SCALA: SUPERVISED CONTRASTIVE LEARNING FOR END-TO-END AUTOMATIC SPEECH RECOGNITION

Li Fu, Xiaoxiao Li, Runyu Wang, Zhengchen Zhang, Youzheng Wu, Xiaodong He, Bowen Zhou
JD AI Research, Beijing, China

ABSTRACT

End-to-end Automatic Speech Recognition (ASR) models are usually trained to reduce the losses of the whole token sequences, while neglecting explicit phonemic-granularity supervision. This could lead to recognition errors due to similar-phoneme confusion or phoneme reduction. To alleviate this problem, this paper proposes a novel framework of Supervised Contrastive Learning (SCaLa) to enhance phonemic information learning for end-to-end ASR systems. Specifically, we introduce the self-supervised Masked Contrastive Predictive Coding (MCPC) into the fully-supervised setting. To supervise phoneme learning explicitly, SCaLa first masks the variable-length encoder features corresponding to phonemes given phoneme forced-alignment extracted from a pre-trained acoustic model, and then predicts the masked phonemes via contrastive learning. The phoneme forced-alignment can mitigate the noise of positive-negative pairs in self-supervised MCPC. Experimental results conducted on reading and spontaneous speech datasets show that the proposed approach achieves 2.84% and 1.38% Character Error Rate (CER) reductions compared to the baseline, respectively.

Index Terms— Supervised contrastive learning, masked contrastive predictive coding, automatic speech recognition

1. INTRODUCTION

End-to-end Automatic Speech Recognition (ASR) systems have been significantly improved in recent years [1, 2]. They are trained to map acoustic feature sequences to token sequences directly. The training normally aims to minimize the losses of the whole character or word sequences, while neglecting explicit phoneme level supervision. The models are powerful enough to learn latent representation partially corresponding to phonemes from each frame [3]. However, these models still suffer from various phonemic issues like similar-phoneme confusion [4] as well as constant or vowel reduction [5]. Inspired by recent development of contrastive learning in speech representation [6–9], we propose a novel framework named Supervised Contrastive Learning (SCaLa) for end-to-end ASR systems. Contrastive learning with guidance of phoneme labels is conducted to enhance ASR models to learn phoneme information.

Masked Contrastive Predictive Coding (MCPC) [10] is a representative method of contrastive learning used in self-supervised ASR systems. It masks a consecutive segment of the encoder features with a fixed/random length, and then selects anchor/positive and negative samples from masked and unmasked indices, respectively. The model is then trained to discriminate the anchor/positive features from a set of negative features via a contrastive task. Self-supervised MCPC has been proved to be able to learn the latent representation of speech units in the pre-training stage [10]. It achieves a remarkable performance via fine-tuning on small amounts of labeled data. Extending self-supervised MCPC to supervised ASR systems may improve the recognition performance. UniSpeech [11] is a first try to combine self-supervised MCPC and supervised ASR systems. However, it follows the pre-training and fine-tuning paradigm, and MCPC is still employed to learn speech representation without using labelling information. Applying MCPC to unlabeled data is challenging in both mask length selection and mitigation of the effect from noisy negative samples. Since phonemes usually have various lengths in speech, masking with a fixed/random length may not help the model learn phonemic information effectively [5, 12]. In contrastive learning, the indices of negative features for contrasting are randomly selected [10]. Hence, there might be noisy negative pairs. For example, the anchor/positive-negative pairs might be from the same phoneme, or both of them may be silence or background noise, etc. As referred in [13], noisy negative samples will compromise the effectiveness of the feature representation.

Unlike previous studies [10, 11], SCaLa applies MCPC in a fully-supervised manner. The above two challenges are addressed by using the phoneme label information. First, the mask length is customized for each particular phoneme duration, i.e., the masked unit can be a phoneme. This will help improve the prediction accuracy of reduced phonemes in speech [5]. Specifically, we perform forced-alignment between the speech and the labeled transcription to obtain the phoneme duration and labels. Then the encoder features are masked in phoneme level to enhance the model to learn phonemic latent representations explicitly. Second, the noisy anchor/positive-negative pairs are eliminated by selecting the negative features based on the phoneme forced-alignment labels. Although the alignments might not be perfect, one can still largely reduce noisy negative pairs, 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, SCaLa includes two sub-tasks: an ASR sub-task is to directly generate character or word sequences from acoustic features, and a contrastive sub-task is to predict masked phonemes for improving 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.

