when mapping an input speech to an output character or word sequence, lacking explicit phoneme level supervision. They are powerful enough to learn latent representation partially corresponding to phonemes from each frame [6]. However, the models are still not robust to phonemic issues like similar-phoneme confusion [7] as well as consonant or vowel reduction [8].

Contrastive learning has shown great potential in addressing the phonemic issues of ASR tasks – a variety of masking and contrasting strategies have been proposed to learn speech representations for downstream tasks [9–13]. Recently, most existing contrastive learning based ASR systems assume a self-supervised setting [14]. Masked Contrastive Predictive Coding (MCPC), proposed in Wav2vec2.0 [15], is one of the most representative methods. It masks a consecutive frame of the encoder features with a fixed/random length, then selects anchor and positive samples to obtain positive pairs from the same masked indices of the defined context features and target features, and randomly selects the negative samples from other indices of the target features to obtain negative pairs. The model is then trained to discriminate the anchor/positive features from a set of negative features via a contrastive task. However, applying MCPC to unlabeled data is challenging in mask length selection and noise reduction (caused by negative samples). As shown in Fig. 1: (1) Since phonemes usually have various lengths in speech, masking with a fixed/random length would ignore the boundaries between adjacent phonemes, which may damage the model on learning phonemic representation effectively [8, 16]. (2) In contrastive learning, the indices of negative features for contrasting are randomly selected [15]. Hence, there might exist noisy negative pairs. For example, the anchor/positive-negative pairs may come from the same phoneme, or both of them may be silence or background noise, etc. As referred in [17], noisy negative samples will compromise the effectiveness of the feature representation.



(a) Phoneme-level forced-alignment of a Mandarin speech

(b) Unlike self-supervised MCPC (in purple/regular font), SCaLa (in red/italic font) masks phonemes based on boundary information on encoder features[1], and constructs contrastive feature pairs from context features and target features with mitigation of noisy negative pairs. If the first "q0" in context features is selected, the second "q0" in target features should not be selected as the negative features for contrasting.

Figure 1: *Advantages of SCaLa using phoneme-level forced-alignment (involves labels and boundaries) as supervision over MCPC [15], in terms of masking and contrasting strategies.*

To address the above challenges in self-supervised MCPC, we propose a novel framework named Supervised Contrastive Learning (SCaLa) for end-to-end ASR systems. Unlike previous self-supervised studies, SCaLa applies MCPC in a fully-supervised manner, which has the following two advantages in masking and contrasting strategies, respectively. First, the mask

[1] Masking is for selecting contrastive pairs and is actually performed on encoded features only; the masks on the context features are presented here merely for showing the features being selected.

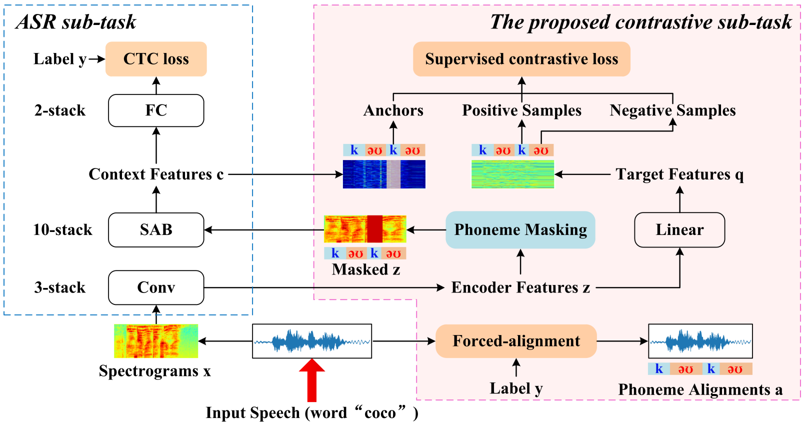


Figure 2: *Model architecture of SCaLa. A CTC-based ASR sub-task is combined with the proposed contrastive sub-task that leverages a forced-alignment model to perform phoneme masking and contrastive learning. The backbone CTC-base ASR network is composed of a successive stack of Convolutional (Conv) layers, Self-Attention Blocks (SABs), and Fully Connected (FC) layers. In the proposed contrastive sub-task, masked items from context features are selected as anchors (associated with phoneme “k”), with the same indices in target features taken as positive samples correspondingly; while items with other indices from target features are selected as negative samples (associated with different phonemes, e.g. “əʊ”).*

