Self-Supervised Learning With Segmental Masking for Speech Representation

Xianghu Yue, Jingru Lin, Fabian Ritter Gutierrez, and Haizhou Li, *Fellow, IEEE*

***Abstract*—Self-supervised learning has achieved remarkable success for learning speech representations from unlabeled data. The masking strategy plays an important role in the self-supervised learning algorithm. Most of the masking techniques operate at a frame level. In linguistics, phone is the smallest unit of sound. Hence, we believe that a masking technique that operates at a phoneme level will effectively encode the phonotactic and prosodic constraints of a spoken language, thus eventually benefits the downstream speech recognition tasks. In this work, we explore a novel segmental masking strategy. Specifically, we mask phonetically motivated speech segments according to the phonetic segmentation in an utterance. By doing so, we implicitly incorporate the properties of a spoken language, such as phonotactic constraints and duration of phonetic segments, into the pre-training. Through extensive experiments, we confirm that the segmental masking strategy consistently outperforms the frame-based masking counterpart. We also further investigate the effect of segmental masking unit size, i.e. phoneme, phoneme span, and lexical word. This work presents an important finding about masking strategy in speech representation learning.**

***Index Terms*—Self-supervised learning, speech representation learning, segmental masking.**

I. Introduction

CHILDREN learn a spoken language by first listening to unlabeled speech continuously to figure out the phonetic and phonotactic structure of a language [1]–[3], then further acquiring lexical and syntactic knowledge to associate with meanings. The former is a typical self-supervised learning (SSL) process [4]–[6], while the latter is achieved via supervised learning.

The recent studies on representation learning of speech signals mimics the unsupervised process of language acquisition by first pre-training a model on a large amount of unlabeled speech data to capture speech dynamics, then fine-tuning the model over a specific downstream task. Benefiting from abundant unlabeled data, this paradigm has shown to be effective for improving many speech-related tasks, such as automatic speech recognition (ASR) [4], [5], [7], speaker verification (SV) [8], [9], speech translation (ST) [10], [11] and speech enhancement (SE) [12], [13]. As the self-supervised learning seeks to discover useful feature representations from unlabeled data, the resulting representations are expected to be more general and robust than those derived from supervised learning, which tend to bias towards downstream applications [14]. Self-supervised learning is therefore an important stepping stone for learning robust and generic representations [6].

Speech signals carry multiple levels of information, e.g. acoustic, phonetic, prosodic, and linguistic units, at different time scale, e.g. short-time frame, phoneme, word, phrase, and sentence [15], [16]. Ideally, a learned model is expected to characterize the speech signals in a way that is more accessible to the downstream tasks. For example, it should capture phonetic, phonotactic, and prosodic knowledge in a speech signal that is useful for a speech recognition task. Furthermore, such learned model could also help improve transfer learning and adaptations across different data distributions and domains [17]–[19].

In practice, the pretext task is at the centre of a self-supervised learning algorithm [4]–[6]. Solving the pretext task, a neural network learns a mapping function that transforms the input speech into latent representations that are potentially useful for various downstream tasks. Therefore, the design of the pretext task plays a crucial role in self-supervised learning. The pretext task should be designed to encourage the model to learn the underlying representations of speech.

Just like the Masked Language Model (MLM) of BERT [20], [21] for language representations, the BERT-style masked reconstruction is studied for learning speech representations [22], [23]. Mockingjay [24] introduces the reconstruction loss for speech, in which it randomly masks the input speech frames into zero to pre-train the Transformer encoder [25] with a masking policy similar to BERT [20] and RoBERTa [21]. Audio ALBERT [26] explores a lite version of the self-supervised speech representation model based on Mockingjay [24]. TERA [5] is another extended version of Mockingjay, where the alterations are introduced and applied on inputs along three dimensions: time, channel and magnitude. In [27], the input features are divided into chunks of four frames, and masks are applied on the chunks at a masking rate of 15%. The above masking strategies directly borrow ideas from natural language processing (NLP),

Manuscript received 31 January 2022; revised 13 May 2022; accepted 4 July 2022. Date of publication 18 July 2022; date of current version 14 October 2022. The work of Haizhou Li was supported by the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen under Grant B10120210117-KP02. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Hung-Yi Lee. *(Corresponding author: Xianghu Yue.)*

Xianghu Yue, Jingru Lin, and Fabian Ritter Gutierrez are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077 (e-mail: xianghu.yue@u.nus.edu; jingru@nus.edu.sg; fabianritterg@nus.edu.sg).

Haizhou Li is with the School of Data Science, The Chinese University of Hong Kong, Shenzhen 518172, China, also with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077, and also with the Kriston AI, Xiamen 361000, China (e-mail: haizhouli@cuhk.edu.cn).

Digital Object Identifier 10.1109/JSTSP.2022.3191845

and have shown effectiveness in speech processing by replacing a discrete text symbol with a continuous speech frame. However, speech and text are inherently different. Speech is a continuous flow of signal, while text symbols are discrete. Moreover, in linguistics, phone is the smallest unit of sound. Speech is seen as a sequence of phones than short-time frames. Thus, it makes more sense for speech-related downstream tasks to have a phoneme as the minimal unit than a speech frame during pre-training.

We herein propose a novel segmental masking strategy for learning speech representations. Specifically, we use phoneme-like speech segments instead of speech frames as the masking units during the self-supervised pre-training. In this way, we expect the model to learn the phonetic sequence, i.e. phonotactic constraint, of a spoken language. With a phoneme-like speech segment, we also take into consideration the phonetic duration. Therefore, the model also learns the prosodic constraint at the same time. We expect that the segmental masking strategy will eventually benefit the downstream speech recognition tasks. Furthermore, motivated by spanBERT [28], we would study the effect of the segmental masking unit size. To the best of our knowledge, this is the first work that investigates the impacts of phonetically motivated masking units on self-supervised speech representation learning.

We demonstrate the effectiveness of the proposed masking strategies on various downstream tasks, which include phoneme classification, speech recognition and keyword spotting, speaker identification, intent classification, and speech emotion recognition. In addition, we also explore the use of different acoustic features, e.g., Mel Frequency Cepstral Coefficients (MFCC), filter-bank (FBANK) features and spectrograms, for reconstruction-based speech representation learning. Furthermore, we extend our masking strategies for contrastive predictive coding (CPC) [6]. Unlike previous CPC works [6], [18], that uses an autoregressive model to predict future frames based on previous context, we explore bidirectional Transformer encoder [25] instead.

This paper is a substantial extension to the study in [29]. We investigate a general segmental masking framework, and provide a comprehensive study of masking units for speech representation learning. We also carry out comprehensive experiments and analysis on large datasets that provide insights into the segmental masking strategies. Furthermore, we extend the proposed segmental masking to CPC. The rest of this paper is organized as follows. In Section II, we introduce the related work to set the stage for our research. In Section III, we discuss the phonotactic and prosodic constraints in self-supervised speech representation learning, and formulate the segmental masking strategies. The experimental setup is provided in Section IV. In Section V, we present and discuss the experimental results. Lastly, we conclude this paper in Section VI.

II. Related Work

Let's review three related studies, that are self-supervised learning, speech representation learning, and segmental structure of speech, to set the stage of this work.

A. *Self-Supervised Learning*

Supervised learning as a deep learning technique has brought phenomenal advances to various research areas such as image recognition [30], [31], speech recognition [32]–[34] and machine translation [25], [35], [36]. Such technique relies heavily on the quality and quantity of annotated datasets for particular applications, that oftentimes are not abundantly available. Furthermore, supervised learning also suffers from issues such as generalization error, domain mismatch, spurious correlations, and adversarial attacks [37]. Therefore, unsupervised or semi-supervised learning strategies become attractive alternatives. In particular, self-supervision allows a model to learn the natural characteristics of the data itself or the relationships between different modalities, and hence capitalise on the raw data without the need of manual annotations.

Self-supervised learning is a form of unsupervised learning that treats the input or modifications of the input as learning targets (pretext tasks). In doing so, it takes advantage of abundantly available unlabeled data for training [4], [5], [20], [38]. Specifically, in Computer Vision (CV), some studies incorporate contrastive objective and self-supervised learning for learning visual representations [38], [39]. In NLP, people exploit self-supervised learning to learn powerful language representations. ELMo [40] is the first work that introduces the concept of contextualized word representations using a bidirectional language model. BERT [20] first introduces the concept of masked language model (MLM) with deep Transformer encoder [25] architecture. XLNet [41], built with different attention strategies, outperforms both autoregressive models and MLM. RoBERTa [21], a BERT model with more data, larger batch size, and better hyperparameters, shows competitive results with XLNet in various downstream tasks. ALBERT [42] reduces the parameters drastically without losing performance compared to BERT [20]. We further the study of masking strategy of MLM in this paper in the context of speech representation.

