mathematics

MDPI

Article

Unsupervised Representation Learning with Task-Agnostic Feature Masking for Robust End-to-End Speech Recognition

June-Woo Kim [1], Hoon Chung [2] and Ho-Young Jung [1,*]

[1] Department of Artificial Intelligence, Kyungpook National University, Daegu 41566, Republic of Korea
[2] Electronics and Telecommunications Research Institute, Daejeon 34129, Republic of Korea
* Correspondence: hoyjung@knu.ac.kr

Abstract: Unsupervised learning-based approaches for training speech vector representations (SVR) have recently been widely applied. While pretrained SVR models excel in relatively clean automatic speech recognition (ASR) tasks, such as those recorded in laboratory environments, they are still insufficient for practical applications with various types of noise, intonation, and dialects. To cope with this problem, we present a novel unsupervised SVR learning method for practical end-to-end ASR models. Our approach involves designing a speech feature masking method to stabilize SVR model learning and improve the performance of the ASR model in a downstream task. By introducing a noise masking strategy into diverse combinations of the time and frequency regions of the spectrogram, the SVR model becomes a robust representation extractor for the ASR model in practical scenarios. In pretraining experiments, we train the SVR model using approximately 18,000 h of Korean speech datasets that included diverse speakers and were recorded in environments with various amounts of noise. The weights of the pretrained SVR extractor are then frozen, and the extracted speech representations are used for ASR model training in a downstream task. The experimental results show that the ASR model using our proposed SVR extractor significantly outperforms conventional methods.

Keywords: speech vector representation; representation learning; unsupervised learning; feature representation extractor; speech recognition; deep learning; neural network; speech processing

MSC: 68T10

Citation: Kim, J.-W.; Chung, H.; Jung, H.-Y. Unsupervised Representation Learning with Task-Agnostic Feature Masking for Robust End-to-End Speech Recognition. *Mathematics* **2023**, *11*, 622. https://doi.org/10.3390/math11030622

Academic Editors: Ravil Muhamedyev and Evgeny Nikulchev

Received: 6 January 2023
Revised: 22 January 2023
Accepted: 23 January 2023
Published: 26 January 2023

1. Introduction

Automatic speech recognition (ASR), which attempts to automatically recognize the speech of all types of individuals, has been an actively expanding topic with ASR services. Generally, end-to-end speech recognition models based on supervised learning are limited by their requirements for a large amount of speech and corresponding labeled text data. In order to achieve high recognition performance, large amounts of paired speech and labeled text data are required for training ASR models [1]. Although such datasets are difficult and expensive to obtain, most ASR systems are still being developed using paired speech datasets. Specifically, most ASR models are trained with refined benchmark speech datasets, such as TIMIT [2], WSJ [3], LibriSpeech [4], and Libri-Light [5], and they are mainly effective on in-domain data. Real speech "in the wild " is very diverse and leads to severe degradation of ASR performance [6–8].

To make matters worse, since most of the benchmark curated speech datasets [2–5] are built with very limited diversity, mostly representing healthy adults, it is challenging to accurately recognize the speech of children [9,10], the elderly [11–13], or those using dialects. Consequently, speech recognition performance suffers in highly variable scenarios, such as far-field or noisy environments [6–8,14,15], where the conditions or personal characteristics [16–19] degrade the performance compared with normal speech. In addition,

recognizing new or trending words is important for ASR systems, but updating already built end-to-end ASR systems every time is time-consuming and resource-intensive. Accordingly, deep-learning-based ASR systems are trained based on labeled speech datasets, so optimal learning is restricted due to the limited amount of curated speech datasets.

To handle the aforementioned problems, an unsupervised or self-supervised learning method using unlabeled speech data has been proposed. In particular, unsupervised learning is an effective approach that enables leveraging large-scale speech data for speech vector representation (SVR) learning without the labeled text [20–31]. This unsupervised pretrained SVR model can then be used as a frozen speech feature extractor, and the extracted speech representations are applied for speech recognition model training in a downstream task. The pretrained SVR-based ASR model demonstrated performance gains over using conventional mel filterbank as input features. Leveraging the unsupervised pretrained SVR showed significant improvement compared with the supervised learning-based ASR model [26–28].

However, the success of the aforementioned studies seems to be limited to unsupervised pretraining methods using well-curated speech datasets. Recent unsupervised SVR learning studies have mostly focused on large-scale English speech datasets, which have relatively clean, curated read speech [20–31]. Despite the importance of the practical tasks of ASR systems, it is difficult to find studies that have focused on pretraining an SVR model with noncurated speech datasets containing diverse conditions.

Speech data in the wild with unexpected noise, dialects, and personal characteristics are very different from those in curated speech datasets. Similar to the supervised ASR model, it is hard to achieve high recognition performance with the SVR extractor pretrained with a curated speech dataset [32]. To obtain a robust ASR model that can be used in real environments, the SVR extractor has to be able to handle the speech data recorded from various speakers in different environments.

To solve these issues, we introduce a novel unsupervised SVR learning approach for a robust ASR model. Our proposed method is designing an unsupervised method of speech masking, inspired by both BERT [33] and SpecAugment [34], to stabilize the SVR extractor learning and to make SVR generalized for ASR tasks. To achieve this, we propose a noise masking strategy that introduces noise and masking into various combinations of the time–frequency regions of speech features and makes the SVR extractor more robust.

For representation learning experiments, the SVR model is trained to reconstruct the original speech from masked speech using unsupervised learning based on our proposed masking policy that only concerns speech data. We use a total of around 18,000 h of various Korean speech datasets recorded with diverse speakers in noisy and normal environments for SVR model training. For the ASR experiments, the pretrained SVR extractor is frozen, and the extracted speech representations are used for ASR model training. We present the overall architecture of our model in Figure 1.

To summarize, the main contributions of this work are listed below:

- We propose a novel unsupervised SVR learning method for robust ASR performance in practical applications.
- We demonstrate that incorporating a noise masking strategy into various combinations of the time–frequency regions of spectrum features makes the SVR extractor more robust, and that the speech recognition performance using our proposed method significantly outperforms existing methods in real conditions.
- We provide ASR performance for the SVR extractor trained with real speech datasets of varying sizes (1 k–18 k h) and present ASR performance for speech datasets that were not used for pretraining of the SVR. To obtain an accurate comparison, we report the speech recognition performance for four different conditions, which experimentally shows that our unsupervised masking method is effective.
- To the best of our knowledge, this is the first attempt at pretraining an SVR model with large-scale real Korean speech. We further explore and provide a wide range of

ablation studies and analyses on the results of practical ASR using various masking combinations of the time–frequency regions and two noise masking techniques.