length is customized for each particular phoneme duration, i.e., the masked unit can be a complete phoneme. This will help improve the prediction accuracy of reduced consonants or vowels in speech [8]. Specifically, we perform forced-alignment between utterances and their labeled transcription to obtain the phoneme labels and boundaries (shown in Fig. 1(a)). Then the encoder features are masked according to phoneme boundaries to help the model learn phonemic representation explicitly. Second, the noisy anchor/positive-negative pairs are mitigated by selecting the negative features based on the phoneme forced-alignment labels (shown in Fig. 1(b)). Although the alignments may not be perfect, the proposed method can still largely reduce noisy negative pairs empirically, e.g., same-phoneme pairs or silence pairs. Hence the ASR model can learn better latent representation by phoneme discrimination.

To train the ASR model, our SCaLa involves two sub-tasks: (1) an ASR sub-task to directly generate character or word sequences from acoustic features, and (2) a contrastive sub-task to predict masked phonemes to improve phoneme discrimination via contrastive learning. We combine the two sub-tasks to help the model learn the representations of character or word sequences and phoneme level information at the same time, to improve the performance on speech recognition tasks. Our main contributions are: 1) To the best of our knowledge, this is the first work extending the self-supervised MCPC approach to fully-supervised ASR systems; 2) We propose a framework named SCaLa for end-to-end ASR to enhance phoneme-level representation learning; 3) We show the effectiveness of our method, with discussion, on both reading and spontaneous speech data.

The remainder of this paper is organized as follows. Section 2 presents the related work. Section 3 is the details of the proposed method. Section 4 shows the experimental results. Finally, the conclusions and future work are given in Section 5.

2. Related Work

Contrastive learning based on Predictive Coding (PC) has been widely used in self-supervised ASR training. Contrastive PC (CPC) [18] and its variants [19] predict future speech segments based on the past ones for pre-training unidirectional networks. The CPC-based method was adopted for ASR tasks in Wav2vec [20] and Vq-wav2vec to learn discrete representation of speech units [21]. Masked PC (MPC) was proposed in [22], which can improve the performance of Transformer based ASR systems by predicting masked encoder features. In Wav2vec2.0 [15], the authors proposed MCPC, which involved contrastive learning on top of MPC, with significant gain in performance on downstream ASR tasks. However, as mentioned in [23], the self-supervised paradigm of Wav2vec2.0 needs to be carefully designed, and the representation is difficult to interpret. To help the model learn a more meaningful speech representation with MCPC, UniSpeech [23] and JUST [24] were proposed by combining labeled and unlabeled speech for training in a multitask manner; however, labels were still not exploited for MCPC performing. To apply contrastive learning for accented ASR, the authors of [25] adopted SimCLR [26] in the computer vision domain, and then generated contrastive positive pairs from the model's output corresponding to letter-level tokens using various data augmentation methods. Differently, our SCaLa uses forced-alignment results to build contrastive pairs, which ensures these contrastive pairs to be independent from the model training. Masking strategy based on confidence has been introduced into self-supervised training [27]. Besides the supervised setting, our SCaLa concerns more about generating contrastive pairs by masking phonemic-level encoder features to learn phonemic latent representations explicitly.

3. Proposed Method

3.1. Model architecture

The model architecture of SCaLa is shown in Fig. 2, which consists of an ASR sub-task and the proposed contrastive sub-task.

