electronics

MDPI

Article

A Self-Supervised Model for Language Identification Integrating Phonological Knowledge

Qingran Zhan [1], Xiang Xie [1,2,*], Chenguang Hu [1] and Haobo Cheng [1,2]

[1] School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China; qingran.zhan@gmail.com (Q.Z.); 3220200551@bit.edu.cn (C.H.); chb@bit.edu.cn (H.C.)
[2] Shenzhen Research Institute, Beijing Institute of Technology, Shenzhen 518000, China
* Correspondence: xiexiang@bit.edu.cn

Abstract: In this paper, a self-supervised learning pre-trained model is proposed and successfully applied in language identification task (LID). A Transformer encoder is employed and multi-task strategy is used to train the self-supervised model: the first task is to reconstruct the masking spans of input frames and the second task is a supervision task where the phoneme and phonological labels are used with Connectionist Temporal Classification (CTC) loss. By using this multi-task learning loss, the model is expected to capture high-level speech representation in phonological space. Meanwhile, an adaptive loss is also applied for multi-task learning to balance the weight between different tasks. After the pretraining stage, the self-supervised model is used for xvector systems. Our LID experiments are carried out on the oriental language recognition (OLR) challenge data corpus and *1 s*, *3 s*, *Full-length* test sets are selected. Experimental results show that on *1 s* test set, feature extraction model approach can get best performance and in *3 s*, *Full-length* test, the fine-tuning approach can reach the best performance. Furthermore, our results prove that the multi-task training strategy is effective and the proposed model can get the best performance.

Keywords: self-supervised learning; phonological knowledge; language identification

Citation: Zhan, Q.; Xie, X.; Hu, C.; Cheng, H. A Self-Supervised Model for Language Identification Integrating Phonological Knowledge. *Electronics* **2021**, *10*, 2259. https://doi.org/10.3390/electronics10182259

Academic Editor: Daniel Hládek

Received: 12 August 2021
Accepted: 7 September 2021
Published: 14 September 2021

1. Introduction

Recently, the self-supervised training has shown to be effective for improving downstream systems [1–4]. Speech signal contains a rich set of acoustic and linguistic information, including phonemes, words, articulatory and even sentiment information. Through self-supervised pre-training, high-level speech representation can be captured from raw speech [1,5]. The learned models could be applied to downstream speech and language processing tasks through feature-based speech representation extraction.

In this work, we propose a self-supervised based pre-trained model where the phonological labels are used as an auxiliary objective. In this model, two objectives are used. First, like most of the self-supervised models do, the masking strategies are applied on the input frames and the L1 Loss is used to minimize reconstruction error between prediction and ground-truth frames. Second, to make the model learn the speech representation in phonological space, we apply the CTC with phoneme and phonological labels. After the pre-training stage, xvectors system is incorporated with pre-trained self-supervised model for LID. The framework is described as Figure 1.

During the self-supervised model training, the input acoustic frames are randomly masked on time and channel axis, the model learns to reconstruct and predict the original frames. In neutral network models, a contrastive loss can induce high-level latent knowledge [5] so the sequence-level CTC loss is used with phoneme and phonological labels for phonological representation learning here. Language identification is very important in our real life communication, for both text LID [6,7] and speech LID [8,9]. In speech LID, phonetic knowledge is often used to improve the system performance. The most

traditional approach is to incorporate the deep bottleneck features (DBF) with LID model, where the DBF is extracted from a well-trained hybrid automatic speech recognition (ASR) system. In our proposed model, we not only use phonetic information, phonological knowledge is also introduced to make the model learn phonological representation. A lot of researches have proved that the phonological knowledge can be shared across different languages by using statistical model but most of the previous works are using "acoustic-to-articulatory(-attribute) modeling [10,11]. When the self-supervised model reconstructs the masked frames, the CTC loss works as a regularization with phonological knowledge at the same time, thus the model can capture high-level representation at phonological space. To balance the different losses, we apply a principled approach to multi-task deep learning which weighs multiple loss functions by considering the homoscedastic uncertainty of each task. Because of the jointly training, the model can learn both acoustic and phonological representation. By incorporating the pre-trained model, the LID system can integrate the phonological representation from source language through model transferring method to improve the LID performance.