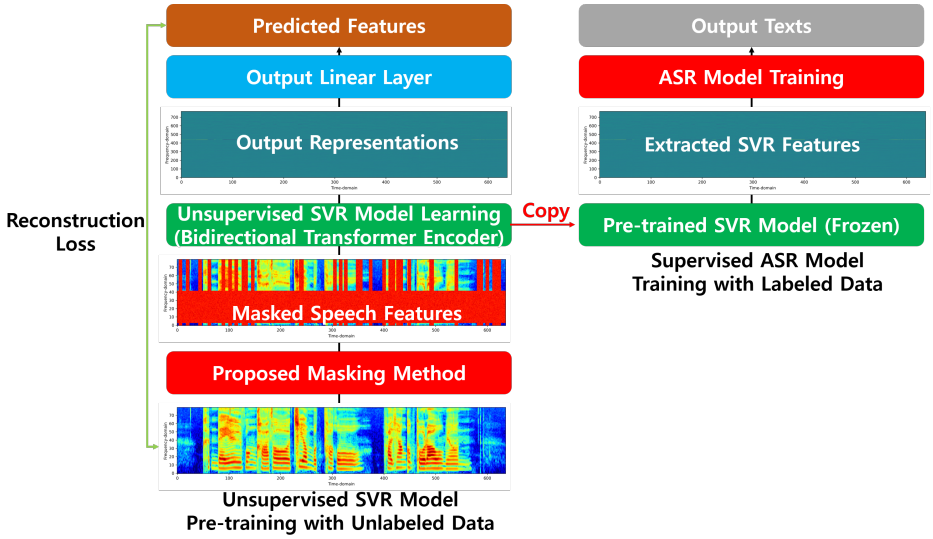

Figure 1. Illustration of the proposed overall architecture for unsupervised SVR pretraining and ASR training.

The rest of this paper is organized as follows: We give first the related works briefly in Section 2. In Section 3, we demonstrate details of the proposed masking method and model architecture. In Section 4, we present details on the experimental setting and datasets. In Section 5, we provide experimental results to demonstrate the effectiveness of our proposed method. Finally, discussion and conclusions are drawn in Sections 6 and 7.

2. Related Work

Recently, unsupervised speech feature representation learning has emerged as an effective approach to utilize large-scale speech data with no supervision [20–25,27–31,35,36]. Two types of SVR approaches are as used: masking-based reconstruction [33] and contrastive predictive coding [20]. In this paper, we focus on the former method.

The most widely known reconstruction method is the masked language model (MLM), which was used by BERT [33] and proposed in the field of natural language processing (NLP). BERT has been shown to acquire strong feature representation through the process of reconstructing the original tokens from contaminated input tokens by MLM approach using the bidirectional Transformer [37] encoders. Furthermore, fine-tuning the pretrained BERT outperforms previous NLP studies, even with a small amount of labeled data in the vast majority of downstream tasks.

Heavily inspired by BERT, recent SVR learning approaches have been studied based on MLM [23–25,27–31,36]. In particular, several time frames are randomly selected, and parts of them are filled with zero values [23–25,29,36], while the remainder of them are replaced with different frames [24,29,36] or specific vectors [27,28,30,31]. The objective function of the pretraining SVR model restores the masked speech frames to the original using reconstruction loss functions. Note that the pretraining of the SVR is performed at the utterance level. Unlike the contrastive predictive coding technique that relies on past or current frames [20], the masking method can predict the current frames by jointly conditioning the past and future information owing to the bidirectional recurrent neural networks [23,25], Transformer encoders [24,27,29,30,36], or Conformers [28,31]. Furthermore, quantizing approaches that convert continuous speech signals into discrete tokens, such as tokenizing techniques in the NLP domain, have also been successfully applied [35].

Fine-tuning the pretrained SVR model or using the model as the feature extractor has shown promising results in speech recognition in the downstream tasks, even with a

limited training dataset [27]. To this end, Connectionist Temporal Classification (CTC) [38] or Cross-Entropy with label smoothing [39] losses are leveraged for the fine-tuning and end-to-end ASR task, respectively.

However, the aforementioned approaches have been limited to pretraining using benchmark speech datasets. Despite the poor performance of the ASR systems in practical tasks, the majority of SVR methods are still focused on relatively clean, well-curated read speech datasets. In particular, most of them are the result of pretraining with 960 h of LibriSpeech [4] or 60 k hours of Libri-Light [5].

To overcome this limitation, we aim to apply the pretrained SVR model as the feature extractor, leading to a robust ASR model using various conditions of large-scale datasets. To this end, we introduce a novel and simple masking method, which we detail in the next section.

3. Method

In this section, we provide both SVR pretraining and ASR downstream training architecture, starting with the masking method for speech representation learning.

3.1. SVR Learning with Masking Method

As we explored in Section 2, the SVR model trained with the masking approach is able to help with several downstream tasks due to its capacity to extract powerful speech representations. Heavily inspired by previous approaches [23–25,29,33,36], we propose a masking-based reconstruction strategy with three parts: time–frequency and noise masks. Figure 2 shows the proposed masking scheme in this paper.

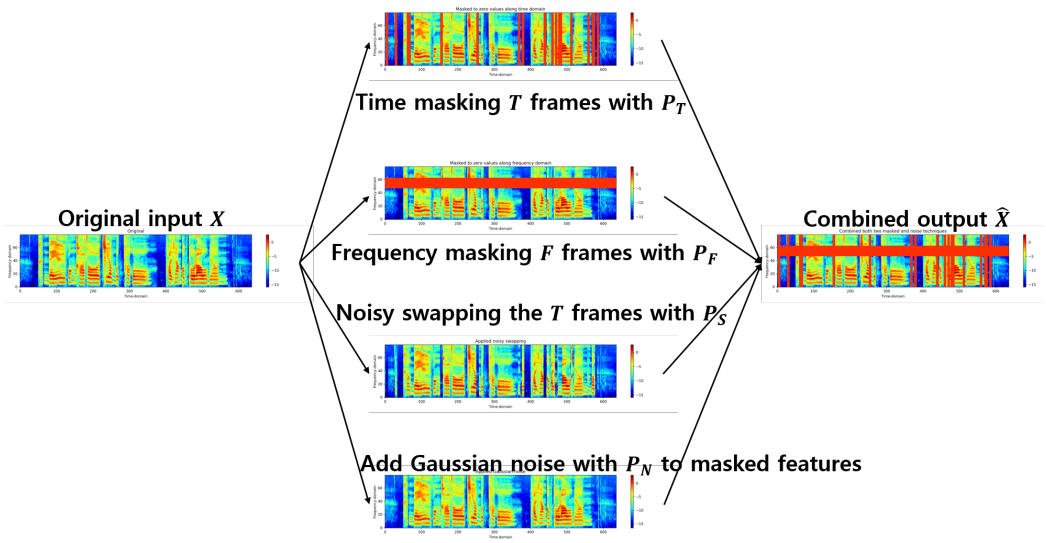

Figure 2. The overall flow of the proposed masking strategy has both the time–frequency mask and noise masking techniques. The masked part is presented in red.

