

HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia,
Ruslan Salakhutdinov, Abdelrahman Mohamed

Abstract—Self-supervised approaches for speech representation learning are challenged by three unique problems: (1) there are multiple sound units in each input utterance, (2) there is no lexicon of input sound units during the pre-training phase, and (3) sound units have variable lengths with no explicit segmentation. To deal with these three problems, we propose the Hidden-Unit BERT (HuBERT) approach for self-supervised speech representation learning, which utilizes an offline clustering step to provide aligned target labels for a BERT-like prediction loss. A key ingredient of our approach is applying the prediction loss over the masked regions only, which forces the model to learn a combined acoustic and language model over the continuous inputs. HuBERT relies primarily on the consistency of the unsupervised clustering step rather than the intrinsic quality of the assigned cluster labels. Starting with a simple k-means teacher of 100 clusters, and using two iterations of clustering, the HuBERT model either matches or improves upon the state-of-the-art wav2vec 2.0 performance on the Librispeech (960h) and Libri-light (60,000h) benchmarks with 10min, 1h, 10h, 100h, and 960h fine-tuning subsets. Using a 1B parameter model, HuBERT shows up to 19% and 13% relative WER reduction on the more challenging dev-other and test-other evaluation subsets.¹

Index Terms—Self-supervised learning, BERT.

I. INTRODUCTION

The north star for many research programs has been learning speech and audio representations through listening and interaction, similar to how babies learn their first language. High fidelity speech representation includes disentangled aspects of the spoken content along with non-lexical information of how it is delivered, e.g., speaker identity, emotion, hesitation, interruptions. Furthermore, reaching a complete situational understanding requires modeling structured noise interleaving and overlapping with the speech signal, e.g., laughter, coughing, lip-smacking, background vehicle engine, birds chirping, or food sizzling sounds.

The need for such high-fidelity representations drove research in self-supervised learning for speech and audio where the targets driving the learning process of a designed pretext task are drawn from the input signal itself. Examples of pretext tasks for self-supervised speech representation learning include distinguishing near-by features from temporally distant ones [1]–[3], next-step prediction of audio features [4], masked prediction of audio features given unmasked context [5], [6]. Besides, self-supervised learning methods do not rely on any linguistic resources during training, allowing them to learn

universal representations since labels, annotations, and text-only material ignores rich information in the input signal.

Learning speech representations without reliance on large volumes of labeled data is crucial for industrial applications and products with ever-increasing coverage of new languages and domains. The time needed to collect large labeled datasets covering each of these scenarios is the real bottleneck in the current fast-moving AI industry, with time-to-market playing a critical role for product success. Building more inclusive applications covering spoken-only dialects and languages is another significant benefit of reducing dependence on linguistic resources. Given their non-standard orthographic rules, many of these languages and dialects have very little or no resources at all.

Pseudo-labeling (PL), also known as self-training and belongs to the family of semi-supervised learning techniques, has been the dominant approach for utilizing unlabeled speech and audio with successful applications dating back to the mid-1990s [7]–[10]. PL starts with some supervised data to train a “teacher” model in one specific downstream task. Pseudo-labels are then generated for the unlabeled data using the teacher model. Next, a student model is trained using the combined supervised and teacher-labeled data either using the standard cross-entropy [9] loss or using a contrastive loss [11] to account for noise in teacher-generated labels. The pseudo-labeling process may be repeated multiple times to improve teacher label quality [12] iteratively.

Without discounting the immense success of pseudo-labeling techniques, self-supervised representations offer two unique advantages: (1) Pseudo-label methods force student models to merely mimic a teacher model, which is limited by its supervised data size and the provided annotation quality. On the other hand, self-supervised pretext tasks force the model to represent the entire input signal by compressing much more bits of information into the learned latent representation. (2) In pseudo-labeling, the supervised data of the teacher model forces the whole learning to be geared towards a single downstream task. On the contrary, self-supervised features show better generalization to a multitude of downstream applications.