B. *Speech Representation Learning*

The self-supervised learning algorithms for speech representations can be generally grouped into two strands: discriminative and generative. Generative approaches learn representations by reconstructing the input speech data or predicting withheld parts of the data, while discriminative approaches learn discriminative representations directly, often based on metric learning-based objectives.

1) Discriminative Approach: Contrastive Predictive Coding (CPC) [6] is one of the typical discriminative approaches, that combines contrastive loss and predictive coding for learning speech representations. With CPC, we use an autoregressive model to learn representations by conditioning on the past context to discriminate the future frames from the negative samples, that is, the contrastive loss is used to pull temporally nearby representations closer and push temporally distant ones further. Following this work, [19] further explores CPC to learn robust and multilingual speech representations, and [18] uses CPC to learn representations that can transfer well across languages for low-resource scenarios.

Wav2vec [43] is a fully convolutional network that is trained with a self-supervised learning strategy. In wav2vec [43], the network is first optimized by the CPC loss, then used for speech recognition. VQ-wav2vec [44] learns discrete speech representation through a two-stage pre-training pipeline. Wav2vec2.0 [4] introduces a framework for self-supervised learning of speech representations which masks latent representation of the raw waveform and solves a contrastive task over the quantized speech representations. It jointly learns discrete speech units with contextualized representations. Motivated by MLM [20], w2v-BERT [45] combines contrastive learning and MLM for speech representation learning, where the former trains the model to discretize continuous speech signal input and the latter trains the model to learn contextualized speech representations via solving a masked prediction task consuming the discretized tokens.

The discriminative approaches are generally effective. However, their masking schemes are mostly based on speech frames, without taking segmental information into consideration.

2) Generative Approach: Similar to Recurrent Neural Network (RNN) LM for text, Autoregressive Predictive Coding (APC) technique [16], [17] learns generic speech representations. It incorporates an autoregressive model to encode the temporal information within an acoustic utterance for reconstructing the target frame conditioning on previous context. To encourage APC to learn more global structures of speech sequences, the APC model is trained to predict a frame that is a few steps ahead of the current frame. In [46], the APC objective is extended to multi-target training. The new objective predicts not only the future frame conditioning on previous context but also past memory through reconstruction. In VQ-APC [47], VQ layer is combined to impose a bottleneck and force the model to learn better speech representations. DeCoAR [48] and DeCoAR 2.0 [49] further the APC technique and study deep contextualized acoustic representations.

Mockingjay [24], Audio ALBERT [26] and TERA [5] learn speech representations using bidirectional Transformer encoder [25], in which the model is trained to reconstruct the current frame through jointly conditioning on both past and future context. These BERT-style masked reconstruction methods are largely inspired from MLM from BERT [20], and adapt the NLP pre-training techniques to continuous speech. They follow similar masking policy to BERT [20] and RoBERTa [21]. [22], [23], [27] follow the standard BERT masking policy to pre-train the ASR encoder, in which 15% of the input frames are randomly chosen to be masked. The chosen frames are replaced with zero vectors for 80% of the time, with frames from random positions 10% of the time, and kept unchanged for remaining 10% of the time. In Speech-XLNet [50], motivated by Permutation Language Modeling (PLM) from XLNet [41], the model learns by reconstructing from shuffled input speech frames rather than masked frames. PASE [51] and PASE+ [7] explore to learn problem-agnostic speech representation from multiple self-supervised tasks, including reconstruction of raw waveform, Low Power Spectrum (LPS), MFCC and prosody and other binary classification tasks.

Similar to the discriminative approaches, the generative approaches also operate with masking schemes based on speech frames. The main difference lies in the manner in which they optimize the model: while generative approaches attempt to reconstruct masked frames, discriminative approaches learn to distinguish masked frames from negative distractors.

The studies of both discriminative and generative suggest that the choice of masking schemes matters in the speech representation learning. In this work, we take two typical discriminative methods, wav2vec [43] and CPC [6], and two typical generative methods, Mockingjay [24] and TERA [5] as the reference baselines.

C. Segmental Structure of Speech

In linguistics, phone is the smallest unit of sound. A sequence of linguistic units, e.g. phonemes, words, and phrases, form a speech signal. Thus, it makes more sense for a model to predict phonemes than speech frames in a running speech. Phoneme segmentation is an important precursor task for many speech processing tasks. Phoneme boundary detection has been explored under both supervised and unsupervised settings [52]–[55]. In [53], self-supervised learning (SSL) is used for phoneme boundary detection, in which the model is optimized to identify spectral changes in the signal using the Noise-Contrastive Estimation (NCE) principle and achieves state-of-the-art performance.

There have been studies on the segmental structure of speech, such as SCPC [56], ACPC [57], and mACPC [58]. SCPC [56] aims at performing phoneme segmentation with self-supervised learning, however, the training is still done at a frame level. ACPC [57] is a modification of CPC which predicts a sequence of latent representations instead of a single future latent vector. mACPC [58] is built on top of [56], [57] by using a frame-level encoder and a segment-level encoder that follows the same principle as that in the encoder in SCPC [56]. Unlike these methods, we directly incorporate segmental structures of speech, e.g. phonemeand word, into the masking strategy of self-supervised speech representation learning, and implicitly inject the phonotactic and prosodic constraints of a spoken language into the pre-training.

D. Contribution of This Work

Generally, the discriminative approach learns the speech representations by distinguishing masked positive samples from negative distractors, while the generative approach learns the representations by predicting the masked samples. Therefore, the masking strategy plays a vital role in the Masked Speech Model (MSM) [4], [5], [24]. A frame-level random sampling strategy was studied that treats a speech frame as the masking unit, similar to a token of text in Masked Language Model [20], [21]. This work is a departure from the prior work in several ways. First, we propose the use of phonetically motivated segment, instead of speech frame, as the masking unit, and show that segmental masking is more effective than masking random speech frames. Second, as the segmental masking involves multiple consecutive speech frames, we further study the effect of masking unit size, e.g., phoneme span and lexical word. Third, we apply the proposed segmental masking strategy to both generative and discriminative approaches using the Transformer

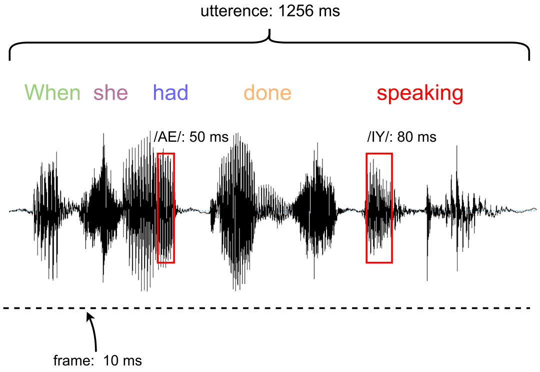

Fig. 1. The hierarchical structure of speech. The waveform is sampled at 16KHz from *train-clean-100* subset of LibriSpeech. A speech frame is a short-time window of 10 ms, which serves as a masking unit in many speech pre-training algorithms. A phoneme is a speech segment of multiple consecutive frames, as illustrated in red boxes. A sentence is typically over one second.

encoder [25]. Finally, we validate the proposed segmental masking strategy on various downstream speech recognition tasks to conclude the study.

III. Self-Supervised Learning With Segmental Masking

We can describe a spoken language by its phonotactic constraint, that dictates how phonemes sequence form syllables and words in a language. In a frame-based masking strategy, the self-supervised learning model learns to predict a frame from its context. In this paper, we propose a segmental masking strategy, with which the self-supervised learning model learns to predict a segment of frames instead.

A. Segmental Masking

For both discriminative and generative approaches, the masking strategy plays an important role. The masking strategy decides what the model learns and what effect of the model has on the downstream tasks.

Typically, a frame-based masking scheme firstly selects 15% of frames, and then 1) masks 80% of the time to zero, 2) replaces 10% of the time with a random frame, 3) leaves 10% of the time unchanged, simply adopting the masking strategy in BERT [20], [21]. While effective, this simple random masking scheme treats speech as a sequence of feature frames, without considering its phoneme-based construct. The model learns to encode speech frame sequences with a focus on local context at a fine resolution that is not directly related to phonetic decoding. Fig. 1 shows an example of the hierarchical structure of a speech utterance.

As the frame-based masking scheme randomly masks the speech frames, it treats all the speech frames in an utterance equally. It is apparent that individual speech frames do not carry equal information because the duration and energy distributions of phonemes vary very much. Therefore, randomly masking speech frames does not benefit from the inherent properties of a spoken language, e.g. phonotactic and prosodic patterns.

In [59], it was shown that data selection matters in the self-supervised pre-training for downstream ASR tasks. The topic of data selection and masking strategy has been not well studied. All the studies and observations above prompts us to look into the way to incorporate speech properties into the masking strategy in speech representation learning.