As for the ASR sub-task, the representative Connectionist Temporal Classification (CTC) framework used in ASR training [15, 20, 23] is adopted as our backbone model. Our experiments use the typical model network, which is composed of a successive stack of Convolutional (Conv) layers, Self-Attention Blocks (SABs), and Fully Connected (FC) layers [28]. Given a sequence of d_s-dimensional acoustic spectrograms $\mathbf{x} \in \mathbf{R}^{d_s \times T}$ with length T, the ASR model tries to predict the labeled character sequence $\mathbf{y} \in \mathbf{L}^N$ with length N, where $\mathbf{L}$ is the size of the finite label character. The output of the last Conv layer is denoted as encoder features $\mathbf{z} \in \mathbf{R}^{d_f \times S}$, where d_f is the latent feature dimension, and S is the sequence length. Note that S is less than T after subsampling by the Conv layers [29]. The phoneme-level forced-alignment results are denoted by $\mathbf{a} \in \mathbf{R}^S$ [30]. Merely for brevity, we assume that each item in $\mathbf{a}$ is a phoneme label of items in the encoder features $\mathbf{z}$ instead of the input speech. The items in $\mathbf{z}$ are masked with some prescribed probability during training. The masked features are fed into SABs that yields context features $\mathbf{c} \in \mathbf{R}^{d_f \times S}$.

Regarding the proposed contrastive sub-task, a linear layer is adopted like in [31, 32] to obtain the contrastive target features $\mathbf{q} \in \mathbf{R}^{d_f \times S}$. Then we select masked items from context features as the anchors [15]. Items in target features with the same indices are taken as the positive samples, while items in target features with indices from other alignment labels are taken as negative samples. For the contrastive sub-task, a contrastive loss is involved (in our loss function described next) to guide the model to decrease the similarities between the anchor-negative pairs, and to increase the similarities between the anchor-positive pairs.

3.2. Loss functions

We combine the ASR sub-task and the contrastive sub-task to help the model learn the representation of character sequences and phoneme-level features simultaneously, and to improve the performance on speech recognition tasks. The two losses of SCaLa corresponding to the two sub-tasks are (1) a CTC loss based on phoneme masking L_{CTC} and (2) a supervised contrastive loss with phonemic-granularity supervision L_{SCL}.

3.2.1. CTC loss based on phoneme masking

Phoneme labels and boundaries are used for masking to help the model enhance phonemic representation [8]. Specifically, a HMM-DNN acoustic model is trained offline using Kaldi [33] to get the phoneme-level label of each frame and the corresponding boundaries. An example is shown in Fig. 1(a). During the training, we randomly sample a set of start indices of encoder features with a certain probability p_e (We set $p_e = 6.5\%$ like in [15]). A total of P phonemes adjacent to the start indices are integrally masked by leveraging the phoneme boundaries given forced-alignment. The choice of P is experimentally studied in Sec. 4.2 (see Fig. 3). For a data-label pair $\{\mathbf{x}, \mathbf{y}\}$ and the forced-alignment result $\mathbf{a}$, the CTC loss based on phoneme masking is obtained as

$$L_{\text{CTC}} = -\log \sum_{\boldsymbol{\pi} \in \boldsymbol{\phi}(\mathbf{x},\mathbf{y})} p(\boldsymbol{\pi} | \mathbf{x}, \mathbf{a}, p_e) \qquad (1)$$

where a valid CTC path $\boldsymbol{\pi}$ is a variant of the transcription $\mathbf{y}$ that allows occurrences of blank tokens and repetitions. The set $\boldsymbol{\phi}(\mathbf{x}, \mathbf{y})$ includes all valid CTC paths [34].

3.2.2. Contrastive loss with phonemic-granularity supervision

Supervised contrastive learning [17] aims to improve the robustness of feature representation by discriminating an anchor/positive phoneme from a set of negative phonemes. In particular, to reduce noisy negative pairs, features having the same phoneme label with the masked phoneme are avoided from the negative phoneme sets. As shown in Fig. 1(b), the second "q0" of target features will not be selected as negative phonemes to be contrasted with the first "q0" of context features. Accordingly, the supervised contrastive loss is defined as