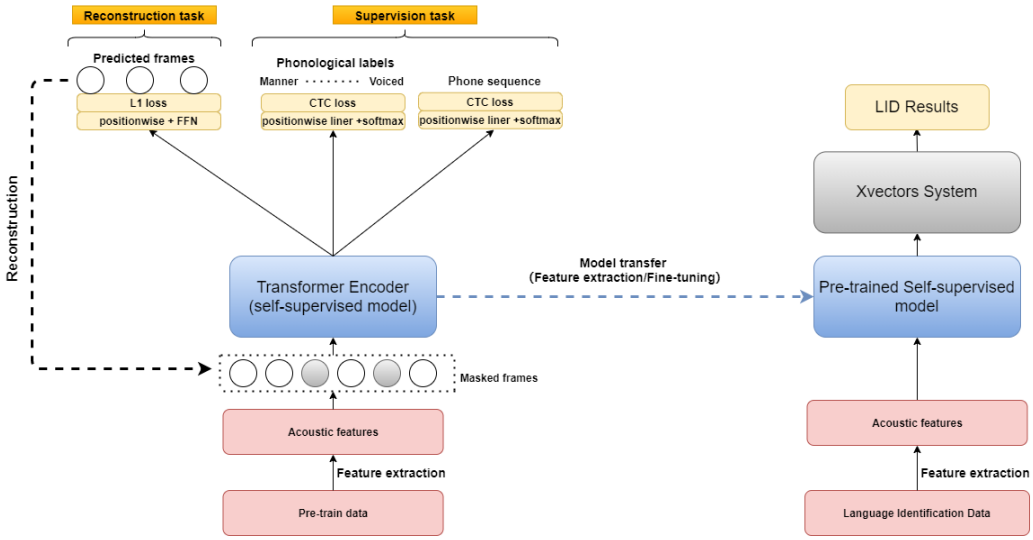


Figure 1. Block diagram of the proposed self-supervised pretrain model for language identification. For model transfer, two approaches are considered: Feature extraction and Fine tuning, which will be described in Section 4.

The rest of the paper is organized as follows: Section 2 presents some related works. Section 3 describes the definition of the phonological. Section 4 gives the model architecture. Section 5 and Section 6 show the experiments set and experimental results. Section 7 concludes the paper.

2. Related Work

Inspired by the Masked Language Model (MLM) task from BERT [12], researches have explored using BERT-style tasks to pretrain speech encoders. In [13], the author proposed a transformer encoder based pretrain model named Mockingjay, where the input frames are masked to zero. The model learns to reconstruct and predict the original frames. In Audio ALBERT [14], Mockingjay is modified to share parameters across Transformer layers. In [13], a pretrained model using time and frequency alteration objective, the results show that the pretrained model can improve several downstream tasks. In PASE [15], a single neural encoder is trained to solve multiple self-supervised tasks at the same time, including reconstruction of waveform, Log power spectrum, MFCC, prosody and other binary discrimination tasks.

The phonological knowledge has been used in many speech tasks. Many researches design neutral network model to map the acoustic features to articulatory features (AFs) by using phonological knowledge. In [16], the authors combined the acoustic and articulatory

features, which can improve the speaker identification performance. [17] applied the articulatory features (AFs) to Deep Bottleneck (DBN) features based ivector and xvector systems, which can get better performance than baseline. Because the AFs are language-independent features, there are many researches focused on multilingual speech recognition [18–20]. Previous studies generally take a bottom-up approach and train phonological feature detectors, and here we jointly train the phonological labels and the acoustic frames reconstruction.

Most traditional solutions are based on ivector which is extracted from Gaussian mixture model (GMM) [21]. Recently, as the development of the deep neural network (DNN), it has been demonstrated that the DNN can bring significant improvement for LID. In [22], the authors developed deep bottleneck features (DBF) based ivectors, which were extracted from a well-trained hybrid automatic speech recognition (ASR) system. In [23], authors proposed a LID model namedPhonetic Temporal Neural Model (PTN) where LSTM-RNN LID system that accepts phonetic features produced by a phone-discriminative DNN as the input. In this work, the self-supervised model is used to learn phonological knowledge and transfer the knowledge to LID model.

3. Phonological Definition

In this experiment, Mandarin is used to train the self-supervised model. According to the previous work [24] and the International Phonetic Alphabet (IPA) [25], we defined the six speech attributes for Mandarin, which can be found in Table 1. These attributes of speech can be comprehended by a collection of information from fundamental speech sounds. For each phoneme, it has six classes: *Place*, *Manner*, *Front*, *Height*, *Roundness* and *Voiced*. In this table, the *nil* means "not specified". For example, the articulatory class *Manner* does not exist in consonants, thus, in consonants phones, this class is defined as *nil*.

Table 1. Articulatory classes definition for Mandarin.