B. Segmental Masking Strategies

1) Phoneme-Based Segmental Masking Strategy: The masking schemes dictate what the pretext task predicts during pretraining. We propose to use a phoneme instead of a speech frame as the masking unit. In this way, the model learns to predict a phonetic segment from its phonetic context in a speech utterance. It should be noted that the proposed technique does not require the phonetic identity of the speech segment, it only uses phoneme boundary to define a speech segment, i.e. multiple consecutive speech frames.

In practice, a speech utterance can be segmented into phonetic segments either in an supervised [52], [60] or unsupervised [53] manner. Once we obtain the phoneme boundary, we can randomly apply the masking at a segment level, i.e. over multiple consecutive speech frames, that we will discuss in further detail next.

We denote the entire speech corpus as $\mathcal{X}$ and the acoustic features of the utterance sampled from $\mathcal{X}$ as X. The length (the number of frames) and the height (the number of channels) of X are denoted as L_x and H_x, respectively. The utterance X contains N phonemes $X = (x_1, x_2, \ldots, x_N)$. To mask the input utterance, we randomly select $\{m : m < N\}$ phonemes without replacement. This is another difference between the phoneme-based phonetic masking scheme and the traditional frame-based masking in the literature.

We first randomly select a set of phonemes. 1) We mask 80% of selected phonemes to zero across all the frames. 2) We replace 10% of them with randomly sampled frames. 3) We leave the remaining 10% unchanged. Similar to the masking policy in [5], [20], [21], [24], the 10% unchanged phonemes are introduced to mitigate the train-test mismatch at run-time as the model only receives unmasked speech frames or phonemes during downstream training stage.

In addition to temporal masking, TERA [5] shows that channel alteration and magnitude alteration are also beneficial for speech representation learning. For channel alteration, we randomly mask a certain percentage of channels to zero for all time steps across the input sequence. For magnitude alteration, we randomly apply sampled Gaussian noise to augment the magnitude of input sequences with a certain probability. These two alterations can be easily combined with our phoneme-based masking strategy.

2) Phoneme Span-Based Segmental Masking Strategy: Given a sequence of phonemes $X = (x_1, x_2, \ldots, x_N)$, we select a subset of phonemes $Y \subseteq X$ by iteratively sampling the spans, i.e. ℓ, of consecutive phonemes until the masking budget (e.g., 20% of X) has been spent. At each iteration, we first sample a span length (number of phoneme) from a geometric distribution $\ell \sim Geo(p)$, which is skewed towards shorter spans. We then randomly (uniformly) select the starting point for the span to be masked. Following preliminary trials, we set $p = 0.4$, and also clip ℓ at $\ell_{max} = 7$. This yields a mean span length of

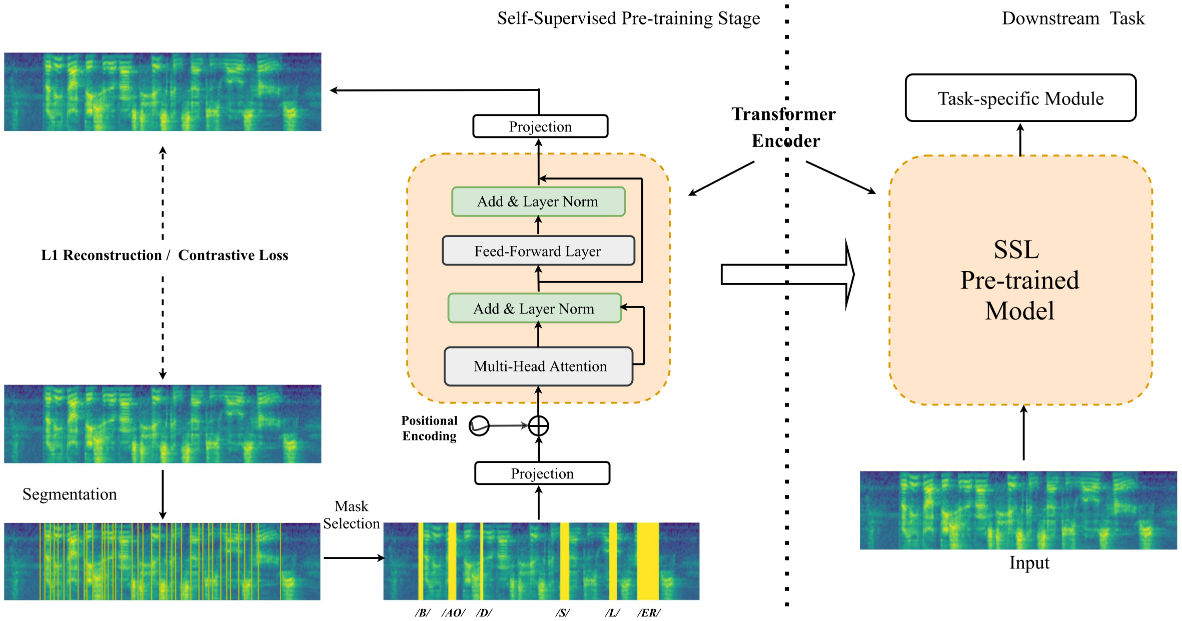

Fig. 2. An overview of the proposed self-supervised learning with segmental masking strategy. The left panel is self-supervised pre-training stage, while the right panel is a downstream application. In the masked spectrogram, some phonemes are selected for masking. We only highlight the phonemes that are randomly masked in the spectrogram.

$mean(\ell) = 2.3$. Fig. 3 shows the distribution of span mask lengths.

As in above phoneme-based masking, we mask 20% of the phonemes in total: replacing 80% of the masked phonemes with zeros, 10% with random frames, and 10% with original frames. However, we perform this replacement at the span level and not for each phoneme individually, i.e. all the phonemes in a span are replaced with zeros or samples frames.

3) Word-Based Segmental Masking Strategy: We may further extend the size of a masking unit to a spoken word. For this to work, we may apply a speech recognizer on the training data to delimit the spoken words. Similar to the phoneme based masking strategy, here, we randomly select a certain proportion of words in the utterance, and mask all speech frames belonging to those selected words.

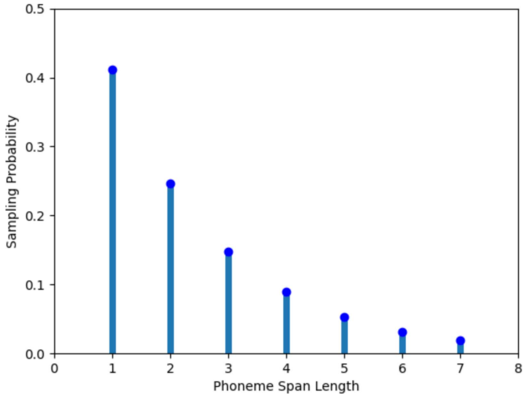

Fig. 3. We sample random phoneme span lengths from a geometric distribution $\ell \sim Geo(p = 0.4)$ clipped at $\ell_{max} = 7$.

C. Segmental Predictive Coding

We propose to use bidirectional Transformer encoder [25], similar to [5], [24], to learn speech representations via masked prediction task as illustrated in Fig. 2. The input acoustic frames and the target predicted frames could be any acoustic frames, such as MFCC, FBANK or spectrogram features. In this work, FBANK features are used if not specified otherwise. Each Transformer encoder layer consists of two sub-layers: (1) a multi-head self-attention module, and (2) a position-wise fully connected feed-forward network (FFN). Each sub-layer has a residual connection, followed by layer normalization [61].

1) Transformer Encoder: Multi-head attention (MHA) is the core module of Transformer encoders, which learns the relationship between queries, keys and values from different representation subspaces at different positions. The basic unit of MHA is self-attention [25].

Since the Transformer encoder is neither recurrent nor convolutional, we use the sinusoidal positional encoding to incorporate the position information of the input acoustic sequence order. Specifically, we first linearly project the input frames into the dimension of the model then we sum the input frames and the positional encoding.

2) Objectives: This work considers two predictive coding methods for learning speech representations: generative predictive coding (GPC) and contrastive predictive coding (CPC). GPC is a generative approach, while CPC is a discriminative approach. In this work, GPC and CPC share a similar methodology as both use the bidirectional transformer encoder to learn representations but differ in the way they optimize the model: while GPC attempts to predict the masked speech frames via minimizing L1 regression loss, CPC incorporates a proposal distribution for drawing negative samples, and learns representations containing information that most discriminate the masked speech frames from the negative samples using contrastive loss, which is defined as:

$$\mathcal{L}_c = -\log \frac{\exp(sim(c_t, q_t)/k)}{\sum_{\tilde{q}\sim Q_t} \exp(sim(c_t, \tilde{q})/k)} \tag{1}$$

where $sim(a, b) = a^T b/\|a\|\|b\|$ is the cosine similarity between the context representations and the target features, Q_t is a set of $K + 1$ candidate representations $\tilde{q} \in Q_t$ which includes q_t and K distractors. The K distractors are drawn from the same sequence X.

D. Integration With Downstream Tasks

A speech representation learning model is useful only if it contributes to a downstream task. There are two typical ways to integrate a pre-trained model with a downstream task.