There have been impressive successes for self-supervised learning in Computer Vision (CV) [13]–[15] and Natural Language Processing (NLP) [16]–[18] applications. Learning representations of discrete input sequences, such as in Natural Language Processing (NLP) applications, uses either masked prediction [19], [20] or auto-regressive generation [18], [21] of input sequences with partial obfuscation. For continuous

¹The code, pre-trained and fine-tuned models are available at <https://github.com/pytorch/fairseq/tree/master/examples/hubert>.

inputs, such as in Computer Vision (CV) applications, representations are often learned through instance classification, in which each image and its augmentations are treated as a single output class to be pulled together [14], [15] or contrasted against other negative samples [22].

Speech signals differ from text and images in that they are *continuous-valued sequences*. Self-supervised learning for the speech recognition domain faces unique challenges from those in CV and NLP. Firstly, the presence of multiple sounds in each input utterance breaks the instance classification assumption used in many CV pre-training approaches. Secondly, during pre-training, there is no prior lexicon of discrete sound units available, as in NLP applications in which words or word pieces are used, hindering the use of predictive losses. Lastly, the boundaries between sound units are not known, which complicates masked prediction pre-training.

In this paper, we introduce **Hidden unit BERT (HuBERT)** that benefits from an offline clustering step to generate noisy labels for a BERT-like pre-training. Concretely, a BERT model consumes masked continuous speech features to predict pre-determined cluster assignments. The predictive loss is only applied over the masked regions, forcing the model to learn good high-level representations of unmasked inputs to infer the targets of masked ones correctly. Intuitively, the HuBERT model is forced to learn both acoustic and language models from continuous inputs. First, the model needs to model unmasked inputs into meaningful continuous latent representations, which maps to the classical acoustic modeling problem. Second, to reduce the prediction error, the model needs to capture the long-range temporal relations between learned representations. One crucial insight motivating this work is the importance of consistency of the targets, not just their correctness, which enables the model to focus on modeling the sequential structure of input data. Our approach draws inspiration from the DeepCluster method for self-supervised visual learning [23]; however, HuBERT benefits from the masked prediction loss over speech sequences to represent their sequential structure.

When the HuBERT model is pre-trained on either the standard Librispeech 960h [24] or the Libri-Light 60k hours [25], it either matches or improves upon the state-of-the-art wav2vec 2.0 [6] performance on all fine-tuning subsets of 10mins, 1h, 10h, 100h, and 960h. We present systematic results on three model sizes pre-trained with HuBERT: BASE (90M parameters), LARGE (300M), and X-LARGE (1B). The X-LARGE model shows up to 19% and 13% relative WER improvement from LARGE models on dev-other and test-other evaluation subsets when pre-trained on the Libri-Light 60k hours.

II. METHOD

A. Learning the Hidden Units for HuBERT

An acoustic model trained on text and speech pairs provides pseudo-phonetic labels for each frame via forced alignment in semi-supervised learning. On the contrary, the self-supervised representation learning setup has access to speech-only data. Nevertheless, simple discrete latent variable models such as

k-means and Gaussian mixture models (GMMs) infer hidden units that exhibit non-trivial correlation with the underlying acoustic units [26] (see also Table V). More advanced systems can achieve better acoustic unit discovery performance using better graphical models [27], [28] or parameterizes the distributions with more powerful neural network models [29]–[33].

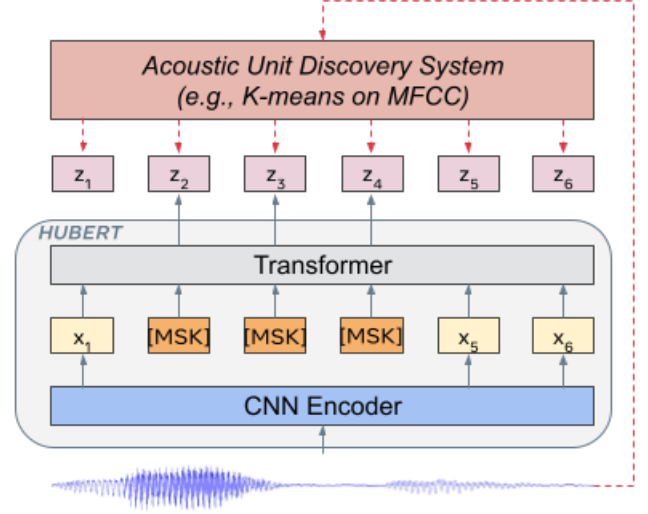


Fig. 1: The HuBERT approach predicts hidden cluster assignments of the masked frames (y_2, y_3, y_4 in the figure) generated by one or more iterations of k-means clustering.