Articulatory Classes	Articulatory Attributes	Phoneme
Place	alveolar	n l d t
	bilabial	b p m
	dental	c iy
	labiodental	f
	palatal	aa o u j q a oo i iz ei uu g ee x v e vv ii
	pos-alveolar	zh r sh ix ch
	retroflection	er
	velar	h
	nil	sil
Manner	fricative	f s sh r x h
	lateral	i
	nasal	m ng n
	stop	t q j b d ch zh c g k p z
	nil	all_vowels sil

Table 1. *Cont.*

Articulatory Classes	Articulatory Attributes	Phoneme
Voiced	voiced	oo uu o n ng ei ix a er i vv ee ii iz r m e u iy aa i v
	unvoiced	s ch p zh z x sh b t g q k c h d j f
	nil	all_consonants nil
Height	height	iz vv i iy ix v u uu ii
	low	ee e oo o
	mid-height	ee e oo o
	mid-low	er er
	nil	all_consonants sil
Round	round	u v uu vv o oo
	unround	ix iy a e ee iz i ii aa er ei
	nil	all_consonants sil
Front	front	i ei v iy vv iz ii
	central	a aa er ix
	back	uu e ee oo o u
	nil	all_consonants sil

4. Model Structure

4.1. Architecture for Self-Supervised Model

In recent years, Transformer model has been applied successfully in mask speech tasks [26,27], so we use a standard multi-layer Transformer encoder with multi-head self-attention for left-and-right bidirectional encoding to train the self-supervised model. Each encoder layer has two sublayers, the first is a multi-head self-attention network, and the second is a feed-forward layer, each sub-layer has a residual connection followed by layer normalization [6]. To make the model aware of the input sequence order, the positional encoding is used. The sinusoidal positional encoding instead of learnable positional embeddings because acoustic features can be arbitrarily long with high variance [13]. After the Transformer encoder, the fully connected layer is used to reduce the dimension of the output vector.

4.2. Multi-Task Learning

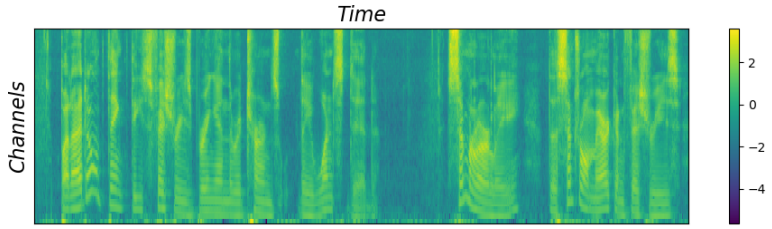
As we describe above, there are two tasks in our proposed model: reconstruction task and supervision task using phonological knowledge. By training the model jointly with these tasks, the model can learn more on phonological space since LID relies on phonological information. The training stage is described in following parts:

- **Reconstruction task**: For reconstruction task, we apply two kinds of masking approaches on the input frames: *Time mask* and *Channel mask*.
  - *Time mask*: Follow BERT and some previous work on self-supervised models with speech, the *Time masking* strategy is used. Through the masking of segments along the time axis, our model learns bidirectional representations from past and future contex. For the *Time mask*, 15% of the input frames are selected. In the selected 15% of frames, (1) 80% of those frames are masked to zero. (2) 10% of those frames are replaced by random frames. (3) leave the frames untouched 10%

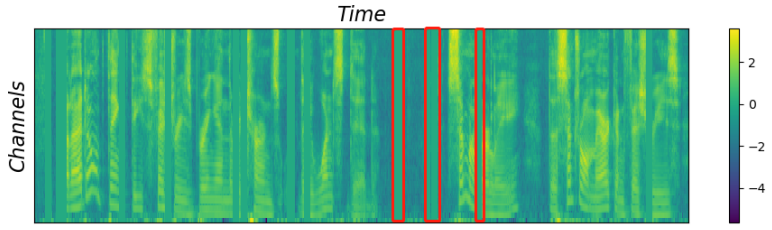
of the time. Through the masking of contiguous segments along the time axis, our model learns bidirectional representations from past and future context.

– *Channel mask*: Inspired by Specaugment [28] and TERA, we also introduce channel mask on top of time mask. For channel mask, a block of consecutive channels is masked to zero on all the time step. In our experiments, the percentage of the masked channels are 20%. In [29], the results showed that using channel masking can make the model to learn more on speaker representation. So by using the channel masking, we want to find whether the channel masking can learn phonological representation since the CTC loss with phonological knowledge is used.