1) Pre-Trained Model as Feature Frontend: In this approach, we use the pre-trained model as a feature extraction frontend of the downstream task. We may use the hidden states of the last Transformer encoder layer and the feature representations. In this way, the pre-trained model is fixed, and not involved in the training of the downstream task.

Since the representations from last layer is not always the best [62], [63], we extract the hidden states from all encoder layers and weighted-sum them through learnable weights as the final representations. In this way, the information encoded in every layer can be better adapted to downstream tasks, yielding better transfer performance. In this work, we use the weighted-sum approach as default if not specified otherwise.

2) Fine-Tuning of Pre-Trained Model: This approach instead involves tuning the parameters of the pre-trained network on the downstream tasks. The fine-tuning often performs better than the feature-based extraction approach. Since we aim to directly evaluate the quality of learned speech representations from pre-training, we do not consider fine-tuning approach that updates the parameters of the pre-trained model if not specified otherwise.

IV. Experimental Setup

In this section, we conduct extensive experiments to evaluate the learned speech representations obtained with different masking strategies in six downstream tasks. We will describe the model architectures and the hyperparameters as part of the experiment setup.

A. Pre-Training Setup

For the pre-training, we use the publicly available LibriSpeech [64] dataset. It consists of three training subsets: *train-clean-100*, *train-clean-360*, and *train-other-500*. We use these subsets to form different pre-training datasets with various size, including 100 hours (*train-clean-100*), 360 hours (*train-clean-360*) and 960 hours (*train-clean-100, train-clean-360, and train-other-500*). Besides the commonly used LibriSpeech for pre-training, we also consider the CommonVoice [65]. For the phoneme segmentation or word segmentation, we use the force-alignment from the LibriSpeech pre-trained model.[1] It should be noted that we only use the phoneme or word boundary information, and not the phonetic labels.

To validate the effectiveness of the proposed method on a zero annotation dataset, we also use the pre-trained unsupervised phoneme segmentation model from [53], which is trained on the train-other-500 subset of LibriSpeech, to obtain pseudo phoneme boundary labels. We use the force-alignment as the default setting if not specified otherwise. All experiments including pre-training and downstream tasks are performed using 80-dimensional FBANK features (normalized to zero mean and unit variance per speaker) computed over the sampled 16 kHz audio. We conducted all the experiments using the s3prl [2] toolkit. The total training steps of pre-training are set to 200 K, 500 K, 1 M for 100 hours, 360 hours and 960 hours of speech data, respectively. For CommonVoice, the training steps is set to 1 M. We use Transformer Encoders with the hidden size of 768, the number of self-attention heads as 12, dropout probability of 0.1, and the hidden size of the intermediate feed-forward layer as 3,072. We use gradient descent algorithm with batch size of 32. The Adam optimizer [66] is employed for updating model parameters, warming up the learning rate for the first 7% of total training steps to a peak of 2e-4 and then linearly decayed. The configurations of pre-training and task-specific downstream training are summarized in Table I. For all our baselines, including AALBERT [26], Mockingjay [24] and TERA [5], we directly use the released versions in the s3prl toolkit. We did not pre-train them from scratch by ourselves.

B. Downstream Tasks

1) Phoneme Classification: We first study how the pre-trained models with our proposed segmental masking strategies contribute to a phoneme classification (PC) task. Following the common setting of previous work [5], [6], we use a linear classifier to evaluate the linear separability of phonemes on the *train-clean-100* subset of LibriSpeech. For a fair comparison, we use the aligned phoneme labels and train/test split provided in the CPC [6] paper, where there are 41 possible phoneme classes obtained using the Kaldi toolkit [67] and pre-trained models on LibriSpeech. Since not all the information encoded from the pre-trained model is linearly accessible, we also train a non-linear classifier with a single hidden layers for phoneme classification, following the same setting in CPC [6] and TERA [5]. These two classifiers are denoted as *PC (Linear)* and *PC (non-Linear)*, respectively. For fair comparison with TERA and CPC, we here only use the hidden states of last encoder layer for classification, and report the results in terms of accuracy (ACC). Since pre-training and downstreaming both on LibriSpeech introduce a bias. To remove this bias, we also perform phoneme classification on Wall Street Journal (WSJ)[68] dataset.

2) Speech Recognition: We further study how the pre-trained model with segmental masking strategies benefits from the encoded knowledge of phonotactic constraint of a language for

[1] [Online]. Available: www.kaldi-asr.org/downloads/build/6/trunk/egs/LibriSpeech/

[2] [Online]. Available: https://github.com/s3prl/s3prl

TABLE I
CONFIGURATIONS OF PRE-TRAINING AND DOWNSTREAM TASKS

Self-Supervised Pre-training	
input features	80-dimensional FBANK
transformer encoder layers	3
attention hidden size	768
multi-heads	12
feed-forward dimension	3072
attention dropout probability	0.1
batch size	32
training steps	200K / 500K / 1M
Phoneme Classification	
phoneme classes	41
one hidden layer dimension	768
batch size	32
training steps	200K
Speech Recognition	
liGRU layers	5
batch size	32
training epochs	24
Keyword Spotting	
keywords classes	12
batch size	32
training steps	200K
Speaker Identification	
speaker classes	251
batch size	32
training steps	200K
Intent Classification	
intent classes	3
batch size	32
training steps	200K
Emotion Recognition	
emotion classes	4
batch size	32
training steps	30K

speech recognition. Here, we perform the evaluation using a Deep Neural Network-Hidden Markov Model (DNN-HMM). The ASR system is built with the PyTorch-Kaldi toolkit [69]. We use the advanced DNN architecture, comprising a 5-layer light-gated recurrent units (liGRU) followed by 2-layers of fully-connected networks. We feed the output of the pre-trained model to the hybrid ASR system while maintaining pre-trained model parameters frozen during training. The supervised ASR training experiments are conducted on the TIMIT [70] dataset. Following the conventional settings, the ASR model on TIMIT is based on 48 phoneme classes, while accuracy is measured after mapping the prediction to a smaller set of 39 phoneme classes. Splits of the dataset are obtained according to the Kaldi TIMIT [67] recipe. The results are reported in terms of Phone Error Rate (PER).

3) Keyword Spotting: Keyword spotting (KWS) is a variant of speech recognition. As the KWS task that seeks to detect the presence of pre-defined keywords in a speech flow, the knowledge of phonotactic constraint in a language is certainly useful. Experiments are conducted with Google Speech Commands Datasets [71]. It contains 65,000 one-second-long utterance files, recorded and labeled with one of 30 target categories. Following Google's implementation, we distinguish 12 classes, namely *"yes," "no," "up," "down," "left," "right," "on," "off," "stop," "go," silence, and unknown*. Using SHA-1 hashed name of the audio files, we split the dataset into training, validation, and test sets, with 80% training, 10% validation, and 10% test, respectively. The results are reported in terms of accuracy (ACC).

4) Speaker Identification: Speaker identification (SID) aims to recognize who speaks in a speech audio. It is therefore a multi-class classification task. To identify the speaker individuality of phonetic and prosodic rendering, we believe that we could benefit from the speech representations that are aware of the phonotactic and prosodic constraint in a language.

We here evaluate the speaker characteristics of the learned speech representations on *train-clean-100* subset of LibriSpeech and Voxceleb1 [72]. For a fair comparison, we first use the *train-clean-100* subset for frame-level and utterance-level speaker classification to follow the common experimental setting [5], [6], [24]. This subset contains 251 speakers, and we use the same train/test split as provided in [6]. For utterance-level speaker identification, the representations of each utterance are first averaged over time. Then, the classifier predicts speaker identity conditioning on the averaged embedding. For frame-level speaker identification, the classifier predicts the speaker identity for each input speech frame. For both tasks, we only extract the representations from the last encoder layer and use a linear model to perform classification for fair comparison with CPC [6] and TERA [5]. The two different levels are denoted as *SID (Frame)* and *SID (Utt)*, respectively. It should be noted that this speaker identification task on the *train-clean-100* subset of LibriSpeech only serves as a simple check for the speaker information in the learned speech representations. Hence, we further include experiments using VoxCeleb1 [72] to evaluate the learned representations from the speaker aspect, following [62] training, development and testing configurations, with 138,316; 6,904; 8,251 utterances respectively. We denote the task on VoxCeleb1 as *SID (Vox)* and report the results in terms of accuracy (ACC).

5) Intent Classification: Intent Classification (IC) takes an utterance, i.e. a sequence of speech features, as input and classifies it into one of the pre-defined intent categories. Besides the speech content, the speech prosody is also intent informing, that is reflected in phonetic duration and energy. A pre-trained model, that captures the knowledge of phonetic, phonotactic, as well as phonetic duration in a spoken language, could potentially bring benefits to the IC task.