The main contributions are: 1) To the best of our knowledge, this is the first work extending the self-supervised MCPC approach to the fully-supervised ASR systems; 2) We propose a framework named SCaLa for end-to-end ASR to enhance phoneme-level information learning; 3) We show the effectiveness of our method with discussion on both reading and spontaneous speech data.

2. RELATED WORK

Contrastive learning based on Predictive coding (PC) has been widely used in self-supervised ASR training. Contrastive PC (CPC) [14] and its variants [15] 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 [16] and Vq-wav2vec to learn discrete representation of speech units [17]. Masked PC (MPC) was proposed in [18], which can improve the performance of Transformer based ASR systems by predicting masked encoder features. In Wav2vec2.0 [10], the authors proposed MCPC, which involved contrastive learning compared to MPC. Significant performance improvements are achieved on downstream ASR tasks. However, as mentioned in [11], 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 combined labeled and unlabeled speech for pre-training in a multitask manner [11]. However, MCPC was still performed without using labels in [11].

Research on contrastive learning for supervised ASR training is relatively new. The motivation of [9] is very similar to our work. Nevertheless, these two works are quite different. First, [9] adopted the SimCLR [19] method in the computer vision domain, while SCaLa utilizes the self-supervised MCPC for ASR. Also, new speech is generated using various data augmentation methods first in [9]. These new speech and the original speech are fed into an ASR framework to output representation corresponding to letter-level tokens. The representations of the same letter are taken as the positive pairs for contrasting. There is a potential problem that the output of the ASR framework may represent a wrong letter. In this case, consequently, the contrastive pairs would be wrong, and such pairs would mislead the model training by itself. Differently, SCaLa uses the forced-alignment results to build contrastive pairs, which ensures the contrastive pairs are independent of the model training. Moreover, we generate the contrastive pairs by masking encoder features at the phonemic level.

Masking phonemes in the data augmentation domain has been introduced into supervised ASR training. The authors of [5] masked off certain phoneme features during model training to simulate the phoneme reduction phenomenon in Uyghur speech. With this method, the ASR model performed better on spontaneous speech. SCaLa employs the same masking strategy, and adds a new contrastive learning loss. The loss is proposed to improve the robustness of representation to phoneme prediction and discrimination.

3. PROPOSED METHOD

3.1. Model architecture

The model architecture of the proposed SCaLa is shown in Fig. 1, which consists of an ASR sub-task and a contrastive sub-task. The representative Connectionist Temporal Classification (CTC) framework used in ASR training [10, 11, 16] 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 [20]. Given a sequence of $d_x$-dimensional acoustic spectrograms $x \in \mathbb{R}^{d_x \times T}$ with length $T$, the ASR model tries to predict the labeled character sequence $y \in \mathbb{L}^N$ with length $N$, where $\mathbb{L}$ is the size of the finite label character. The output of the last Conv layer is denoted as encoder features $z \in \mathbb{R}^{d_f \times S}$, where $d_f$ is the latent feature dimension, and $S$ is the sequence length. Note that $S$ is less than $T$ after subsampled by the Conv layers. The phoneme-level forced-alignment results are denoted by $a \in \mathbb{R}^S$ [21]. To make the expression concise, we assume that each item in $a$ is a phoneme label of items in the encoder features $z$ instead of the input speech. The items in $z$ are masked with a certain probability during training. The masked features are fed into SABs that yields context features $c \in \mathbb{R}^{d_f \times S}$. Regarding the contrastive sub-task, a linear layer is adopted as done in [22] to obtain the contrastive target features $q \in \mathbb{R}^{d_f \times S}$. Then we select masked items from context features as anchors [10], take items in target features with same indices as anchors as positive samples, and select legal items with unmasked indices from target features as negative samples.