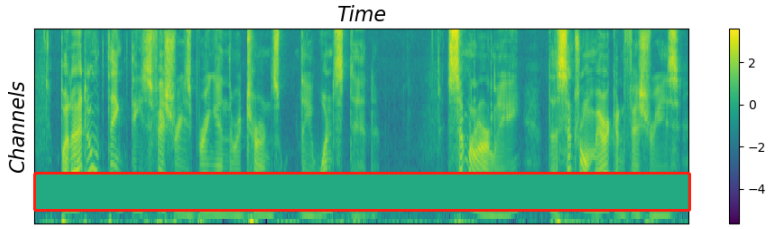
To better illustrate the *Time mask* and *Channel mask*, we visualize different masking strategies, as plotted in Figure 2.



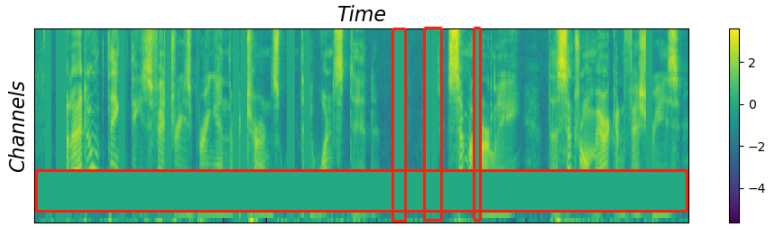
(**a**)Original Fbank features



(**b**)Time masking



(**c**)Channel masking



(**d**)Combining *Time masking* and *Channel masking*

Figure 2. Visualization of the inputs frames on different masking strategies. The masking parts are highlighted in *Red*.

For both masking strategies, we follow RoBERTa [30] and generate new masking patterns for each batch. Finally, we reconstruct all the frames to induce acoustic

information at all positions and more explicitly train the model. The loss is described as follows:

$$L_{rec} = \frac{1}{T}\sum_{t=1}^{T}|x_t - z_t| \quad (1)$$

where z_t are outputs of the transformer encoder and x_t is the input.

- **Supervision task**: To make the model learn phonological representation from speech, the CTC loss function is applied and the phoneme and phonological labels are used. The CTC approach is an objective function for sequence labeling problem [31], which doesn't rely on force alignment between input and output labels. For the phoneme output, the loss of the phoneme based CTC can be calculated as:

$$L_{phn-ctc}(y|x) = \sum_{\pi\epsilon\Omega(y)} P(\pi|x) \quad (2)$$

where the x is the input acoustic features and y is the phoneme sequence.
There are 6 classes for phonological labels, so for ith class, the loss is:

$$L_{pho-ctc-i}(y|x) = \sum_{\pi\epsilon\Omega(y)} P(\pi|x) \quad (3)$$

where the x is the input acoustic features and y is the phonological labels sequence.
So the multitask learning loss is:

$$L = \lambda_{rec}L_{phn-ctc}(y|x) + \lambda_{phn}L_{pf-ctc-1}(y|x) + \lambda_i L_{pf-ctc-i}(y|x) \quad (4)$$

The λ represents the weight of the corresponding task.

When using the multi-task learning, the performance of the model can be sensitive to the weight between different tasks and finding optimal values can be expensive. To better train the model, we propose to use the adaptive loss function derived in to automatically weight the task-specific loss functions [32], i.e.,

$$L_{ada}(\sigma_1, \sigma_2\sigma_i) = \frac{1}{\sigma_1^2}L_{rec} + \frac{1}{\sigma_2^2}L_{phn-ctc} + \frac{i}{\sigma_1^2}L_{pf-ctc-i} + log\sigma_1\sigma_2\sigma_i \quad (5)$$

Thus this adaptive loss is used for model training.

4.3. Xvector System

For the xvector system, the Time Delay Neutral Network (TDNN) based xvector system is chosen because it can get the state-of-art results and always considered as the baseline for LID [33,34].

The first five layers are the extended context layers while following a statistical pooling layer then accumulates all frame-level outputs. Then the outputs are calculated the mean and standard deviation and the segment-level fixed-dimension representation is obtained. The segment-level statistics are passed to the fully connected hidden layers. There are main two ways to incorporate the pretrained model to language identification tasks.

4.4. Incorporating with Language Identification Tasks

- **Feature Extraction**: The first approach is to extraction the features from the last layer of transformer encoder. The extracted features are fed to LID system as input. Parameters of the self-supervised model is frozen when training the LID system in this approach. In later experiments, we denote this approach as FE.
- **Fine-tuning**: The second approach is to fine-tune the self-supervised model with LID model. Here the output of the self-supervised model is connected to the xvector model. We then update the pretrained model with random initialized xvectors model. In later experiments, we denote this approach as FT.