We perform the IC experiments on the Fluent Speech Commands (FSC) [73] dataset, where each utterance is tagged with three intent labels: action, object and location. We follow the FSC protocol, having 23,132; 3,118, and 3,793 utterances for training, validation and testing, respectively. Following the setting of previous work [62], we use *MeanPooling* followed by a linear classifier for utterance-level classification and report the results in terms of accuracy (ACC).

6) Speech Emotion Recognition: Speech Emotion Recognition (SER) is the task to classify the emotion of a speech into one of the pre-defined categories. As the SER task is supposed to be independent of the speech content. The phonetic, phonotactic and prosodic patterns play a role, just like the intent classification task.

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) [74] dataset, which includes five sessions of utterances

TABLE II
COMPARISON OF SELF-SUPERVISED PRE-TRAINING MODELS ON VARIOUS DOWNSTREAM TASKS, INCLUDING PHONEME CLASSIFICATION (PC), KEYWORD SPOTTING (KWS), SPEAKER IDENTIFICATION (SID), INTENT CLASSIFICATION (IC), SPEECH EMOTION RECOGNITION (SER). THE *TRAIN-CLEAN-100* SUBSET OF LIBRISPEECH OR COMMONVOICE IS USED FOR PRE-TRAINING. THE EXPERIMENT RESULTS FOR DOWNSTREAM TASKS ARE REPORTED IN TERMS OF ACCURACY (%). OUR MODEL IS TRAINED WITH L1 RECONSTRUCTION LOSS

ID	Pre-trained Models	PC (Linear)	PC (non-Linear)	KWS	SID (Frame)	SID (Utt)	SID (Vox)	IC	SER
1	FBANK	42.1	46.9	8.63	0.6	5.4	8.5E-4	9.1	35.4
2	CPC [6]	64.6	72.5	-	97.4	-		-	-
3	Mockingjay [24]	64.3	76.8	82.6	68.4	96.1	48.4	53.3	57.8
4-1	TERA [5]	65.1	77.3	83.1	98.9	99.2	55.9	56.4	60.8
4-2	TERA [5] - FT	90.7	91.1	93.8	99.7	99.0	75.1	97.6	55.4
4-3	TERA [5] - CV	68.1	76.7	88.3	98.7	99.6	51.4	62.0	59.8
5	Ours (word)	69.2	78.6	84.2	99.9	100.0	52.0	56.4	61.2
6	Ours (phoneme span)	71.8	80.3	83.6	99.9	100.0	56.4	59.7	61.6
7-1	Ours (phoneme)	72.8	80.9	86.4	100.0	100.0	58.7	65.6	62.5
7-2	Ours (phoneme) - FT	89.7	91.0	93.7	99.8	99.4	75.4	98.5	58.2
7-4	Ours (phoneme) - Unsupervised	71.4	80.2	84.6	100.0	100.0	58.4	70.0	62.5
7-3	Ours (phoneme) - CV-Unsupervised	69.1	78.2	89.6	99.5	100.0	52.1	62.9	61.3

between two speakers (one male and one female), is adopted in this task. Following the conventional evaluation setting [62], we exclude the imbalanced emotion classes, and only consider four classes (neutral, happy, sad, angry) with a similar amount of data points and cross-validates on five folds of the standard splits. Similar to IC, we use *MeanPooling* to obtain the utterance-level embedding, and then use this embedding for recognition. The results are reported in terms of accuracy (ACC).

V. EXPERIMENTAL RESULTS AND ANALYSIS

We now report the experimental results on various downstream tasks to show the effectiveness of our proposed segmental masking strategies, including the phoneme-based masking, the phoneme span-based masking and the word-based masking, in Section V-A. Then in Section V-B, we present the ablation study of different phoneme masking rates. In Section V-C, we compare the performance when pre-trained with different acoustic features under the phoneme-based masking. Finally, we further explore the effects of different objectives (e.g., contrastive or reconstruction) in Section V-D.

A. Effectiveness of Segmental Masking Strategies

Table II shows the evaluation results on various downstream tasks including phoneme classification (PC), keyword spotting (KWS), speaker identification (SID) on two datasets (i.e. the *train-clean-100* subset of LibriSpeech and VoxCeleb1), intent classification (IC) and emotion recognition (SER).

Among the tasks, Mockingjay [24] and TERA [5] are based on reconstruction loss, while Mockingjay is equivalent to TERA when the alteration is done temporally. Hence, for fair comparison, we here use L1 reconstruction loss for our pre-trained model and the masking probability is set to 20%. All models are pre-trained on the *train-clean-100* subset of LibriSpeech. Similar to TERA [5], we perform alteration from three dimensions (e.g., time, channel, and magnitude). The suffix FT denotes the fine-tuning results, the suffix CV denotes the pre-training on CommonVoice, and the suffix Unsupervised denotes the phoneme segmentation is obtained using pre-trained unsupervised phoneme segmentation model from [53].

Firstly, as expected, all pre-trained models (ID 2-7) outperform the surface feature (ID 1) across all downstream tasks. Secondly, our method (ID 5,6,7) consistently improves all downstream tasks over Mockingjay (ID 3), TERA (ID 4) and CPC (ID 2). The improvements observed in the downstream tasks support our belief that phoneme-based masking outperforms frame-based masking as a pre-training strategy.

Specifically, our phoneme-based method (ID 7) outperforms TERA (ID 4), e.g. from 65.1% / 77.3% to 72.8% / 80.9% for phoneme classification, and outperforms CPC (ID 2) by a large margin. This suggests that phoneme-based masking in the pre-trained model strengthens the phoneme prediction in the downstream task. As far as masking unit size is concerned, we can see that the phoneme-based (ID 7) (72.8% / 80.9%) masking performs the best, followed by phoneme span (ID 6) (71.8% / 80.3%), and lastly, word-based masking (ID 5) (69.2% / 78.6%) on phoneme classification. Empirically, we observe that phoneme serves as the suitable masking unit, that coincides with the general understanding that phoneme is the smallest unit of sound.

In speaker identification (SID) task, our three segmental masking methods obtain almost 100.0% accuracy on *train-clean-100* subset (IDs 5,6,7), suggesting that the pre-trained model with segmental masking better characterizes the speakers than other pre-trained models. Besides the *train-clean-100* subset, we also evaluate the learned speaker characteristics on VoxCeleb1 [72]. The results show that our phoneme-based masking (ID 7) consistently contributes to the downstream speaker identification task, yielding 58.7% accuracy, outperforming Mockingjay (ID 3, 48.4%) and TERA (ID 4, 55.9%).

However, we did not find the similar gain with the phoneme span-based masking (ID 6) and the word-based masking (ID 5). They only gives 56.4% and 52.0% results, respectively. An explanation for this result is that speakers characteristics differ from one speaker to the other on phoneme level, given by the vocal tract characteristics of each speaker.

In both IC and SER tasks, similar to PC and SID, the phoneme-based masking (ID 7) performs the best, yielding 65.6% / 62.5% accuracy, respectively, and outperforms 53.3% / 57.8% of Mockingjay, and 56.4% / 60.8% of TERA. In addition, among the

TABLE III
Phoneme Classification Results (Accuracy %) on LibriSpeech Dataset, With Surface Features (MFCC, FBANK, and fMLLR), and Pre-Trained Speech Representations (CPC, Mockingjay, TERA, and Phoneme-Based Segmental Masking), for Both Linear Classifier and Non-Linear Classifier (1 Hidden Layer) With Different Pre-Training Data Size

Surface Features	0 h		Pre-trained Models	100 h		360 h		960 h	
	Linear	non-Linear		Linear	non-Linear	Linear	non-Linear	Linear	non-Linear
			CPC [6]	64.6	72.5	-		-	
MFCC	39.7	59.9	Mockingjay [24]	64.3	76.8	64.4	77.0	67.0	79.1
FBANK	42.1	46.9	TERA [5]	65.1	77.3	66.4	78.3	66.4	78.9
fMLLR	52.6	68.4	Ours	72.8	80.9	74.7	82.2	75.6	82.4

TABLE IV
Phoneme Classification Results (Accuracy %) on WSJ Dataset With TERA and Phoneme-Based Segmental Masking, for Both Linear Classifier and Non-Linear Classifier (1 Hidden Layer) With Different Pre-Training Data Size

	TERA		ours (phoneme)	
Pre-training Data	Linear	non-Linear	Linear	non-linear
100 hr	76.7	82.8	77.8	84.6
360 hr	-		79.2	85.4
960 hr	77.4	83.5	79.2	85.6

TABLE V
Comparison of the Proposed Method With Other Recent Approaches on TIMIT. All Pre-Training Data are From LibriSpeech Dataset, If Not Specified Otherwise

Models	Pre-training Data	PER
CNN + HMM [75]	None	16.5
CNN + TD-filterbanks [76]	None	18.0
liGRU + MFCC [77]	None	16.7
liGRU + FBANK [77]	None	15.8
liGRU + fMLLR [77]	None	14.9
wav2vec [43]	960 hr	15.6
wav2vec [43]	960 hr + WSJ 81 hr	14.7
liGRU + TERA [5]	100 hr	15.2
liGRU + TERA [5]	360 hr	14.9
liGRU + TERA [5]	960 hr	14.5
Ours (word)	100 hr	15.2
Ours (phoneme span)	100 hr	14.7
Ours (phoneme)	100 hr	14.1
Ours (phoneme)	360 hr	13.9
Ours (phoneme)	960 hr	13.6
Ours (phoneme) - Unsupervised	100 hr	14.6
Ours (phoneme) - Unsupervised	360 hr	14.2
Ours (phoneme) - Unsupervised	960 hr	14.0

three segmental masking implementations, it is obvious that the phoneme-based masking beats the phoneme span-based masking and the word-based masking by a large margin, especially on IC tasks (65.6% v.s. 56.4% and 65.6% v.s 59.7%). Overall, these results suggest that the phoneme-based masking is effective for downstream IC and SER tasks.