![Fig. 1. Model architecture of SCaLa. A CTC-based ASR sub-task is combined with a 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. The contrastive sub-task selects masked items from context features as anchors, takes items in target features with same indices as anchors as positive samples, and selects legal items with unmasked indices from target features as negative samples.](image-url)
3.2. Loss functions

To train the ASR model, we combine the ASR sub-task and the contrastive sub-task to help the model learn the representation of character sequences and phoneme-level information at the same time, and to improve the performance on speech recognition tasks. The total loss of SCaLa is a combination of a CTC loss based on phoneme masking $L_{CTC}$ and a contrastive loss $L_C$, as follows:

$$L_{total} = \lambda L_{CTC} + (1 - \lambda) L_C$$  \hspace{1cm} (1)

where hyperparameter $\lambda$ is used to balance these terms.

3.2.1. CTC loss based on phoneme masking

Phoneme lengths are used for masking to help the model learn more phonemic information [5]. Specifically, a HMM-DNN acoustic model is trained offline using Kaldi [23] to get the phoneme-level label of each frame. An example is shown in Fig. 2. During the training, we randomly sample start indices of encoder features with a certain probability $p_n$ (We set $p_n = 7.5\%$ same as [10]). A total of $P$ phonemes adjacent to the start indices are integrally masked by leveraging the forced-alignment choices. The choice of $P$ is experimentally studied in Sec. 4.2 (see Fig. 3). For a data-label pair $(x, y)$ and the forced-alignment result $a$, the CTC loss based on phoneme masking is obtained as

$$L_{CTC} = - \log \sum_{\pi \in \phi(x, y)} p(\pi | x, a, p_c)$$  \hspace{1cm} (2)

where a valid CTC path $\pi$ is a variant of the transcription $y$ that allows occurrences of blank tokens and repetitions. The set $\phi(x, y)$ includes all valid CTC paths [24].

3.2.2. Contrastive loss

Supervised contrastive learning 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 removed from the negative phoneme sets. As shown in Fig. 2, the second “q0” of target features will not be selected as negative phonemes to be contrasted with the first “q0” of context features. Then the contrastive loss is defined as

$$L_C = - \frac{1}{|M|} \sum_{m \in M} \log \frac{\exp(\text{sim}(c_m, q_m) / \tau)}{\sum_{n \in N_m} \exp(\text{sim}(c_m, q_n) / \tau)}$$ \hspace{1cm} (3)

where $M$ is the set of all masked indices of encoder features, and $|M|$ is the number of masked indices; $c_m$ and $q_m$ are the $m$th vectors in context features $c$ and target features $q$, respectively; sim$(\alpha, \beta) = \alpha^T \beta / (||\alpha|| ||\beta||)$ is the cosine similarity; $\tau$ is a temperature SCaLa. The index set $N_m$ consists of the masked index $m$ and a negative index set $K$, which are uniformly sampled from all indices except those having the same alignment label with the masked phoneme $a_m$, i.e. $a_k \neq a_m, \forall \in K$. We set $\tau = 0.1$ and $|K| = 100$ in our experiments same as [10].

| Testing data | Aishell-1 | JD-tel |
|---------------|-----------|--------|
| Speaking style | reading | spontaneous |
| Chain model [28] | 7.45 | 15.94 |
| Wav2vec2.0 [10, 30] | 5.30 | 16.30 |
| WeNet CTC-conformer [29] | 5.91 | 15.62 |
| w/ CTC prefix beam search | 5.30 | 14.74 |
| w/ attention rescoring | 6.74 | 15.34 |
| CTC [20] | 5.11 | 14.80 |
| CTC+phoneme mask [5] | 3.90 | 13.96 |

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 [25]; and 2) spontaneous speech data: a JD in-house Mandarin conversational telephony dataset (JD-tel) which contains 1500 hours of speech with 8kHz sampling rate. For Aishell-1, the original train-test splits are used. For the other one, 10% of the speech 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. There are 3 Conv layers, 10 SABs, and 2 FCs in our model as shown in Fig. 1. For more model details refer to [20]. A 4-gram language model trained with the KenLM toolkit [26] is used for inference.