5. Experiment Setup

5.1. Datasets

For the self-supervised model, the THCHS30 [35] dataset is used and the details of the dataset are listed in Table 2. The LID experiments are conducted on the second oriental language recognition (OLR) challenge AP17-OLR [34]. The training set contains 10 different languages: Mandarin, Cantonese, Indonesian, Japanese, Russian, Korean, Vietnamese, Kazakh, Tibetan and Uyghur. In these languages, the male and female speakers are balanced. For the training set, it is recorded by mobile phones, with a sampling rate of 16 kHZ and a sample size of 16 bits.

Table 2. Statistics of THCHS-30 database.

Dataset	Speaker	Utterances	Duration (Hours)
Train	30	10,893	27.23
Test	10	2496	6.24

Our systems are evaluated on AP17 challenge's development set , which is selected apart from the training set. The development set contains three test sets of different conditions: 1 s, 3 s and full length utterance condition, which are denoted as *1 s*, *3 s*, and *full-length*. The test utterances of the 1 s and 3 s are randomly excerpted from the full-length utterance. If a test utterance is not sufficient long for the excerption, it is simply discarded. The details of the dataset are described in Table 3.

Table 3. Statistics of AP17 challenge database.

Languages	Train		Test	
	No. of Speakers	Total Utt	No. of Speakers	Total Utt
Kazakh	86	4200	86	1800
Tibetan	34	111,000	34	1800
Uyghur	353	5800	353	1800
Cantonese	24	7559	6	1800
Mandarin	24	7198	6	1800
Indonesian	24	7671	6	1800
Japanese	24	7662	6	1800
Russian	24	7190	6	1800
Korean	24	7196	6	1800
Vietnamese	24	7200	6	1800

5.2. Self-Supervised Model Setup

For the self-supervised model training, the input to our network is cepstral mean-normalized Fbank of speech utterances. All of the features are extracted using open-source toolkit Kaldi [36] , using windows of 25 ms and an overlap of 10 ms. We stack every 3 frames to reduce the memory cost of long sequences [37]. The self-supervised Transformer encoder architecture has 12 self-attention layers and the number of multi-head attention is 12. Gradient descent training with mini-batches of size 16 is used to find model parameters. The Adam optimizer [38] is employed for updating model parameters, where learning rate is warmed up over training.

All the experiments are conducted on Pytorch [39].

5.3. LID Systems

To compare our proposed model, different systems are introduced:

- **Xvectors**: For the xvector system, the open-source toolkit asv-subtools is used [40]. The acoustic features for xvector system is 23-dim MFCCs and before feeding to the xvector system, a frame-level energy-based voice activity detection (VAD) is used to select voiced speech frames. The xvector system contains 6-layers TDNN layers, the details of the TDNN configuration is shown in Table 4. To get the xvectors, 512-dimensional embedding features are extracted at the layer segment6 of the network before the nonlinearity. We apply the Linear Discriminant Analysis (LDA) to reduce the dimension of output vectors. For the back-end classifiers are used: Logistic Regression(LR) and Probabilistic Linear Discriminant Analysis(PLDA).

Table 4. The standard xvector architecture. The T represent the speech frames to input the DNN, and t is the current frame.

Layer	Layer Context	Context	Input Dim	Output Dim
Frame 1	$(t-2, t+2)$	5	200	512
Frame 2	$(t-2, t, t+2)$	9	1536	512
Frame 3	$(t-3, t, t+3)$	15	1536	512
Frame 4	(t)	15	512	512
Frame 5	(t)	15	512	1500
stats pooling	[0,T)	T	1500T	3000
segments 6	(0)	T	3000	512
segments 7	(0)	T	512	512
softmax	(0)	T	512	6

- **SSL**$_{xv}$: The self-supervised learning pre-trained model is used to incorporate xvector system, as shown in Figure 1.
- **PTN**: Phonetic Temporal Neural Model (PTN) [23] an auxiliary phonetic model produces phonetic feature, and an RNN LID model is used to identify the language. The PTN is also the baseline for AP17 OLR challenge [34]. Meanwhile, in the results reported by this model, the same source data THCHS30 is used.
- **IM-LSTM-PTN**: The structure of the IM-LSTM-PTN is described in [41]. This model is the submitted model to AP17 OLR challenge which ranked 4th among all the participated teams (http://cslt.riit.tsinghua.edu.cn/mediawiki/index.php/OLR_Challenge_2017). Based on the PTN, The IM-LSTM-PTN uses a modified LSTM which has a top-down connection from time t to time $t+1$.