3.1.1. Masking Time–Frequency Feature

The speech masking strategy is performed on the spectral feature level, and the log mel filterbank features are used. Given a speech utterance $X = (x_1, \ldots, x_L)$ of length L, we first randomly select T frames with a probability P_T from each speech X. We then mask 80% of the selected T frames with zero values, as suggested in [24,29,36]. The selected frames are consecutively masked with C frames and can be overlapped.

In the frequency region, we also randomly select F bins with a probability P_F from each X. Given, for example, P_F as 0.25, since we use an 80-dimensional log mel filterbank, up to 20 frequency bins are masked with zero values.

Reconstruction loss is used to learn contextualized speech representations, which allows the SVR model to predict original spectral features from our proposed masking

strategy. Owing to the bidirectional attention, the masked frames based on the time and frequency regions can be restored as original speech frames by the SVR model, capturing the contextual information from both directions.

3.1.2. Noise Masking Techniques

Here, we propose two noise masking techniques to perturb the data: noisy swapping and noise addition. We perform swapping on T frames from X with probability P_S. Note that noisy swapping is only used in the case that zero masking for the time region has not been applied. Unlike BERT, which performs word-level token replacement randomly, our method swaps selected frames based on the contamination of the corresponding filterbank.

Furthermore, we use the technique of adding Gaussian noise to disturb SVR model learning. In particular, after time–frequency regions masking and noisy swapping, we add a Gaussian noise task with a probability P_N. The Gaussian noise we set up is sampled from a normal distribution with zero mean and 0.2 variance and is added to each element of masked features.

By adding Gaussian noise, we can considerably increase the amount of pretraining data for learning SVR, which can be considered as a data augmentation technique. Furthermore, adding noise also provides another variation for speech frames replaced with zero values. The selected frames of time–frequency regions will be able to take on various nonzero values. Thus, our proposed noise masking techniques provide an advantage for practical ASR systems due to the increased diversity of the input data.

3.2. SVR Architecture

During SVR learning, only unlabeled speech data are used. Our SVR model architecture for pretraining is composed of several bidirectional Transformer encoders [33,37] that extract latent speech vector representations from a log mel filterbank. The left part of Figure 1 illustrates the unsupervised SVR model pretraining. The proposed network is a stack of bidirectional Transformer layers, where each layer has an identical configuration to those from the reconstruction objective functions. Our SVR model can directly extract high-level contextualized speech representations.

We obtain the corrupted speech features according to the proposed masking strategy with the 80-dimensional log mel filterbank. Since we set the default hidden size of the SVR model as 768, the 80-dimensional masked features are then fed into a 768-dimensional linear layer and transformed. Afterward, a sinusoidal positional encoding [37] is added to 768-dimensional features. In this paper, we use 768 dimensions for the location positional encoding size and 1500 for the maximum length.

The 768-dimensional vectors added with positional encoding values are then injected into the bidirectional Transformer encoder. Our proposed SVR network is composed of 3 Transformer encoder layers with a hidden size of 768, 12 multihead attentions, and a 2-layer feed-forward neural network with a hidden size of 3072 dimensions, followed by a dropout proportion of 0.1. Conventional Transformer encoders have leveraged look-up masks to input sequence to perform masking. We do not apply the look-up mask so that we can obtain the bidirectional properties. Through this, we conceived to jointly utilize information from the past and the future of speech features.

The 768-dimensional output vectors of the SVR model are then passed through the 80-dimensional linear layer that reduces them to 80, which is the size of the original mel filterbank. In other words, in order to reconstruct the original frames from the corrupted input mel filterbank, the 768-dimensional output vectors have to match the size of the target mel filterbank. This output linear layer in the left part of Figure 1 is only used during unsupervised SVR learning for reconstruction.

To learn contextualized speech vector representations for predicting original speech frames from masked samples, we use the reconstruction loss. To restore the final output of the SVR model to the original speech frames, L1 or L2 can be used as the reconstruction loss. Through this, the SVR model can recover the original speech frames from the masked

inputs. With this process, we expect that based on the pretrained SVR model, we can extract robust speech feature representations.

More formally, the L1 objective function of pretraining the SVR model is

$$\mathcal{L}_1 = \frac{1}{|X|} \sum_{x_L \in X} \sum_{i=1}^{L} |x_i^s - f(x_i^m)|, \tag{1}$$

where x_i^s are the original frames that are selected by two masking strategies for the time–frequency regions with both probabilities of P_T and P_F from each speech data x_L with length L. f is the output of the SVR model, and x_i^m are the corrupted input speech features, which are masked or noisy swapped frames of x_L selected by masking policy. The SVR model is trained by reducing the error between the corrupted x_i^m and the original x_i^s. For L2 loss, the following objective function is used:

$$\mathcal{L}_2 = \frac{1}{|X|} \sum_{x_L \in X} \sum_{i=1}^{L} (x_i^s - f(x_i^m))^2. \tag{2}$$

3.3. ASR Architecture

In order to show the effectiveness of the proposed method, we check whether performance is improved compared with an end-to-end ASR model trained with basic mel filterbank as model inputs. To this end, we introduce end-to-end ASR architecture that is able to receive mel filterbank or extracted features from the SVR model as inputs.

During ASR training, we use both speech and labeled data. We apply our pretrained SVR extractor, kept frozen to two tasks: KsponSpeech [40] and KForeignWordSpeech [41]. As the feature processor of the ASR network, we apply a 7-layer CNN with 3 max pooling layers, followed by batch normalization and ReLU.

The baseline ASR model that receives the 80-dimensional log mel filterbank as input is composed of strides [1, 1, 1, 1, 1, 1, 1] and kernel widths [3, 3, 3, 3, 3, 3, 3]. On the other hand, the ASR model using the 768-dimensional feature representation extracted from the SVR model consist of strides [3, 3, 1, 1, 1, 1, 1] and kernel widths [10, 5, 5, 4, 4, 3, 3]. Both ASR models have [16, 16, 32, 32, 32, 64, 64] channels and max pooling with strides [2, 2, 2] and kernel widths [2, 2, 2], respectively.

The feature maps generated by CNNs are fed into 2 linear layers of 512 and 256 dimensions, after which they are further sent to a 3-layer of 256-dimensional LSTM. Following that, they are then passed through an encoder. Our ASR network's encoder utilizes 3 Transformer encoder layers with a hidden size of 256, 8 multihead attentions, and a 2-layer feed-forward neural network having a 1024-dimensional hidden size.

The decoder is a component to map the text from a given speech feature. It consists of a text embedding layer and 6 Transformer decoder layers with the same parameters as the encoder. The weight of each part of the given text and speech feature can be determined using multihead attention between the encoder–decoder.

4. Experimental Setup

In this section, we introduce the details of the speech datasets used to learn the SVR and ASR, as well as implementations, respectively.