$$L_{\text{SCL}} = -\frac{1}{|\boldsymbol{M}|} \sum_{m \in \boldsymbol{M}} \log \frac{e^{sim(\mathbf{c}_m, \mathbf{q}_m)/\tau}}{\sum_{n \in \boldsymbol{N_m}} e^{sim(\mathbf{c}_m, \mathbf{q}_n)/\tau}} \qquad (2)$$

where $\boldsymbol{M}$ is the set of all masked indices of encoder features, and $|\boldsymbol{M}|$ is the number of masked indices; $\mathbf{c}_m$ and $\mathbf{q}_m$ are the m^{th} vectors in context features $\mathbf{c}$ and target features $\mathbf{q}$, respectively; $sim(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{\alpha}^T \boldsymbol{\beta} / (||\boldsymbol{\alpha}|| \ ||\boldsymbol{\beta}||)$ is the cosine similarity; τ is a temperature scale. The index set $\boldsymbol{N_m}$ consists of the masked index m and a negative index set $\boldsymbol{K}$, which are uniformly sampled from all indices except those having the same alignment label with the masked phoneme $\mathbf{a}_m$, i.e. $\mathbf{a}_k \neq \mathbf{a}_m$, $\forall k \in \boldsymbol{K}$. We set $\tau = 0.1$ and the number of negative indices $|\boldsymbol{K}| = 100$ in our experiments – the same as [15].

3.3. Model training

Following [35], an alternate minimization training method is employed in our proposed method as well. The training losses L_{CTC} and L_{SCL} are minimized alternately with a balanced ratio, i.e. 1:1, to update the model parameters. The main advantage of alternating training is that the learning rates of ASR sub-task optimizer and contrastive sub-task optimizer are separated [35]. In our experiments, we find that a single optimizer resulting from a weighted sum of the two loss functions would result in a slightly higher CER, and the ratio does not significantly influence the performance of our method.

4. Experiments and Discussion

4.1. Experimental setup

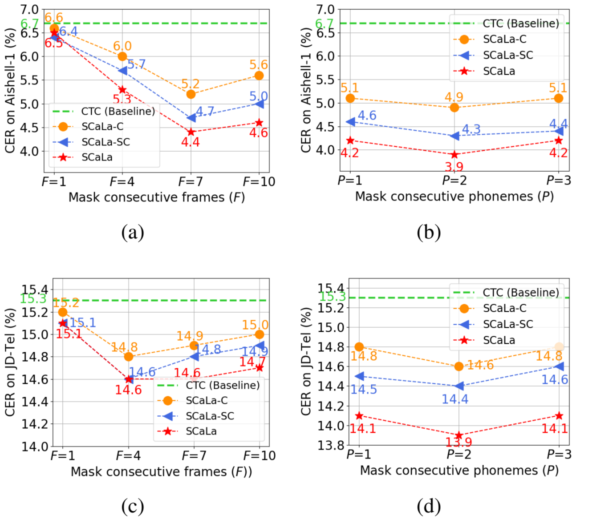
Two datasets with different speaking styles are used in our experiments: 1) reading speech data: the open-source Aishell-1 which contains 170 hours of Mandarin speech with 16kHz sampling rate [36]; and 2) spontaneous speech data: an in-house Mandarin conversational Telephony (JD-Tel) dataset which contains 1500 hours of speech with 8kHz sampling rate. For Aishell-1, the original train-test split is used. For the other one, 10% of the samples are randomly selected for testing.

The 80-dimensional Mel-spectrograms are used as the input to the network. The frame size and step size are 20ms and 10ms, respectively. Our model contains 3 Conv layers, 10 SABs, and 2 FCs, as shown in Fig. 2. Please refer to [28] for more details about the model.

Our models are trained on 4 NVIDIA V100 GPUs with mini-batch size 128. For the alternate loss minimization in Sec. 3.3, the Learning Rate (LR) for L_{CTC} is 2.5×10^{-5}, and 5×10^{-4} for L_{SCL}. The LR of L_{CTC} is unchanged, while the LR of L_{SCL} is decayed to 5×10^{-5} ultimately.