6. Experimental Results

6.1. Language Identification Results

First we list all the results on proposed methods in Table 5. In this table, it can be found that for back-end classifiers, the LR always performs better than PLDA in all the LID systems. For model transferring approaches, in short duration test condition *1 s*, the FE outperforms than FT. This is because that the 1s utterance is too short so it is hard to do fine-tuning on pretrained model. In *3 s* and *Full-length* test condition, FT can get better performance. When applying the channel masking on top of time masking, the performance of the LID always worse than only using time masking. The reason is that the LID often requires linguistic knowledge which is on time axis and the CTC loss is a sequence-level loss. Previous works also show that the channel masking mainly helps to encode speaker information [29]. Then we compare our best results with some other results reported in some previous works, which are listed in Table 6. Among al the results reported, our proposed model still can get the best performance. More specifically, when comparing our best results with PTN which is also a phonetic based LID model, our results still have significant improvement.

Table 5. LID results (EER %) on different test set. In this table, FT means fine-tuning and FE means feature extraction. "+" means the corresponding approach is used and "−" means not. "+ *channel*" means the channel masking is applied on the corresponding model.

Methods	LID Sytem	LR	PLDA	1 s	3 s	Full-Length
Baseline	Xvector $_{PLDA}$	−	+	12.9	5.9	4.9
	Xvector$_{LR}$	+	−	10.9	5.6	3.4
Proposed Methods	SSL$_{xv+PLDA}$(FT)	−	+	10.9	4.5	3.2
	+ *channel*	−	+	11.6	4.6	3.5
	SSL$_{xv+PLDA}$(FE)	−	+	11.9	3.8	2.3
	+ *channel*	−	+	13.8	5.1	3.6
	SSL$_{xv+LR}$ (FT)	+	−	10.0	**2.5**	**1.6**
	+ *channel*	+	−	10.1	4.6	3.4
	SSL$_{xv+LR}$(FE)	+	−	**9.7**	3.3	2.4
	+ *channel*	+	−	11.8	3.5	2.8

Table 6. LID results (EER %) on proposed model and some results from previous works. In this table, the results of proposed model are the best from Table 5.

Methods	LID Sytem	1 s	3 s	Full-Length
Proposed Methods	SSL$_{xv+LR}$ (FT)	10.0	**2.5**	**1.6**
	SSL$_{xv+LR}$(FE)	**9.7**	3.3	2.4
Competive Methods	PTN [23]	12.3	8.2	8.0
	LSTM-LID [23]	11.7	8.0	7.8
	TDNN-LID [23]	15.6	15.4	14.6
	ivector+SVM [23]	12.6	4.7	3.3
	IM-LSTM-PTN [41]	11.7	8.0	7.8
	stacked-SDC-Resnet [42]	14.4	11.1	10.1

6.2. Analysis of Different Tasks

To analyze the influence of different tasks in the training stage, we train our model with single task and apply to LID. To simplify the experiments, we conduct the experiments using SSL$_{xv+LR}$ (FT). The results are listed in the Table 7.

It shows that all the single tasks can improve the LID results. Meanwhile, the model only using reconstruction loss performs worse than only using CTC loss does which is because the phonetic knowledge always benefits the LID system. By combining all the losses and using weighted adapted loss, the LID system can get the best performance. Considering all the results above, it can be concluded that through the reconstruction task, the model can capture contextual representation and when the supervision task is applied, the phonological representation is learned. So by jointly training these two tasks, our proposed method can improve the performance on LID.

Table 7. Multi-task learning results versus different single tasks (EER %). The results of SSL$_{xv+LR}$ (FT) are taken from Table 5.

LID Systems	1 s	3 s	Full-Length
Xvector	12.9	5.9	4.9
SSL$_{xv+LR}$ (FT), only $L_{phn-ctc}$	10.5	2.9	2.0
SSL$_{xv+LR}$ (FT), only $L_{pho-ctc}$	10.7	3.2	2.3
SSL$_{xv+LR}$ (FT), only L_{rec}	11.1	3.6	2.6
SSL$_{xv+LR}$ (FT)	10.0	2.5	1.6

7. Conclusions

In this paper, a self-supervised model integrating phonological knowledge is proposed for language identification. The proposed model achieves training speech perception and speech production jointly by using self-supervised approach and our model can get significant improvement on downstream task (language identification task). In the self-supervised model, the reconstruction loss and CTC loss with phonological labels are jointly used to train the model. For the reconstruction loss, we apply masking on time and channel axis and use L1 loss to reconstruct the output frames. For the CTC loss, the phoneme labels and phonological labels are used to train the model. By doing the jointly training, the model aims to learn high-level speech representation at phonological space. The final results show that our proposed model can get best performance with features extraction (FE) model transfer approach and LR as the back-end classifier. Our future work will explore on applying our proposed model on other speech tasks.