Same with [27], the alternate minimization training method is employed. The training losses $L_{CTC}$ and $L_C$ in Eq. (1) are used

1The results on Aishell-1 are obtained from published papers except [5]. The other results of the methods are re-implemented by our self.
4.2. Experimental Results

As shown in Table 1, we compare SCaLa with state-of-the-art ASR systems including hybrid [28], end-to-end [5, 29], and self-supervised learning [10, 30]. Numerically, SCaLa outperforms the traditional CTC models [20] with 2.84% and 1.38% CER reductions on reading and spontaneous speech data, respectively. The detailed analysis of our SCaLa is presented in Sec. 4.3. Experimental results also show that SCaLa significantly outperforms hybrid chain models [28], end-to-end CTC-Conformer systems [29], self-supervised systems learning [10, 30], and methods with phoneme masking [5].

Parameter searching and ablation study are conducted. We compare two masking methods: masking $F$ consecutive frames and masking $P$ consecutive phonemes where $F \in \{1, 4, 7, 10\}$ and $P \in \{1, 2, 3\}$. Different contrastive strategies including SCaLa, SCaLa without supervision in the contrastive sub-task (SCaLa-SC), and SCaLa without the contrastive sub-task (SCaLa-C) are also compared. The results are shown in Fig. 3. The baseline method in Fig. 3 is the traditional CTC method. One can see that both masking with phoneme information and supervised contrastive learning can improve system performance significantly. SCaLa achieved the best performance when $P = 2$.

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. 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. This might be that recognition tasks on spontaneous speech are more challenging, and the given forced-alignment results might not be accurate enough. Nevertheless, our method can still improve the performance of speech recognition.

Noisy negative reduction. The forced-alignment results show that the noisy negative rates of self-supervised MCPC are 10.21% and 14.60% in the two data sets, respectively. The number is non-negligible [13]. As shown in Fig. 3, SCaLa improves system performance compared to SCaLa-SC. SCaLa achieves significant reductions on substitution and deletion errors. The detailed CER, including substitution (SUB), deletion (DEL) and insert (INS) error rates, of SCaLa and the baseline CTC method on reading (Aishell-1) and spontaneous (JD-tel) speech data are shown in Table 2. The results show that SCaLa achieves significant reductions on substitution and deletion errors compared to the baseline CTC method.

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 shows 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 the fully-supervised setting. The labels are effectively leveraged to enhance ASR models to learn phoneme information. SCaLa significantly improved the performance on both reading and spontaneous Mandarin speech data compared to the baseline methods. In the future, more other masking strategies will be investigated, and the performance on other languages will be evaluated.
6. REFERENCES

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

[2] J. Li, Y. Wu, Y. Gaur, et al., “On the comparison of popular end-to-end models for large SCoLa speech recognition,” in Proc. Interspeech, 2020.

[3] Y. Belinkov, and J. Glass, “Analyzing hidden representations in end-to-end automatic speech recognition systems,” in Proc. NeurIPS, 2017.

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

[5] 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.

[6] A. Baevski, and A. Mohamed, “Effectiveness of self-supervised pre-training for ASR,” in Proc. ICASSP, 2020.

[7] M. Ravanelli, J. Zhong, S. Pascual, et al., “Multi-task self-supervised learning for robust speech recognition,” in Proc. NeurIPS, 2020.

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

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

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

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

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

[13] P. Khosla, P. Teterwak, C. Wang, et al., “Supervised contrastive learning,” in Proc. NeurIPS, 2020.

[14] A. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” arXiv preprint arXiv:1807.03748, 2018.

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

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

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

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

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

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

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

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

[23] D. Povey, A. Ghoshal, G. Boulianne, et al., “The kaldi speech recognition toolkit,” in Proc. ASRU, 2011.

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

[25] 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.

[26] K. Heafield, “KenLM: Faster and smaller language model queries,” in Proc. SMT, 2011.

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

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

[29] 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.

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