Table 1: *System performance in CER (%).*

Testing data	**Aishell-1**	**JD-Tel**
Speaking style	**reading**	**spontaneous**
Chain model [37]	7.5	15.9
Wav2vec2.0 [15, 39]	5.3	16.3
WeNet CTC-conformer [38]		
w/ CTC prefix beam search	5.9	15.6
w/ attention rescoring	5.3	14.7
CTC (Baseline) [28]	6.7	15.3
CTC+phoneme mask [8]	5.1	14.8
SCaLa	**3.9**	**13.9**



(a) (b)

(c) (d)

Figure 3: *Ablation study where CERs for SCaLa compared with: 1) SCaLa using random samplings for the contrastive sub-task without supervision (SCaLa-SC), and 2) SCaLa without the contrastive sub-task (SCaLa-C), for a variety of mask settings on reading and spontaneous speech data. The performance of the traditional CTC method (Baseline) is included for reference.*

4.2. Main experiment and ablation study

As shown in Table 1, we compare SCaLa with state-of-the-art ASR systems including hybrid (i.e. the chain model, a type of improved DNN-HMM) [37], end-to-end [8, 28, 38], and self-supervised learning [15,39]. Experimental results show that existing end-to-end systems [8, 28, 38] and self-supervised learning [15, 39] largely outperform the chain model that based on forced-alignment [37]. The performance of CTC models [28] can also benefit from phoneme mask strategies [8]. Compared with the existing methods, our SCaLa achieves the best performance. Numerically, it outperforms the traditional CTC models [28] with 2.8 and 1.4 points CER absolute reductions on reading and spontaneous speech data, respectively.

As ablation studies on phoneme masking and contrastive learning: (1) We compare two masking methods: masking F consecutive frames (fixed-length masks in [15]) and masking P consecutive phonemes (phoneme masks in SCaLa) where $F \in \{1, 4, 7, 10\}$ and $P \in \{1, 2, 3\}$. In the experiments, the forced-alignment results show that the average phoneme lengths are 3.3 and 2.6 frames for the reading and spontaneous speech data, respectively. (2) Different contrastive strategies including the proposed SCaLa, SCaLa using random samplings for the contrastive sub-task without supervision (SCaLa-SC), and SCaLa without the contrastive sub-task (SCaLa-C) are also compared. The results for these ablation studies are shown in Fig. 3. The CERs curves of different methods indicate that supervised contrastive learning can significantly reduce CERs. Moreover, the results of different masking strategies on both sides of the figures show that phoneme masking outperforms the fixed-length masking in ASR tasks. The results also indicate that our SCaLa achieved the best performance when $P = 2$. We infer that the optimal value of P is related to the fact that a character in Mandarin usually contains two phonemes [40]. Larger mask length creates a heavy burden on the model learning because there is an overly large variety in the masked contents, while smaller ones influence the training very little [15].

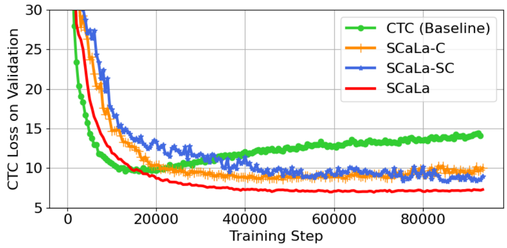


Figure 4: *Regularization effect of SCaLa with $P = 2$ on Aishell-1 (Each training epoch contains 936 steps in the experiments).*

4.3. Analysis of SCaLa

To further evaluate the effectiveness of SCaLa, we analyze the proposed method from the following three perspectives:
Robustness to phonemic issues. The phonemic issues, such as similar-phoneme confusion and phoneme-reduction usually introduce substitution and deletion errors [8]. The detailed CER, including substitution (SUB), deletion (DEL) and insert (INS) error rates, of SCaLa and the baseline CTC method are shown in Table 2. The results show that SCaLa achieves significant reductions on substitution and deletion errors. We also observe that the performance improvement of SCaLa on reading speech data is more than that on spontaneous speech data. The reasons may be that recognition tasks on spontaneous speech are more challenging, and that the given forced-alignment results may not be accurate enough. Nevertheless, our method can still improve the performance of speech recognition.