We also report two fine-tuning results in Table II, one is TERA [5], another is our proposed method. We can see that TERA and our method obtain comparable performance on most downstream tasks, except that TERA gives 55.4% accuracy on SER task, while our method gives 58.2% accuracy. On the other side, using the pre-trained model as feature frontend, our method achieves significantly better performance on all downstream tasks, suggesting that our method learns much better speech representations.

In Table III, we report the phoneme classification performance as a function of the pre-training data size. In this experiment, we adopt phoneme as the masking unit. The results show that the phoneme accuracy linearly increases as the pre-training data increases, from 72.8% on 100 hours, to 74.7% on 360 hours, to 75.6% on the full LibriSpeech 960 hours of data. This trend is not observed in Mockingjay and TERA, suggesting that our method exploits the available data more effectively. By using only 100 hours of pre-training data, we (72.8% / 80.9%) even surpass the results of TERA [5] (66.4% / 78.8%) and Mockingjay [24] (67.0% / 79.1%) that use 960 hours of data for pre-training.

Table IV shows the phoneme classification results on WSJ dataset. Our method (77.8% / 84.6%) outperforms TERA (76.7% / 82.8%) when the pre-training data is 100 hours, even better than TERA (77.4% / 83.5%) that use 960 hours of data for pre-training.

In Table V, we summarize experiment results on TIMIT in terms of Phone Error Rate (PER), that include two recent self-supervised learning techniques, e.g., TERA [5] and wav2vec [43], and several competitive supervised baselines [75]–[77]. We can see that our phoneme-based masking with 960 hours of pre-training data yields the best PER (13.6%), outperforming TERA (14.5%) and wav2vec (15.6%) by 0.9% and 2% absolute PER reduction, respectively. When only 100 hours of pre-training data, among all our proposed segmental masking strategies, the word-based masking gives the worst result (15.2%) as expected, while the phoneme-based masking strategy gives the best result (14.1%). What is more, when only using 100 hours of pre-training data, our phoneme-based masking outperforms TERA [5] (14.5%) pre-trained on 960 hours and wav2vec [43] (14.7%) pre-trained on both full LibriSpeech and Wall Street Journal (WSJ) datasets. Finally, increasing the data for pre-training benefits learning with a steady PER decrease from 14.1% on 100 hours, 13.9% on 360 hours, to 13.6% on 960 hours.

In both Table II and Table V, we also report the results where the unsupervised phoneme segmentation is used for the pre-training. We can see that our method does not need very accurate phoneme segmentation to capture the phonotactic constraints during the pre-training, but consistently benefits the downstream tasks. However, more accurate boundary information (e.g., via force alignment) gives better pre-training results.

In Table VI, we report the speaker identification performance with different pre-training data size. We also include the results of three surface features, including MFCC, FBANK, and feature-space Maximum Likelihood Linear Regression (fMLLR) [78]. However, these surface features perform very bad on SID task, indicating there is almost no speaker information

TABLE VI
SPEAKER IDENTIFICATION RESULTS ON THE *TRAIN-CLEAN-100* SUBSETS OF LIBRISPEECH AND VOXCELEB1 BY VARYING AMOUNT OF PRE-TRAINING DATA. OUR MODEL IS TRAINED WITH PHONEME BASED SEGMENTAL MASKING

Surface Features	0 h		Pretrained Models	100 h			360 h			960 h		
	Frame	Utt		Frame	Utt	Vox	Frame	Utt	Vox	Frame	Utt	Vox
			CPC [6]	97.4	-			-			-	
			AALBERT [26]		-		98.8	99.1	-		-	
MFCC	17.6	10.8	Mockingjay [24]	68.4	96.1	48.4	86.9	97.3	-	99.3	99.7	44.7
FBANK	0.6	5.4	TERA [5]	98.9	99.2	55.9	99.0	99.5	-	99.5	99.8	57.6
fMMLR	0.4	2.6	Ours	100.0	100.0	58.7	100.0	100.0	62.1	100.0	100.0	63.7

TABLE VII
ABLATION STUDY OF DIFFERENT PHONEME MASKING RATE

Mask Rate	PC (Linear)	PC (non-Linear)	KWS	SID (Vox)	IC	SER
10%	68.1	76.5	85.1	48.8	64.9	61.8
15%	69.5	79.4	84.9	49.6	65.4	61.4
20%	71.3	80.0	85.5	47.2	65.6	62.9
25%	70.1	79.3	85.3	52.3	56.6	59.6
30%	69.0	79.4	84.0	49.2	55.5	59.2
40%	65.7	79.0	81.4	48.3	47.0	57.1
50%	64.9	76.9	75.6	46.1	41.0	59.6

TABLE VIII
ABLATION STUDY OF PHONEME SPAN LENGTH, WORD MASKING RATE, AND ACOUSTIC FEATURES. ALL MODELS ARE PRE-TRAINED ON *TRAIN-CLEAN-100* SUBSET OF LIBRISPEECH

Masking Unit	Masking Rate	Span Length	PC (Linear)	PC (non-Linear)	KWS	SID (Vox)	IC	SER
Phoneme (FBANK)	20%	-	72.8	80.9	86.4	58.7	68.5	62.5
Phoneme (MFCC)	20%	-	73.9	81.2	86.8	56.5	57.4	59.0
Phoneme (Spectrogram)	20%	-	70.3	79.0	65.2	56.1	55.5	58.5
Phoneme Span	20%	7	71.6	80.3	83.1	59.5	56.4	61.6
	20%	5	71.8	80.3	83.6	56.4	59.7	61.6
	20%	3	71.8	80.3	83.1	56.2	59.7	61.5
Word	10%	-	70.6	79.4	84.1	56.3	62.7	60.6
	15%	-	69.6	78.1	84.2	56.2	61.1	61.2
	20%	-	69.2	78.6	85.1	52.0	59.4	60.6

encoded in surface features. On the other hand, similar to the phoneme classification and speech recognition, the speaker characteristics also benefit from a larger amount of pre-training data. The accuracy on VoxCeleb1 increases from 58.7% on 100 hours of pre-training data, to 62.1% on 360 hours of pre-training data, to 63.7% on full 960 hours LibriSpeech data.

B. Effect of Masking Rate

Just like in frame-based masking, the masking rate within an utterance also plays a role. By varying the phoneme masking rate from 10% to 50%, we hope to observe the effect. The previous studies on Mockingjay [24] and TERA [5] show that masking 15% of the speech frames is ideal, that has been adopted in subsequent works [22], [23], [26], [27].

Table VII shows the downstream results as we vary the phoneme masking rate. In this set of experiments, we only did the time alteration like Mockingjay [24], and the channel and magnitude alterations are not applied. The results show that the overall performance is better as the masking rate increases from 10% to 20%, with 20% giving the best performance, which is different from that of the frame-based masking strategy (15%). It is worth mentioning that the performance of 10% masking is already better than the frame-based masking method with 15% masking rate. From 20% onward, the performance will start to drop if we keep increasing the masking rate. The reason is that excessive masked phonemes will increase the difficulty of the pre-training and hence hurt the quality of the learned speech representations. For instance, when the masking rate is 50%, the phoneme classification accuracy is only 64.9% / 76.9%, compared to the best accuracy of 71.3% / 80.0% that pre-trained with 20% masking rate.

Table VIII shows the ablation study of the phoneme span length for the phoneme span-based masking and the masking rate for the word-based masking. For the phoneme span, we select a subset of phonemes by iteratively sampling span of phoneme until the masking budget (20% of the phoneme sequence) is used up and pre-train the model with different span length. Interestingly, we find that the length of the phoneme span does not affect the performance of the downstream tasks as long as they follow the same masking rate (e.g., 20%).

In Table VIII, we also report the performance of word-based masking at different masking rates. We observe that increasing the word masking rate (e.g., from 10% to 20%) adversely affects the performance, with 10% giving the best results as opposed to 20% for phoneme-based masking. This could be due to the fact that word is a much larger acoustic unit than phoneme in terms of duration.