4.1. Data

We collected approximately 18,000 h of audio for pretraining the SVR extractor for the practical Korean ASR system. The Korean government-sponsored AIHub website (https://aihub.or.kr/, accessed on 6 January 2023) has several open speech datasets [40–47]. To facilitate effective learning, we omit the samples shorter than two seconds, as they contain relatively long silent beginnings and endings without useful signals. Note that the speech datasets presented in Table 1 only include samples longer than 2 s.

The KsponSpeech [40] corpus is composed of 921 h of voice recordings from adult males and females, which amounts to 517,144 training examples. The age and regional breakdown of the participants were not considered. The Korean Foreign Word Speech (KForeignWordSpeech) [41] corpus consists of a total of 3155 h of audio from 2,484,843 speech files recorded by native Korean speakers in different noisy environments. In the Korean Command Speech (KCommandSpeech) and Korean Free Conversation Speech (KFreeConvSpeech) datasets, which consist of audio recordings from adults [42,43], children [44,45], and elderly [46,47] from different areas, there are 8072 h and 6242 h of audio recordings bearing 5,770,446 and 6,142,743 training samples, respectively. Both datasets contain the speech of native Korean users in the presence of noise. We combined the three age groups (adult, child, and elderly). We provide detailed specifications of speech datasets for SVR learning in Table 2.

We use a total of four datasets on AIHub website for pretraining the SVR extractor. To comprehend the impact of the amount of pretraining data on ASR performance, Table 3 details how we combined speech datasets for pretraining.

Table 1. Four Korean speech datasets for pretraining the SVR model.

Name	Hours	No. Utterances
KsponSpeech [40]	921	517,144
KForeignWordSpeech [41]	3155	2,484,843
KCommandSpeech (Adult) [42]	1718	1,742,211
KCommandSpeech (Child) [44]	2114	2,262,551
KCommandSpeech (Elderly) [46]	2410	2,137,981
KFreeConvSpeech (Adult) [43]	3186	2,235,385
KFreeConvSpeech (Child) [45]	2377	2,389,409
KFreeConvSpeech (Elderly) [47]	2511	1,145,652
Sum	18,392	14,915,176

Table 2. The detailed specification of Four Korean speech datasets.

Dataset	File Format	Sampling Frequency	Mode	Average Duration (s)	Max Duration (s)	No. Speakers (Male/Female)
[40]	pcm/wav	16 kHz	Mono	6.41	31.00	923/1077
[41]		44 and 16 kHz		4.57	24.96	1000/1000
[42]		48 kHz		3.55	21.42	1751/1751
[44]		48 kHz		3.36	24.18	1500/1500
[46]		48 kHz		4.06	24.90	1500/1500
[43]		44 and 16 kHz		5.13	24.20	1000/1000
[45]		44 and 16 kHz		3.58	17.32	500/500
[47]		44 and 16 kHz		7.88	24.99	500/500

Table 3. Different amount of pretraining data for SVR learning.

Name	Datasets	Hours	No. Utterances
SVR1K	[40]	921	517,144
SVR3K	[41]	3155	2,484,843
SVR4K	[40,41]	4076	3,001,987
SVR7K	[40,42,44,46]	7163	6,659,887
SVR9K	[40,43,45,47]	8995	6,287,590
SVR10K	[40–42,44,46]	10,318	9,144,730
SVR18K	[40–47]	18,392	14,915,176

We use two downstream ASR datasets for evaluation: KsponSpeech and KForeign-WordSpeech. In addition, to evaluate how the quantity of data used in the pretraining of the SVR extractor affects its performance, we define four conditions, as shown in Figure 3:

- Baseline condition: The ASR model trained using the mel filterbank that is directly converted from waveform (no pretraining).
- Matched condition: The same dataset is both utilized in the pretraining phase of the SVR model and the training phase of the ASR model using extracted features from the SVR model.
- Unmatched condition: The dataset that was employed for pretraining the SVR model and for training the ASR model with features extracted from the SVR model is disparate.
- Multi condition: The SVR model is pretrained with additional speech datasets, in addition to those used for ASR training. In other words, we pretrain the SVR model with more datasets than the ASR model.

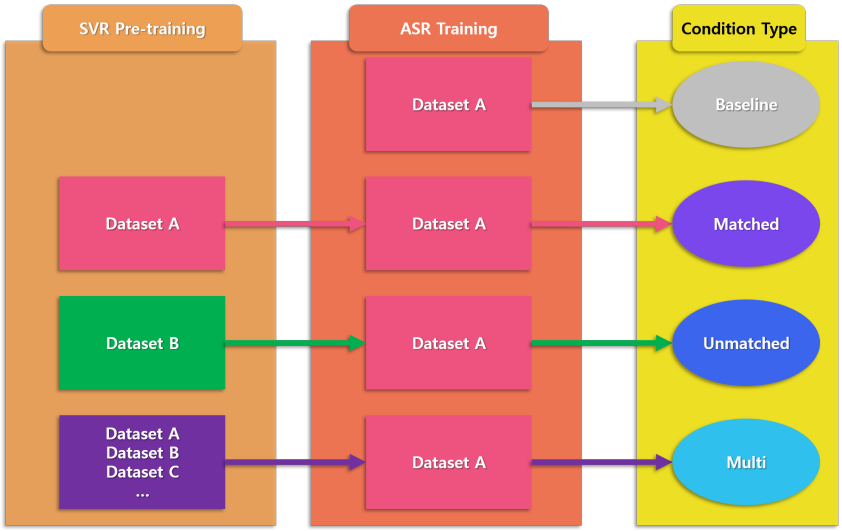

Figure 3. The illustrations of the proposed SVR model learning with unsupervised masking approach.

4.2. Pretraining Details

4.2.1. Masking

For the proposed masking strategy with the given log mel filterbank, we set P_T and P_F as 0.15 and 0.4, respectively. We use the consecutive masking parameter C as 7 for the time region. For the respective noisy swapping and Gaussian noise, we set P_S and P_N as 0.1 and 0.1 in every training sample. For the ablation studies, we further experiment with various combinations from $P_T = [0.1, 0.2, 0.3, 0.4]$, $P_F = [0.0, 0.1, 0.15, 0.2, 0.3, 0.4]$, and $P_N = [0.0, 0.1, 0.2, 0.3, 0.4, 0.5]$, respectively.

4.2.2. Optimization

For all the speech data in this work, we use 16 kHz quality. We further up- or down-sampled all the given audio files with the mismatching sampling rate used in our experiment to match the 16 kHz. For pretraining of the SVR model, we use an 80-dimensional log mel filterbank extracted with a 25 ms window and 10 ms overlap as the input. The SVR model is trained using a batch size of 16 on an NVIDIA RTX A6000 GPU with 48 GB of memory. We utilize 4 GPUs and AdamW optimizer [48] with $\beta = (0.9, 0.999)$ and $\epsilon = 0.00000001$, and the learning rate gradually increases from 0 until it reaches its peak of 0.0002 after 7% of the training steps are completed and then decays back to 0. Furthermore, we use gradient accumulation steps and gradient clipping as 4 and 5 to obtain the optimal model parameters that minimize the L1 loss for reconstruction.