Table 2: *Detailed CER (%), including substitution (SUB), deletion (DEL) and insert (INS) error rates, for SCaLa and the baseline CTC method on reading (Aishell-1) and spontaneous (JD-Tel) speech data.*

Testing data	Methods	SUB	DEL	INS	CER
Aishell-1	CTC (Baseline) [28]	5.2	1.4	**0.1**	6.7
	SCaLa-SC	3.8	0.4	**0.1**	4.3
	SCaLa	**3.5**	**0.3**	**0.1**	**3.9**
JD-Tel	CTC (Baseline) [28]	10.2	4.7	0.4	15.3
	SCaLa-SC	9.9	4.2	**0.3**	14.4
	SCaLa	**9.7**	**3.9**	**0.3**	**13.9**

Noisy negative reduction. We use the forced-alignment results to count the proportion of noisy negative pairs among all the negative pairs for contrasting. We find that the noisy negative rates of self-supervised MCPC are 10.21% and 14.60% for the two datasets, respectively. The number is non-negligible [17]. As shown in Table 2, SCaLa improves system performance compared to SCaLa-SC by reducing the noisy negatives.
Regularization effect. Fig. 4 shows the CTC losses of SCaLa, SCaLa-SC, SCaLa-C and the baseline CTC method on the validation data of Aishell-1. SCaLa obtains the lowest and smoothest loss curve compared with the other three methods, which indicates the regularization effect to the training.

5. Conclusion

In this paper, a novel framework named SCaLa has been proposed for ASR training. It extends the self-supervised MCPC approach to a fully-supervised setting. The labels are effectively leveraged to enhance ASR models to learn phoneme representation. SCaLa significantly improved the performance on both reading and spontaneous Mandarin speech data compared to the baseline methods. In the future work, the performance on more other masking strategies and languages will be evaluated.

6. References

[1] W. Chan, N. Jaitly, Q. Le, et al., "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. ICASSP*, 2016.

[2] D. Bahdanau, J. Chorowski, D. Serdyuk, et al., "End-to-end attention-based large vocabulary speech recognition," in *Proc. ICASSP*, 2016.

[3] R. Prabhavalkar, K. Rao, T. Sainath, et al., "A comparison of sequence-to-sequence models for speech recognition," in *Proc. Interspeech*, 2017.

[4] J. Li, Y. Wu, Y. Gaur, et al., "On the comparison of popular end-to-end models for large scale speech recognition," in *Proc. Interspeech*, 2020.

[5] J. Li, "Recent advances in end-to-end automatic speech recognition," *arXiv preprint arXiv:2111.01690*, 2021.

[6] Y. Belinkov, and J. Glass, "Analyzing hidden representations in end-to-end automatic speech recognition systems," in *Proc. NeurIPS*, 2017.

[7] A. Fang, S. Filice, N. Limsopatham, et al., "Using phoneme representations to build predictive models robust to ASR errors," in *Proc. SIGIR*, 2020.

[8] G. Ma, P. Hu, J. Kang, et al., "Leveraging phone mask training for phonetic-reduction-robust end-to-end Uyghur speech recognition," in *Proc. Interspeech*, 2021.

[9] A. Baevski, and A. Mohamed, "Effectiveness of self-supervised pre-training for ASR," in *Proc. ICASSP*, 2020.

[10] M. Ravanelli, J. Zhong, S. Pascual, et al., "Multi-task self-supervised learning for robust speech recognition," in *Proc. ICASSP*, 2020.

[11] A. Liu, S. Yang, P. Chi, et al., "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *Proc. ICASSP*, 2020.

[12] W. Hsu, B. Bolte, Y. Tsai, et al., "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 29:3451-3460, 2021.

[13] A. Pasad, J. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *Proc. ASRU*, 2021.