TABLE IX
COMPARISON OF DIFFERENT PRE-TRAINING OBJECTIVES. ALL MODELS ARE PRE-TRAINED ON *TRAIN-CLEAN-100* SUBSET OF LIBRISPEECH AT A MASKING RATE OF 15%

Downstream	L1 loss		Contrastive loss	
	frame	phoneme	frame	phoneme
PC (Linear)	64.3	71.3	66.6	72.7
KWS	82.6	85.5	82.6	85.0
SID (Vox)	48.4	47.2	45.1	51.6
IC	53.3	68.5	53.4	66.0
SER	57.8	62.9	58.4	61.6

C. Effect of Different Speech Features

In order to explore the impact of different acoustic features under the phoneme-based masking, we pre-train the model to reconstruct different acoustic features. In this study, we experiment with three different features, including 80-dimension FBANK, 13-dimension MFCC, and 160-dimension Spectrogram. The results are summarized in the upper part of Table VIII. We find that pre-training with MFCC features achieves best performance on phoneme classification, giving 73.9% accuracy on phoneme classification, this is also the state-of-the-art result. However, for other downstream tasks, especially intent classification, pre-training with FBANK leads to better results (58.7% on SID, 68.5% on IC, and 62.5% on SER) compared to pre-training with MFCC (56.5% on SID, 57.4% on IC, and 59.0% on SER) and Spectrogram (56.1% on SID, 55.5% on IC, and 58.5% on SER). Among these three features, pre-training with spectrogram yields worst overall performance. We conclude that despite the same model architecture and objective, pre-training with different acoustic features will significantly affect the quality of the learned speech representations.

D. Effect of Different Pre-Training Objectives

We have used L1 reconstruction loss as the default setting in this work. To further validate the effectiveness of the segmental masking strategy, we extend this study to a contrastive framework. In this set of experiments, only time alteration is performed following the setting in Mockingjay [24]. For the contrastive framework, we follow the wav2vec2.0-style [4] CPC except that we did not use the quantization module, and we randomly sample 50 negative samples from the same utterance and compute the contrastive loss.

In Table IX, we report the results between L1 reconstruction loss and contrastive loss. All masking rates are set to 15% for both objectives. The *train-clean-100* subset of LibriSpeech is used for pre-training. With both frame-level masking and phoneme-based masking, the contrastive loss yields better overall performance. Furthermore, the study on both objectives confirms that segmental masking strategy outperforms frame-based masking on all downstream tasks.

E. Continual Pre-Training on wav2vec2.0

To benchmark the segmental masking constraint against the state-of-the-art, e.g., wav2vevc2.0 [4] and HuBERT [79], we replace the random frame-level masking in wav2vec2.0 with the proposed phoneme-based segmental masking strategy in a continual pre-training. In this set of experiments, the masking probability is set to 50% and only the subset *train-other-500* of LibriSpeech is used for a quick turn-around. We conduct automatic speech recognition (ASR) experiment from SUPERB [62] benchmark to compare the masking strategies.

TABLE X
RESULTS OF THE CONTINUAL PRE-TRAINING ON WAV2VEC2.0 USING THE PROPOSED SEGMENTAL MASKING STRATEGY

Models	ASR	KWS	SID(Vox)	IC	SER
HuBERT [79]	6.4	96.3	81.4	98.3	64.9
wav2vec2.0 [4]	6.4	96.2	75.2	92.4	63.4
wav2vec2.0 (phoneme)	6.2	96.4	74.7	94.1	64.0

Table X shows the results of the continual pre-training on wav2vec2.0. It is apparent that the continual pre-training with the segmental masking strategy consistently outperforms the wav2vec2.0 baseline, especially on the ASR, KWS, and IC tasks, which depend more on the phonetic and semantic information. The experiments validate the benefit of encoding phonotactic constraints by the segmental masking strategy.

VI. CONCLUSION

We propose a novel segmental masking strategy, which uses phonetically motivated segments as the masking units to encode the phonotactic and prosodic constraints during the learning process. We explore three different masking units, e.g., phoneme, phoneme span and word, and provide a comprehensive study into the effect of masking units. We have demonstrated the effectiveness of the segmental masking strategy on various downstream tasks over frame-based masking strategy. To the best of our knowledge, this is the first work that investigates the impacts of different masking schemes. Finally, we extend our segmental masking scheme to contrastive learning to achieve competitive performance.

REFERENCES

[1] K. L. Sakai, "Language acquisition and brain development," *Science*, vol. 310, no. 5749, pp. 815–819, 2005.

[2] P. K. Kuhl, "Brain mechanisms in early language acquisition," *Neuron*, vol. 67, no. 5, pp. 713–727, 2010.

[3] K. Hirsh-Pasek, D. G. K. Nelson, P. W. Jusczyk, K. W. Cassidy, B. Druss, and L. Kennedy, "Clauses are perceptual units for young infants," *Cognition*, vol. 26, no. 3, pp. 269–286, 1987.

[4] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. 33th Conf. Neural Inf. Process. Syst.*, 2020, pp. 12449–12460.

[5] A. Liu, S.-W. Li, and H.-Y. Lee, "TERA: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2351–2366, Jul. 2021.

[6] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," in 2018, *arXiv:1807.03748*.

[7] M. Ravanelli et al., "Multi-task self-supervised learning for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6989–6993.

[8] J. Huh, H. S. Heo, J. Kang, S. Watanabe, and J. S. Chung, "Augmentation adversarial training for self-supervised speaker recognition," in *Proc. 34th Conf. Neural Inf. Process. Syst.*, SAS Workshop, 2020.

[9] N. Arsha, C. J. Son, A. Samuel, and Z. Andrew, "Disentangled speech embeddings using cross-modal self-supervision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6829–6833.

[10] A. Wu, C. Wang, J. Pino, and J. Gu, "Self-supervised representations improve end-to-end speech translation," in *Proc. Interspeech*, 2020, pp. 1491–1495.

[11] H. Nguyen, F. Bougares, N. Tomashenko, Y. Estéve, and L. Besacier, "Investigating self-supervised pre-training for end-to-end speech translation," in *Proc. Int. Conf. Mach. Learn.*, SAS Workshop, 2020.

[12] Y.-C. Wang, S. Venkataramani, and P. Smaragdis, "Self-supervised learning for speech enhancement," in *Proc. Int. Conf. Mach. Learn.*, SAS Workshop, 2020.

[13] A. Sivaraman, S. Kim, and M. Kim, "Personalized speech enhancement through self-supervised data augmentation and purification," in *Proc. Interspeech*, 2021, pp. 2676–2680.

[14] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2041–2053, Dec. 2019.

[15] Y.-A. Chung, Y. Belinkov, and J. Glass, "Similarity analysis of self-supervised speech representations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 3040–3044.

[16] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," in *Proc. Interspeech*, 2019, pp. 146–150.

[17] Y.-A. Chung and J. Glass, "Generative pre-training for speech with autoregressive predictive coding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 3497–3501.

[18] M. Riviére, A. Joulin, P.-E. Mazar'e, and E. Dupoux, "Unsupervised pretraining transfers well across languages," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7414–7418.

[19] K. Kawakami, L. Wang, C. Dyer, P. Blunsom, and A. van den Oord, "Learning robust and multilingual speech representations," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2020, pp. 1182–1192.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT:Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Chapter Assoc. Comput. Linguist.*, 2019, pp. 4171–4186.

[21] Y. Liu et al., "Roberta: A robustly optimized bert pretraining approach," 2019, *arXiv:1907.11692*.

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

[23] L. Liu and Y. Huang, "Masked pre-trained encoder base on joint ctc-transformer," 2020, *arXiv:2005.11978*.

[24] A. Liu, S.-W. Yang, P.-H. Chi, P.-C. Hsu, and H.-Y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6419–6423.

[25] A. Vaswani et al., "Attention is all you need," in *Proc. 31th Conf. Neural Inf. Process. Syst.*, 2018, pp. 6000–6010.

[26] P.-H. Chi, P.-H. Chung, T.-H. Wu, C.-C. Hsieh, S.-W. Li, and H. yi Lee, "Audio albert: A lite bert for self-supervised learning of audio representation," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 344–350.

[27] D. Jiang et al., "A further study of unsupervised pre-training for transformer based speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6538–6542.

[28] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "SpanBERT: Improving pre-training by representing and predicting spans," *IEEE Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 64–77, 2020.

[29] X. Yue and H. Li, "Phonetically motivated self-supervised speech representation learning," in *Proc. Interspeech*, 2021, pp. 746–750.

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.

[31] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2015, pp. 3367–3375.

[32] A. Graves, A. Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 6645–6649.

[33] W. Shinji, H. Takaaki, K. Suyoun, H. J. R., and H. Tomoki, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1240–1253, Dec. 2017.

[34] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014.

[35] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.

[36] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Empirical Methods Nat. Lang. Process.*, 2014, pp. 1724–1734.