We set the total pretraining epochs of the SVR model as 100 for all SVR experiments (SVR1K, SVR3K, . . . , and SVR18K). Table 4 presents the total training steps in relation to 100 epochs, based on the size of the pretraining dataset.

Table 4. Total pretraining steps of various dataset combinations until 100 epochs.

Name	Datasets	Hours	No. Utterances	Training Steps (100 epochs)
SVR1K	[40]	921	517,144	808,050
SVR3K	[41]	3155	2,484,843	3,882,575
SVR4K	[40,41]	4076	3,001,987	4,690,625
SVR7K	[40,42,44,46]	7163	6,659,887	10,406,075
SVR9K	[40,43,45,47]	8995	6,287,590	9,824,375
SVR10K	[40–42,44,46]	10,318	9,144,730	14,288,650
SVR18K	[40–47]	18,392	14,915,176	23,304,975

4.3. ASR Training Details

For training the ASR model with the pretrained SVR model, the SVR model was maintained in the frozen state, and the speech representations were extracted from its deepest layer, which is the hidden state of the final Transformer Encoder layer. We train the ASR model with supervision and a total batch size of 64 on 4 GPUs. We employ the same setting of AdamW optimizer in the pretraining stage, with a Transformer learning rate schedule as described in [37]. We use the label smoothing [39] parameter $\alpha = 0.1$ and Cross-Entropy loss to optimize the ASR model until 50 epochs. The sizes of the text embedding layers including special tokens (e.g., <sos>, <eos>, <mask>, <pad>, and <unknown>) in the ASR trained with KsponSpeech and KForeignWordSpeech are 2311 and 1784, respectively.

4.4. Software Details

In order to train both the SVR and ASR models, we utilized `Python` version 3.9.15 on a GPU server with `Ubuntu` 18.04.6 LTS. For deep learning frameworks and libraries, we employed `PyTorch` [49] version 1.12.1 with `CUDA` version 11.3 and `CuDNN` version 8.21. For speech preprocessing, we used `TorchAudio` [50] version 0.12.1, `Numpy` [51] version 1.23.5, and `SoundFile` version 0.11.0. To evaluate the speech recognition performance, we employed `python-Levenshtein` version 0.20.9 to use edit distance.

5. Experimental Results and Analysis

In this section, we present the end-to-end ASR model performance when trained with two speech datasets without an external language model or beam search, followed by ablation studies. To measure the performance of the ASR model, we use a character error rate (CER) as a metric.

5.1. Main Results

Firstly, we evaluate the ASR performance trained with the KsponSpeech benchmark based on various pretrained models. To this end, we perform pretraining of the respective six previous methods [20,22,24,29,36,52] and our proposed model with the KsponSpeech dataset (SVR1K), following the same settings described in the S3PRL Toolkit. (https://github.com/s3prl/s3prl, accessed on 5 January 2023). We only provide results in the frozen setting due to the computation costs of finding optimal hyperparameters and fine-tuning these pretrained models in the ASR tasks. For a fair comparison, we train until 100 epochs using a single GPU and set all extracted hidden sizes of the pretrained models to 768. Furthermore, all ASR models using extracted features from the frozen SVR model have the same settings as described in Section 4.3.

As shown in Table 5, the proposed pretrained SVR-based ASR model outperforms all its counterparts. Specifically, our proposed method achieved significantly better performance than the methods using only time region masking [24,36]. Furthermore, we observed that using the proposed noise masking in the pretraining stage outperforms the other pretrained models for practical ASR.

Table 5. Comparison of ASR performance based on different pretrained methods. All the counterparts are pretrained with KsponSpeech-SVR1K with S3PRL Toolkit settings. Recurrent and Parallel represent using RNN or Transformer-based neural networks for pretraining, respectively. Bold indicates the best performance.

Pretraining Methods	Network	No. Model Params	CER ↓ (%)
CPC [20]	Recurrent	12,931,584	13.94
APC [22]	Recurrent	9,107,024	14.78
NPC [52]	Recurrent	19,380,560	13.36
Mockingjay [24]	Parallel	22,226,928	16.95
AALBERT [36]	Parallel	7,805,264	17.25
TERA [29]	Parallel	21,981,008	13.86
Ours	Parallel	21,981,008	**12.32**

In Table 6, we present our results under each of the four conditions on KsponSpeech dev sets, using the whole dataset as the supervised data. Error reduction rate (ERR) in Table 6 provides an indication of the relative improvement of the CER from Matched, Unmatched, and Multi conditions compared with the Baseline condition.

Table 6. The ASR results of KsponSpeech (965 h) for the four conditions were trained with the whole dataset as the supervised data. The relative error with respect to the Baseline condition is also indicated. Bold denotes the best performance.

Conditions	Name	No. Unlabeled Data (h)	CER ↓ (%)	ERR ↑ (%)
Baseline	-	-	15.17	-
Matched	SVR1K	921	12.32	18.79
Unmatched	SVR3K	3155	13.18	13.10
Multi	SVR4K	4076	12.23	19.37
	SVR7K	7163	12.54	17.32
	SVR9K	8995	12.09	20.31
	SVR10K	10,318	12.38	18.38
	SVR18K	18,392	**11.72**	**22.77**

The ASR performance in the Baseline condition achieved a CER of 15.17%, which is the baseline for the ERR. Utilizing the pretrained SVR extractor for training the ASR model

in the matched condition (Matched-SVR1K), we obtain a CER of 12.32%. This indicates a significant improvement of 18.79% when compared with the Baseline condition. While we acquired a CER of 13.18% in the Unmatched condition (Unmatched-SVR3K), which is a higher value than that of the Matched condition (Unmatched-SVR1K) by 0.86%, it improved by 13.10% compared with the Baseline condition. We conjecture that this result, derived from the speech dataset used for the pretraining of the SVR model, differs in terms of its distribution compared with KsponSpeech. In addition, we observed that the ASR performance improved when we increased the amount of pretraining data (Multi-SVR4K, SVR9K, SVR18K). Especially, when using approximately 18,000 h of speech data for pretraining (Multi-SVR18K), the ASR outcomes demonstrated the best result in our experiments, with a CER of 11.72% and an ERR of 22.77%.

Table 7 illustrated the results of ASR experiments trained with 3155 h of the KForeignWordSpeech dataset, and we use the same settings of masking hyperparameters described in Section 4.2.1. The ASR performance without pretrained SVR extractor achieved a CER of 4.77%. We hypothesize that the better ASR performance in the Baseline condition in Table 7 is likely due to the larger size of KForeignWordSpeech, which is approximately three times that of KsponSpeech. In the Matched condition (Matched-SVR3K) ASR experiment, a CER of 1.09% was observed, which corresponds to an ERR of 18.79%, which is relatively reduced compared with the Baseline condition. In addition, we obtained a CER of 4.24% and an ERR of 11.11% in the Unmatched condition (Unmatched-SVR1K) ASR experiment. Surprisingly, this shows that despite using a significantly lower amount of data for pretraining the SVR model than the ASR experiment, we found that it yields improved results compared with the Baseline condition. Comparison between the three Multi condition ASR experiments (Multi-SVR4K, Multi-SVR10K, Multi-SVR18K) showed that each ASR performance is nearly the same.