Author Contributions: Conceptualization, Q.Z., X.X., C.H. and H.C.; methodology, Q.Z.; implementation, Q.Z. and C.H.; validation, Q.Z., X.X., C.H. and H.C.; writing—original draft preparation, Q.Z.; supervision, X.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by National Nature Science Foundation of China (No.11590772), Science and Technology Innovation Foundation of Shenzhen (JCYJ20180504165826861).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: AP17 language identification dataset: Tang, Z.; Wang, D.; Chen, Y.; Chen, Q. 2017; AP17-OLR Challenge: Data, Plan, and Baseline. arXiv:cs.CL/1706.09742. THCHS30 dataset: Wang, D.; Zhang, X. 2015, THCHS-30: A Free Chinese Speech Corpus. arXiv:cs.CL/1512.01882.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CTC	Connectionist Temporal Classification
DNN	Deep Neural Networks
EER	Equal Error Rate
Fbank	Filterbank
FE	Feature-extraction
FT	Fine-tuning
MFCC	Mel-Frequency Cepstral Cofficients
LID	Language Identification
LDA	Linear Discriminant Analysis
LR	Logistic Regression
OLR	Oriental Language Recognition
PLDA	Probabilistic Linear Discriminant Analysis

References

1. Ebbers, J.; Kuhlmann, M.; Cord-Landwehr, T.; Haeb-Umbach, R. Contrastive Predictive Coding Supported Factorized Variational Autoencoder for Unsupervised Learning of Disentangled Speech Representations. *arXiv* **2021**, arXiv:2005.12963.
2. Schneider, S.; Baevski, A.; Collobert, R.; Auli, M. wav2vec: Unsupervised Pre-training for Speech Recognition. *arXiv* **2019**, arXiv:1904.05862.
3. Baevski, A.; Schneider, S.; Auli, M. vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations. *arXiv* **2020**, arXiv:1910.05453.
4. Baevski, A.; Auli, M.; Mohamed, A. Effectiveness of self-supervised pre-training for speech recognition. *arXiv* **2020**, arXiv:1911.03912.
5. van den Oord, A.; Li, Y.; Vinyals, O. Representation Learning with Contrastive Predictive Coding. *arXiv* **2019**, arXiv:1807.03748.
6. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.