[14] S. Liu, A. Mallol-Ragolta, E. Parada-Cabeleiro, et al., "Audio self-supervised learning: A survey," *arXiv preprint arXiv:2203.01205*, 2022.

[15] A. Baevski, Y. Zhou, A. Mohamed, et al., "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020.

[16] C. Wang, Y. Wu, Y. Du, et al., "Semantic mask for transformer based end-to-end speech recognition," in *Proc. Interspeech*, 2019.

[17] P. Khosla, P. Teterwak, C. Wang, et al., "Supervised contrastive learning," in *Proc. NeurIPS*, 2020.

[18] A. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[19] Y. Chung, W. Hsu, H. Tang, et al., "An unsupervised autoregressive model for speech representation learning," in *Proc. Interspeech*, 2019.

[20] S. Schneider, A. Baevski, R. Collobert, et al., "Wav2vec: Unsupervised pre-training for speech recognition," in *Proc. Interspeech*, 2019.

[21] A. Baevski, S. Schneider, and M. Auli, "Vq-wav2vec: Self-supervised learning of discrete speech representations," in *Proc. ICLR*, 2020.

[22] D. Jiang, X. Lei, W. Li, et al., "Improving transformer-based speech recognition using unsupervised pre-training," *arXiv preprint arXiv:1910.09932*, 2019.

[23] C. Wang, Y. Wu, Y. Qian, et al., "UniSpeech: Unified speech representation learning with labeled and unlabeled data," in *Proc. ICML*, 2021.

[24] J. Bai, B. Li, Y. Zhang, et al., "Joint unsupervised and supervised training for multilingual asr," in *Proc. ICASSP*, 2022.

[25] T. Han, H. Huang, Z. Yang, et al., "Supervised contrastive learning for accented speech recognition," *arXiv preprint arXiv:2107.00921*, 2021.

[26] T. Chen, S. Kornblith, M. Norouzi, et al., "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, 2020.

[27] M. Baskar, A. Rosenberg, B. Ramabhadran, et al., "Ask2Mask: Guided data selection for masked speech modeling," *arXiv preprint arXiv:2202.12719*, 2022.

[28] L. Fu, X. Li, L. Zi, et al., "Incremental learning for end-to-end automatic speech recognition," in *Proc. ASRU*, 2021.

[29] S. Amodei, R. Ananthanarayanan, J. Anubhai, et al., "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proc. ICML*, 2016.

[30] K. Gorman, J. Howell, and M. Wagner, "Prosodylab-aligner: A tool for forced alignment of laboratory speech," *Canadian Acoustics*, 39(3):192-193, 2011.

[31] Y. Zhang, J. Qin, D. Park, et al., "Pushing the limits of semi-supervised learning for automatic speech recognition," in *Proc. NeurIPS*, 2020.

[32] J. Bai, B. Li, Y. Zhang, et al., "Joint unsupervised and supervised training for multilingual ASR, *arXiv preprint arXiv:2111.08137*, 2021.

[33] D. Povey, A. Ghoshal, G. Boulianne, et al., "The kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[34] A. Graves, S. Fernandez, F. Gomez, et al., "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006.

[35] C. Talnikar, T. Likhomanenko, R. Collobert, et al., "Joint masked CPC and CTC training for ASR," in *Proc. ICASSP*, 2021.

[36] H. Bu, J. Du, X. Na, et al., "AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline," in *Proc. O-COCOSDA*, 2017.

[37] F. Yu, and K. Chen, "Non-autoregressive transformer-based end-to-end asr using bert," in *arXiv preprint arXiv:2104.04805*, 2021.

[38] Z. Yao, D. Wu, X. Wang, et al., "WeNet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," in *Proc. Interspeech*, 2021.

[39] J. Yuan, X. Cai, D. Gao, et al., "Decoupling recognition and transcription in Mandarin ASR," *arXiv preprint arXiv:2108.01129*, 2021.

[40] M. Huang, Y. Lu, L. Wang, et al., "Exploring model units and training strategies for end-to-end speech recognition," in *Proc. ASRU*, 2019.