Table 7. The ASR results of KForeignWordSpeech (3155 h) for the four conditions trained with the whole dataset as the supervised data. Bold indicates the best performance.

Conditions	Name	No. Unlabeled Data (h)	CER ↓ (%)	ERR ↑ (%)
Baseline	-	-	4.77	-
Matched	SVR3K	3155	1.09	77.15
Unmatched	SVR1K	921	4.24	11.11
Multi	SVR4K	4076	0.98	79.45
	SVR10K	10,318	0.95	80.08
	SVR18K	18,392	**0.9**	**81.13**

As a result of the experiment, the ASR performance using our unsupervised pretrained SVR extractor showed a significant improvement between the Baseline condition and the three other conditions. Moreover, two experiments validate the effectiveness of the proposed SVR model, that is, the ASR performance of Multi condition with SVR18K outperforms the baseline in various noisy environments.

5.2. Ablation: Impact of Time Masking Hyperparameter

In this section, we conduct an ablation study to understand the effectiveness of time region masking in the pretraining of the SVR model. Therefore, we investigate the effects of different time region masking hyperparameters P_T on ASR performance. To this end, other hyperparameters for the pretraining of the SVR model, such as P_F, P_S, P_N, and C, are fixed as 0.2, 0.1, 0.1, and 7. Table 8 shows the results of KsponSpeech ASR performance (Matched-SVR1K) trained with the extracted features from SVR models pretrained with various P_T.

Table 8. ASR Performance comparison according to the variations of P_T in the pretraining of the SVR extractor. All the results are from training the ASR model using the pretrained SVR extractor kept frozen. We report on the KsponSpeech ASR performance (Matched-SVR1K). The respective bolds represent the best hyperparameter and performance.

P_T	P_F	P_S	P_N	CER ↓ (%)	ERR ↑ (%)
0.1	0.2	0.1	0.1	13.87	8.57
0.15				13.86	8.64
0.2				13.17	13.18
0.3				**12.94**	**14.70**
0.4				13.45	11.33

As demonstrated in Table 8, we keep the pretrained SVR model frozen and use its representations to train the ASR model with the KsponSpeech dataset. In Table 8, the best performance is obtained when P_T is 0.3. This indicates that while all the results outperform the Baseline condition in Table 6, using a small P_T is not effective when P_F is set as 0.2. In particular, we found that the practical ASR results are degraded when $P_T = 0.15$ is used to pretrain the SVR model, as in previous studies [24,29,33,36].

5.3. Ablation: Impact of Frequency Masking Hyperparameter

To analyze the effect of frequency region masking in the pretrained SVR extractor, we conducted a second ablation study with different values of $P_F = [0.0, 0.1, 0.15, 0.2, 0.3, 0.4]$. During this ablation study, the remaining masking hyperparameters stayed fixed at $P_T = 0.15$, $P_S = 0.1$, $P_N = 0.1$, and $C = 7$. We also keep the pretrained SVR model frozen and use it for training on the KsponSpeech dataset.

As shown in Table 9, we acquired the best ASR performance when P_F was 0.4. When the $P_F = 0.0$ was used in the SVR model pretraining, we obtained a CER of 15.31%, which is worse than the Baseline condition in Table 6. As a result, we found that as P_F is bigger; the practical ASR seems to be more robust. In addition, we demonstrate that frequency region masking can significantly help to obtain better ASR performance.

Table 9. ASR Performance comparison according to the variations of P_F in the pretraining of the SVR model. All the results are from training the ASR model using the pretrained SVR extractor kept frozen. We report on the KsponSpeech ASR performance (Matched-SVR1K). The respective bolds denote the best hyperparameter and performance.

P_T	P_F	P_S	P_N	CER ↓ (%)	ERR ↑ (%)
0.15	0	0.1	0.1	15.31	−0.92
	0.1			13.85	8.70
	0.15			13.80	9.03
	0.2			13.86	8.64
	0.3			13.24	12.72
	0.4			**12.32**	**18.79**

5.4. Ablation: Impact of Two Noise Masking Hyperparameters to Perturb the Speech

To investigate the impact of the two proposed noise masking tactics used in the masking procedure for SVR model training, we further consider two setups: variations of P_N and P_S. We froze the masking hyperparameters as $P_T = 0.15$, $P_F = 0.2$, and $C = 7$. In the first setup, we use $P_N = [0.0, 0.1, 0.2, 0.3, 0.4, 0.5]$ and fix the other masking hyperparameters for training the SVR model. The other one is to freeze the first setup and change the $P_S = 0.0$ to learn the SVR model. This process creates 12 experiments, which are presented in Table 10.

Table 10. ASR Performance comparison according to the variations of P_S and P_N in the pretraining of the SVR model. All the results are from training the ASR model using the pretrained SVR extractor kept frozen. We report on the KsponSpeech ASR performance (Matched-SVR1K). The respective bolds represent the best hyperparameter and performance.

P_T	P_F	P_S	P_N	CER ↓ (%)	ERR ↑ (%)	Avg CER ↓ (%)
0.15	0.2	0.0	0.0	13.74	9.43	13.65
			0.1	**13.08**	**13.78**	
			0.2	13.23	12.79	
			0.3	13.81	8.97	
			0.4	14.16	6.66	
			0.5	13.90	8.37	
		0.1	0.0	13.86	8.64	**13.34**
			0.1	**12.98**	**14.44**	
			0.2	13.13	13.45	
			0.3	13.58	10.48	
			0.4	13.14	13.38	
			0.5	13.36	11.93	

When the $P_N = 0.0$ setting was used in the $P_S = 0.0$ experiments, we acquired a CER of 13.08%, which is the best performance in the same group. Even in the experiment where $P_S = 0.1$, the best CER of 12.98% was obtained in the second group when $P_N = 0.1$, and this result is similar to $P_S = 0.0$. We observe that including too much ($P_N = 0.5$) or no ($P_N = 0.0$) noise masking degrades the SVR model training. As a result, we can obtain that training the SVR model with $P_S = 0.1$ outperforms $P_S = 0.0$ in average CER performance by 0.31%. As a result, the experimental results in Table 10 show that using the two proposed noise masking methods for SVR model pretraining is effective.

6. Discussion

The goal of this article is to create a pretrained SVR model that can provide strong ASR performance in real-world applications. To achieve this, we investigated a wide range of approximately 18,000 h of Korean speech and proposed a novel unsupervised SVR learning method with a task-agnostic feature masking method. To the best of our knowledge, this is the first attempt at pretraining an SVR model with large-scale real Korean speech. The results show that the proposed SVR model outperformed previous similar methods, including those using only time region masking. Additionally, we observed that incorporating noise masking during the pretraining stage further improves the model's performance for practical ASR.