[37] X. Liu et al., "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, early access, Jun. 22, 2021, doi: 10.1109/TKDE.2021.3090866.

[38] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[39] H. Kaiming, F. Haoqi, W. Yuxin, X. Saining, and G. Ross, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2020, pp. 9729–9738.

[40] M. E. Peters et al., "Deep contextualized word representations," in *Proc. North Amer. Chapter Assoc. Comput. Linguist.*, 2018, pp. 2227–2237.

[41] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xl-Net: Generalized autoregressive pretraining for language understanding," in *Proc. 33th Conf. Neural Inf. Process. Syst.*, 2019, pp. 5753–5763.

[42] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite BERT for self-supervised learning of language representations," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–17.

[43] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. Interspeech*, 2019, pp. 3465–3469.

[44] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–12.

[45] Y.-A. Chung et al., "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," 2021, *arXiv:2108.06209*.

[46] Y.-A. Chung and J. Glass, "Improved speech representations with multi-target autoregressive predictive coding," in *Proc. 56th Annu. Meet. Assoc. Comput. Linguist.*, 2020, pp. 2353–2358.

[47] Y.-A. Chung, H. Tang, and J. Glass, "Vector-quantized autoregressive predictive coding," in *Proc. Interspeech*, 2020, pp. 3760–3764.

[48] S. Ling, Y. Liu, J. Salazar, and K. Kirchhoff, "Deep contextualized acoustic representations for semi-supervised speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6429–6433.

[49] S. Ling and Y. Liu, "Decoar 2.0: Deep contextualized acoustic representations with vector quantization," 2020, *arXiv:2012.06659*.

[50] X. Song, G. Wang, Y. Huang, Z. Wu, D. Su, and H. Meng, "Speech-xlnet: Unsupervised acoustic model pretraining for self-attention networks," in *Proc. Interspeech*, 2020, pp. 3765–3769.

[51] S. Pascual, M. Ravanelli, J. Serrá, A. Bonafonte, and Y. Bengio, "Learning problem-agnostic speech representations from multiple self-supervised tasks," in *Proc. Interspeech*, 2019, pp. 161–165.

[52] K. Felix, S. Yaniv, K. Joseph, and A. Yossi, "Phoneme boundary detection using learnable segmental features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 8089–8093.

[53] F. Kreuk, J. Keshet, and Y. Adi, "Self-supervised contrastive learning for unsupervised phoneme segmentation," in *Proc. Interspeech*, 2020, pp. 3700–3704.

[54] Y.-H. Wang, C.-T. Chung, and H. yi Lee, "Gate activation signal analysis for gated recurrent neural networks and its correlation with phoneme boundaries," in *Proc. Interspeech*, 2017, pp. 3822–3826.

[55] P. Godard et al., "Unsupervised word segmentation from speech with attention," in *Proc. Interspeech*, 2018, pp. 2678–2682.

[56] S. Bhati, J. Villalba, P. Żelasko, L. Moro-Velázquez, and N. Dehak, "Segmental contrastive predictive coding for unsupervised word segmentation," in *Proc. Interspeech*, 2021, pp. 366–370.

[57] J. Chorowski et al., "Aligned contrastive predictive coding," in *Proc. Interspeech*, 2021, pp. 976–980.

[58] S. Cuervo et al., "Contrastive prediction strategies for unsupervised segmentation and categorization of phonemes and words," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.,*2021, pp. 3189–3193.

[59] W.-N. Hsu et al., "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," in *Proc. Interspeech*, 2021, pp. 721–725.

[60] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Proc. Interspeech*, 2017, pp. 498–502.

[61] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[62] S. wen Yang et al., "Superb: Speech processing universal performance benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

[63] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," in *Proc. 34th Conf. Neural Inf. Process. Syst.*, 2021, pp. 27826–27839.

[64] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “LibriSpeech: An ASR corpus based on public domain audio books,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 5206–5210.

[65] R. Ardila et al., “Common voice: A massively-multilingual speech corpus,” in *Proc. Lang. Resour. Eval. Conf.*, 2020, pp. 4211–4215.

[66] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.

[67] D. Povey et al., “Purely sequence-trained neural networks for asr based on lattice-free mmi,” in *Proc. Interspeech*, 2016, pp. 2751–2755.

[68] D. B. Paul and J. M. Baker, “The design for the wall street journal-based CSR corpus,” in *Proc. Workshop Speech Nat. Lang.*, 1992, pp. 357–362.

[69] M. Ravanelli, T. Parcollet, and Y. Bengio, “The pytorch-kaldi speech recognition toolkit,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6465–6469.

[70] K.-F. Lee and H.-W. Hon, “Speaker-independent phone recognition using hidden markov models,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 37, no. 11, pp. 1641–1648, Nov. 1989.

[71] P. Warden, “Speech commands: A public dataset for single-word speech recognition,” 2017, [Online]. Available: https://ai.googleblog.com/2017/08/launching-speech-commands-dataset.html

[72] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, “Voxceleb: Large-scale speaker verification in the wild,” *Comput. Speech Lang.*, vol. 60, 2020, Art. no. 101027.

[73] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, “Speech model pre-training for end-to-end spoken language understanding,” in *Proc. Interspeech*, 2019, pp. 814–818.

[74] C. Busso et al., “Iemocap: Interactive emotional dyadic motion capture database,” *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[75] L. Tóth, “Phone recognition with hierarchical convolutional deep maxout networks,” *Eurasip J. Audio Speech Music Process.*, vol. 2015, no. 1, pp. 1–13, 2015.

[76] N. Zeghidour, N. Usunier, I. Kokkinos, T. Schaiz, G. Synnaeve, and E. Dupoux, “Learning filterbanks from raw speech for phone recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5509–5513.

[77] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, “Light gated recurrent units for speech recognition,” *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 2, no. 2, pp. 92–102, Apr. 2018.

[78] M. Gales, “Maximum likelihood linear transformations for hmm-based speech recognition,” *Comput. Speech Lang.*, vol. 12, no. 2, pp. 75–98, 1998.

[79] W.-N. Hsu et al., “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, Oct. 2021.

Xianghu Yue received the B.E. degree in control science and engineering from the Beijing Institute of Technology, Beijing, China, in 2016. He is currently working toward the Ph.D. degree with Human Language Technology Laboratory, Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His research interests include self-supervised speech representation learning and automatic speech recognition.

Jingru Lin received the B.Sc. degree in data science and analytics in 2021 from the National University of Singapore, Singapore, where she is currently working toward the Ph.D. degree with Human Language Technology Laboratory, Department of Electrical and Computer Engineering. She is currently a Research Engineer with the Human Language Technology Laboratory, Department of Electrical and Computer Engineering, National University of Singapore. Her research interests include automatics speech recognition and speech separation.

Fabian Ritter Gutierrez received the B.Eng. degree in acoustic engineering from the Universidad Austral de Chile, Valdivia, Chile, in 2016, and the M.Sc. degree in artificial intelligence from The University of Edinburgh, Edinburgh, U.K., in 2019. He is currently a Research Engineer with the Human Language Technology Laboratory, Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His research interests include self-supervised learning for speech processing and automatic speech recognition for domain shifts.

Haizhou Li (Fellow, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electrical and electronic engineering from the South China University of Technology, Guangzhou, China, in 1984, 1987, and 1990, respectively. He is currently a Professor with the School of Data Science, The Chinese University of Hong Kong, Shenzhen, China, and the Department of Electrical and Computer Engineering, National University of Singapore (NUS), Singapore. His research interests include automatic speech recognition, speaker and language recognition, and natural language processing. Prior to joining NUS, he taught with the University of Hong Kong, Hong Kong, during 1988–1990 and South China University of Technology, during 1990–1994. He was a Visiting Professor with CRIN, France, during 1994–1995, a Research Manager with Apple-ISS Research Centre during 1996–1998, the Research Director with Lernout & Hauspie Asia Pacific during 1999–2001, the Vice President with InfoTalk Corp. Ltd., during 2001–2003, and the Principal Scientist and Department Head of human language technology with the Institute for Infocomm Research, Singapore, during 2003–2016. He was the Editor-in-Chief of the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING during 2015–2018, a Member of the Editorial Board of Computer Speech and Language during 2012–2018. He was an Elected Member of IEEE Speech and Language Processing Technical Committee during 2012–2014, the President of International Speech Communication Association during 2015–2017, the President of Asia Pacific Signal and Information Processing Association during 2015–2016, and the President of the Asian Federation of Natural Language Processing during 2017–2018. He was the General Chair of ACL 2012, INTERSPEECH 2014, ASRU 2019, and ICASSP 2022. Dr. Li is a Fellow of ISCA. He was the recipient of the National Infocomm Award 2002 and the President’s Technology Award 2013 in Singapore. He was named one of the two Nokia Visiting Professors in 2009 by the Nokia Foundation, and Bremen Excellence Chair Professor in 2019.