7. Pitenis, Z.; Zampieri, M.; Ranasinghe, T. Offensive Language Identification in Greek. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 13–15 May 2020; European Language Resources Association: Marseille, France, 2020; pp. 5113–5119.
8. Zissman, M.A. Comparison of four approaches to automatic language identification of telephone speech. *IEEE Trans. Speech Audio Process.* **1996**, *4*, 31. [CrossRef]
9. Lopez-Moreno, I.; Gonzalez-Dominguez, J.; Plchot, O.; Martinez, D.; Gonzalez-Rodriguez, J.; Moreno, P. Automatic language identification using deep neural networks. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 5337–5341.
10. Lee, C.H.; Clements, M.A.; Dusan, S.; Fosler-Lussier, E.; Johnson, K.; Juang, B.H.; Rabiner, L.R. An overview on automatic speech attribute transcription (ASAT). In Proceedings of the Eighth Annual Conference of the International Speech Communication Association, Antwerp, Belgium, 27–31 August 2007.
11. Morris, J.; Fosler-Lussier, E. Combining phonetic attributes using conditional random fields. In Proceedings of the Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, 17–21 September 2006.
12. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
13. Liu, A.T.; Yang, S.W.; Chi, P.H.; Hsu, P.C.; Lee, H.Y. Mockingjay: Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020. [CrossRef]
14. Chi, P.H.; Chung, P.H.; Wu, T.H.; Hsieh, C.C.; Chen, Y.H.; Li, S.W.; Lee, H.Y. Audio ALBERT: A Lite BERT for Self-supervised Learning of Audio Representation. *arXiv* **2021**, arXiv:2005.08575.
15. Pascual, S.; Ravanelli, M.; Serrà, J.; Bonafonte, A.; Bengio, Y. Learning Problem-agnostic Speech Representations from Multiple Self-supervised Tasks. *arXiv* **2019**, arXiv:1904.03416.
16. Hong, Q.B.; Wu, C.H.; Wang, H.M.; Huang, C.L. Combining Deep Embeddings of Acoustic and Articulatory Features for Speaker Identification. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7589–7593. [CrossRef]
17. Yu, J.; Guo, M.; Xie, Y.; Zhang, J. Articulatory Features Based TDNN Model for Spoken Language Recognition. In Proceedings of the 2019 International Conference on Asian Language Processing (IALP), Shanghai, China, 15–17 November 2019; pp. 308–312. [CrossRef]
18. Zhan, Q.; Motlicek, P.; Du, S.; Shan, Y.; Ma, S.; Xie, X. Cross-lingual Automatic Speech Recognition Exploiting Articulatory Features. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 1912–1916. [CrossRef]
19. Tong, S.; Garner, P.N.; Bourlard, H. Fast Language Adaptation Using Phonological Information. *Interspeech* **2018**, *C4*, 2459–2463.
20. Abraham, B.; Umesh, S.; Joy, N.M. Joint Estimation of Articulatory Features and Acoustic Models for Low-Resource Languages. *Interspeech* **2017**, 2153–2157. [CrossRef]
21. Dehak, N.; Torres-Carrasquillo, P.; Reynolds, D.; Dehak, R. Language Recognition via I-Vectors and Dimensionality Reduction. In Proceedings of the Twelfth Annual Conference of the International Speech Communication Association, Florence, Italy, 28–31 August 2011; pp. 857–860.
22. Song, Y.; Jiang, B.; Bao, Y.; Wei, S.; Dai, L. I-vector representation based on bottleneck features for language identification. *Electron. Lett.* **2013**, *49*, 1569–1570. [CrossRef]
23. Tang, Z.; Wang, D.; Chen, Y.; Li, L.; Abel, A. Phonetic temporal neural model for language identification. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **2017**, *26*, 134–144. [CrossRef]
24. Zhang, C.; Liu, Y.; Lee, C.H. Detection-based accented speech recognition using articulatory features. In Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2011, Proceedings, Waikoloa, HI, USA, 11–15 December 2011. [CrossRef]
25. Wikipedia contributors. International Phonetic Alphabet—Wikipedia, The Free Encyclopedia. 2021. Available online: https://en.wikipedia.org/w/index.php?title=International_Phonetic_Alphabet&oldid=1041762539 (accessed on 23 June 2021).
26. Dong, L.; Xu, S.; Xu, B. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5884–5888.
27. Karita, S.; Chen, N.; Hayashi, T.; Hori, T.; Inaguma, H.; Jiang, Z.; Someki, M.; Soplin, N.E.Y.; Yamamoto, R.; Wang, X.; et al. A comparative study on transformer vs rnn in speech applications. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 449–456.
28. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *arXiv* **2019**, arXiv:1904.08779.
29. Liu, A.T.; Li, S.W.; yi Lee, H. TERA: Self-Supervised Learning of Transformer Encoder Representation for Speech. *arXiv* **2020**, arXiv:2007.06028.
30. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.

31. Graves, A.; Fernández, S.; Gomez, F. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the International Conference on Machine Learning, ICML 2006, Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376.
32. Kendall, A.; Gal, Y.; Cipolla, R. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. *arXiv* **2018**, arXiv:1705.07115.
33. Snyder, D.; Garcia-Romero, D.; McCree, A.; Sell, G.; Povey, D.; Khudanpur, S. Spoken Language Recognition using X-vectors. *Odyssey* **2018**, 105–111. [CrossRef]
34. Tang, Z.; Wang, D.; Chen, Y.; Chen, Q. AP17-OLR Challenge: Data, Plan, and Baseline. *arXiv* **2017**, arXiv:1706.09742.
35. Wang, D.; Zhang, X. THCHS-30: A Free Chinese Speech Corpus. *arXiv* **2015**, arXiv:1512.01882.
36. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi Speech Recognition Toolkit. In Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Waikoloa, HI, USA, 11–15 December 2011.
37. Salazar, J.; Kirchhoff, K.; Huang, Z. Self-attention Networks for Connectionist Temporal Classification in Speech Recognition. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May . [CrossRef]
38. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2017**, arXiv:1412.6980.
39. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv* **2019**, arXiv:1912.01703.
40. Tong, F.; Zhao, M.; Zhou, J.; Lu, H.; Li, Z.; Li, L.; Hong, Q. ASV-Subtools: Open Source Toolkit for Automatic Speaker Verification. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6184–6188.
41. Zhan, Q.; Zhang, L.; Deng, H.; Xie, X. An improved LSTM for language identification. In Proceedings of the 2018 14th IEEE International Conference on Signal Processing (ICSP), Beijing, China, 12–16 August 2018; pp. 609–612.
42. Vuddagiri, R.K.; Vydana, H.K.; Vuppala, A.K. Improved Language Identification Using Stacked SDC Features and Residual Neural Network. In *SLTU*; ISCA: Gurugram, India, 2018; pp. 210–214.