Furthermore, we explored how the quantity of data used in the pretraining of the SVR model impacts its ASR performance. As shown in Table 6, we found that increasing the amount of pretraining data improved the ASR performance (Multi-SVR4K, Multi-SVR9K, Multi-SVR18K). Interestingly, these findings indicated that the proposed SVR model is effective even when the amount of data used for training is limited. In particular, using approximately 18,000 h of speech data for pretraining (Multi-SVR18K) resulted in the best performance, with a CER of 11.72% and an ERR of 22.77% in the KsponSpeech ASR experiments. Surprisingly, we observed that the proposed method outperformed the baseline, even when the domains are different (Unmatched-SVR3K).

On the other hand, comparing the ASR performance of the three Multi conditions of Table 7 (Multi-SVR4K, Multi-SVR10K, Multi-SVR18K) showed that the performance was similar across all three Multi conditions. When the amount of data used for ASR training is sufficient, the results show that the Multi condition-based pretrained SVR model can perform similarly or outperform the Matched condition. Furthermore, we discovered that even when the amount of data used for ASR training is limited, the proposed SVR

method outperforms the Baseline condition. In light of these results, it can be inferred that the proposed SVR model is able to extract generalizable speech feature key points, which contribute to its strong task-agnostic ASR system.

Our findings in this paper demonstrate that the proposed SVR model can effectively be used as a speech feature extractor for robust Korean speech recognition tasks.

7. Conclusions

In this paper, we present a novel unsupervised approach for learning the SVR extractor for robust end-to-end speech recognition. The SVR model can learn high-level speech representations and deal with various environments using our proposed noise masking strategy for unlabeled large-scale data. We showed that the ASR model performance using our pretrained SVR as a frozen feature extractor significantly outperforms conventional methods in real-world conditions. Furthermore, we demonstrated how the quantity of data used in the pretraining of the SVR extractor affects the generalization of task-agnostic ASR.

In conclusion, this paper focused on offline ASR services in real-world scenarios. For online or streaming ASR services in practical applications, however, RNN-Transducer [53–55] and streaming [56–59] technology are needed. Therefore, further studies will be required to address this in the future. In addition, we will apply temporal algorithms using our pretrained SVR model to predict future frames in order to obtain the best task-specific performance. We will leave it as future work.

Author Contributions: Conceptualization, J.-W.K.; methodology, J.-W.K.; software, J.-W.K. and H.-Y.J.; validation, J.-W.K., H.C. and H.-Y.J.; formal analysis, J.-W.K.; investigation, J.-W.K.; resources, J.-W.K. and H.-Y.J.; data curation, J.-W.K. and H.-Y.J.; writing—original draft preparation, J.-W.K.; writing—review and editing, J.-W.K., H.C. and H.-Y.J.; visualization, J.-W.K.; supervision, J.-W.K. and H.-Y.J.; project administration, J.-W.K., H.C. and H.-Y.J.; funding acquisition, H.C. and H.-Y.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Research Foundation (NRF), Korea, under project BK21 FOUR, in part by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-2020-0-01808) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation), and in part by the MSIT through the IITP Program (Development of Semi-Supervised Learning Language Intelligence Technology and Korean Tutoring Service for Foreigners) under Grant 2019-0-00004.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The KsponSpeech dataset is available at [40], the KForeignWordSpeech dataset is available at [41], the KCommandSpeech (Adult) dataset is available at [42], the KFreeConvSpeech (Adult) dataset is available at [43], the KCommandSpeech (Child) dataset is available at [44], the KFreeConvSpeech (Child) dataset is available at [45], the KCommandSpeech (Elderly) dataset is available at [46], and the KFreeConvSpeech (Elderly) dataset is available at [47].

Acknowledgments: The authors would like to the reviewers for their helpful feedback and discussion on this document. The authors fully appreciate Miika Toikkanen for the helpful feedback and discussion for this paper.

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

References

1. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust speech recognition via large-scale weak supervision. *arXiv* **2022**, arXiv:2212.04356.
2. Garofolo, J.; Lamel, L.; Fisher, W.; Fiscus, J.; Pallet, D.; Dahlgren, N. *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM*; Linguistic Data Consortium: Philadelphia, PA, USA, 1993.
3. Garofolo, J.; Graff, D.; Paul, D.; Pallett, D. *CSR-I (WSJ0) Complete LDC93S6A*; Web Download; Linguistic Data Consortium: Philadelphia, PA, USA, 1993.

4. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An asr corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; pp. 5206–5210.
5. Kahn, J.; Rivière, M.; Zheng, W.; Kharitonov, E.; Xu, Q.; Mazaré, P.E.; Karadayi, J.; Liptchinsky, V.; Collobert, R.; Fuegen, C.; et al. Libri-light: A benchmark for asr with limited or no supervision. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7669–7673.
6. Gong, Y. Speech recognition in noisy environments: A survey. *Speech Commun.* **1995**, *16*, 261–291. [CrossRef]
7. Rajnoha, J.; Pollák, P. ASR systems in noisy environment: Analysis and solutions for increasing noise robustness. *Radioengineering* **2011**, *20*, 74–84.
8. Li, J.; Deng, L.; Gong, Y.; Haeb-Umbach, R. An overview of noise-robust automatic speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 745–777. [CrossRef]
9. Potamianos, A.; Narayanan, S.; Lee, S. Automatic speech recognition for children. In Proceedings of the Fifth European Conference on Speech Communication and Technology, Rhodes, Greece, 22–25 September 1997.
10. Potamianos, A.; Narayanan, S. Robust recognition of children's speech. *IEEE Trans. Speech Audio Process.* **2003**, *11*, 603–616. [CrossRef]
11. Wilpon, J.G.; Jacobsen, C.N. A study of speech recognition for children and the elderly. In Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Atlanta, GA, USA, 9 May 1996; Volume 1, pp. 349–352.
12. Anderson, S.; Liberman, N.; Bernstein, E.; Foster, S.; Cate, E.; Levin, B.; Hudson, R. Recognition of elderly speech and voice-driven document retrieval. In Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings ICASSP99 (Cat. No. 99CH36258), Phoenix, AZ, USA, 15–19 March 1999; Volume 1, pp. 145–148.
13. Kim, J.W.; Yoon, H.; Jung, H.Y. Linguistic-Coupled Age-to-Age Voice Translation to Improve Speech Recognition Performance in Real Environments. *IEEE Access* **2021**, *9*, 136476–136486. [CrossRef]
14. Shrawankar, U.; Thakare, V. Adverse Conditions and ASR Techniques for Robust Speech User Interface. *Int. J. Comput. Sci. Issues (IJCSI)* **2011**, *8*, 440.
15. Chavan, K.; Gawande, U. Speech recognition in noisy environment, issues and challenges: A review. In Proceedings of the 2015 International Conference on Soft-Computing and Networks Security (ICSNS), Coimbatore, India, 25–27 February 2015; pp. 1–5.
16. Weintraub, M.; Taussig, K.; Hunicke-Smith, K.; Snodgrass, A. Effect of speaking style on LVCSR performance. In Proceedings of the 4th International Conference on Spoken Language Processing, Philadelphia, PA, USA, 3–6 October 1996; Volume 96, pp. 16–19.
17. Benzeghiba, M.; De Mori, R.; Deroo, O.; Dupont, S.; Erbes, T.; Jouvet, D.; Fissore, L.; Laface, P.; Mertins, A.; Ris, C.; et al. Automatic speech recognition and speech variability: A review. *Speech Commun.* **2007**, *49*, 763–786. [CrossRef]
18. Young, V.; Mihailidis, A. Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assist. Technol.* **2010**, *22*, 99–112. [CrossRef] [PubMed]
19. Kim, J.W.; Yoon, H.; Jung, H.Y. Improved Spoken Language Representation for Intent Understanding in a Task-Oriented Dialogue System. *Sensors* **2022**, *22*, 1509. [CrossRef] [PubMed]
20. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
21. Schneider, S.; Baevski, A.; Collobert, R.; Auli, M. wav2vec: Unsupervised pretraining for Speech Recognition. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 3465–3469. [CrossRef]
22. Chung, Y.A.; Hsu, W.N.; Tang, H.; Glass, J. An Unsupervised Autoregressive Model for Speech Representation Learning. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 146–150. [CrossRef]
23. Ling, S.; Liu, Y.; Salazar, J.; Kirchhoff, K. Deep contextualized acoustic representations for semi-supervised speech recognition. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6429–6433.
24. 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; pp. 6419–6423.
25. Wang, W.; Tang, Q.; Livescu, K. Unsupervised pretraining of bidirectional speech encoders via masked reconstruction. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6889–6893.
26. Park, D.S.; Zhang, Y.; Jia, Y.; Han, W.; Chiu, C.C.; Li, B.; Wu, Y.; Le, Q.V. Improved Noisy Student Training for Automatic Speech Recognition. *Proc. Interspeech 2020* **2020**, 2817–2821. [CrossRef]
27. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12449–12460.
28. Zhang, Y.; Qin, J.; Park, D.S.; Han, W.; Chiu, C.C.; Pang, R.; Le, Q.V.; Wu, Y. Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition. In Proceedings of the NeurIPS 2020 workshop: Self-Supervised Learning for Speech and Audio Processing, Virtual Conference, 11 December 2020; arXiv:2010.10504.
29. Liu, A.T.; Li, S.W.; Lee, H.Y. Tera: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 2351–2366. [CrossRef]

30. Hsu, W.N.; Bolte, B.; Tsai, Y.H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3451–3460. [CrossRef]
31. Chung, Y.A.; Zhang, Y.; Han, W.; Chiu, C.C.; Qin, J.; Pang, R.; Wu, Y. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pretraining. In Proceedings of the 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, 13–17 December 2021; pp. 244–250.
32. Rivière, M.; Dupoux, E. Towards unsupervised learning of speech features in the wild. In Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 19–22 January 2021; pp. 156–163.
33. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1 (Long and Short Papers), pp. 4171–4186. [CrossRef]
34. 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. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 2613–2617. [CrossRef]
35. Baevski, A.; Schneider, S.; Auli, M. vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
36. 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. In Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 19–22 January 2021; pp. 344–350.
37. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
38. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376.
39. Müller, R.; Kornblith, S.; Hinton, G.E. When does label smoothing help? In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
40. Bang, J.U.; Yun, S.; Kim, S.H.; Choi, M.Y.; Lee, M.K.; Kim, Y.J.; Kim, D.H.; Park, J.; Lee, Y.J.; Kim, S.H. Ksponspeech: Korean spontaneous speech corpus for automatic speech recognition. *Appl. Sci.* **2020**, *10*, 6936. [CrossRef]
41. AIHub. Korean Foreign Word Speech. Available online: https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=131 (accessed on 6 January 2023).
42. AIHub. Korean Command Speech (Adult). Available online: https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=96 (accessed on 6 January 2023).
43. AIHub. Korean Free Conversation Speech (Adult). Available online: https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=109 (accessed on 6 January 2023).
44. AIHub. Korean Command Speech (Child). Available online: https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=95 (accessed on 6 January 2023).
45. AIHub. Korean Free Conversation Speech (Child). Available online: https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=108 (accessed on 6 January 2023).
46. AIHub. Korean Command Speech (Elderly). Available online: https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=94 (accessed on 6 January 2023).
47. AIHub. Korean Free Conversation Speech (Elderly). Available online: https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=107 (accessed on 6 January 2023).
48. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
49. 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. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
50. Yang, Y.Y.; Hira, M.; Ni, Z.; Astafurov, A.; Chen, C.; Puhrsch, C.; Pollack, D.; Genzel, D.; Greenberg, D.; Yang, E.Z.; et al. Torchaudio: Building blocks for audio and speech processing. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 6982–6986.
51. Harris, C.R.; Millman, K.J.; Van Der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362. [CrossRef] [PubMed]
52. Liu, A.H.; Chung, Y.A.; Glass, J. Non-Autoregressive Predictive Coding for Learning Speech Representations from Local Dependencies. In Proceedings of the Interspeech 2021, Brno, Czech Republic, 30 August–3 September 2021; pp. 3730–3734. [CrossRef]
53. Graves, A. Sequence transduction with recurrent neural networks. *arXiv* **2012**, arXiv:1211.3711.
54. Rao, K.; Sak, H.; Prabhavalkar, R. Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017; pp. 193–199.

55. Chen, X.; Wu, Y.; Wang, Z.; Liu, S.; Li, J. Developing real-time streaming transformer transducer for speech recognition on large-scale dataset. 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. 5904–5908.
56. Kim, K.; Lee, K.; Gowda, D.; Park, J.; Kim, S.; Jin, S.; Lee, Y.Y.; Yeo, J.; Kim, D.; Jung, S.; et al. Attention based on-device streaming speech recognition with large speech corpus. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 956–963.
57. He, Y.; Sainath, T.N.; Prabhavalkar, R.; McGraw, I.; Alvarez, R.; Zhao, D.; Rybach, D.; Kannan, A.; Wu, Y.; Pang, R.; et al. Streaming end-to-end speech recognition for mobile devices. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 April 2019; pp. 6381–6385.
58. Moritz, N.; Hori, T.; Le, J. Streaming automatic speech recognition with the transformer model. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6074–6078.
59. Shi, Y.; Wang, Y.; Wu, C.; Yeh, C.F.; Chan, J.; Zhang, F.; Le, D.; Seltzer, M. Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition. 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. 6783–6787.