Inspired by this, we propose to use acoustic unit discovery models to provide frame-level targets. Let X denote a speech utterance $X = [x_1, \dots, x_T]$ of T frames. Discovered hidden units are denoted with $h(X) = Z = [z_1, \dots, z_T]$, where $z_t \in [C]$ is a C -class categorical variable and h is a clustering model, e.g. k-means.

B. Representation Learning via Masked Prediction

Let $M \subset [T]$ denote the set of indices to be masked for a length- T sequence X , and $\tilde{X} = r(X, M)$ denote a corrupted version of X where x_t is replaced with a mask embedding \tilde{x} if $t \in M$. A masked prediction model f takes as input \tilde{X} and predicts a distribution over the target indices at each timestep $p_f(\cdot | \tilde{X}, t)$. There are two decisions to be made for masked prediction: *how to mask* and *where to apply the prediction loss*.

Regarding the first decision, we adopt the same strategies used in SpanBERT [34] and wav2vec 2.0 [6] for mask generation, where $p\%$ of the timesteps are randomly selected as start indices, and spans of l steps are masked. To address the second decision, we denote the cross-entropy loss computed over masked and unmasked timesteps as L_m and L_u , respectively. L_m is defined as:

$$L_m(f; X, M, Z) = \sum_{t \in M} \log p_f(z_t | \tilde{X}, t), \quad (1)$$

and L_u is of the same form except that it sums over $t \notin M$. The final loss is computed as a weighted sum of the two

terms: $L = \alpha L_m + (1 - \alpha) L_u$. In the extreme case when $\alpha = 0$, the loss is computed over the unmasked timesteps, which is similar to acoustic modeling in hybrid speech recognition systems [35]–[38]. In our setup, this limits the learning process to mimicking the clustering model.

In the other extreme with $\alpha = 1$, the loss is only computed over the masked timesteps where the model has to predict the targets corresponding to the unseen frames from context, analogous to language modeling. It forces the model to learn both the acoustic representation of unmasked segments and the long-range temporal structure of the speech data. We hypothesize that the setup with $\alpha = 1$ is more resilient to the quality of cluster targets, which is demonstrated in our experiments (see Table V).

C. Learning with Cluster Ensembles

A simple idea to improve target quality is to utilize multiple clustering models. While an individual clustering model may perform terribly, cluster ensembles can provide complementary information to facilitate representation learning. For example, an ensemble of k-means models with different codebook sizes can create targets of different granularity, from manner classes (vowel/consonant) to sub-phone states (senones). To extend the proposed framework, let $Z^{(k)}$ be the target sequences generated by the k -th clustering model. We can now re-write L_m as:

$$L_m(f; X, \{Z^{(k)}\}_k, M) = \sum_{t \in M} \sum_k \log p_f^{(k)}(z_t^{(k)} | \tilde{X}, t), \quad (2)$$

and similarly for the unmasked loss L_u . This is analogous to multi-task learning, but with tasks created by unsupervised clustering.

Additionally, ensembling is intriguing because it can be used alongside product quantization (PQ) [39], where a feature space is partitioned into multiple subspaces, and each subspace is quantized separately. PQ allows effective Euclidean distance-based quantization such as k-means for high-dimensional features and heterogeneous features whose scale differs significantly between subspaces. In this case, the theoretical size of the target space is the product of all codebooks' sizes.

D. Iterative Refinement of Cluster Assignments

In addition to using cluster ensembles, another direction for improved representation is *refining* the cluster assignments throughout the learning process. Since we expect a pre-trained model to provide better representations than the raw acoustic feature such as MFCCs, we can create a new generation of clusters by training a discrete latent model over the learned latent representations. The learning process then proceeds with the newly discovered units.

E. Implementation

Our pre-trained models follows the wav2vec 2.0 architecture [6], with a convolutional waveform encoder, a BERT encoder [19], a projection layer and a code embedding layer.

We consider HuBERT in three different configurations: BASE, LARGE, and X-LARGE. The first two follow the architectures of wav2vec 2.0 BASE and LARGE closely. The X-LARGE architecture expands the model size to about 1 billion parameters, similar to the size of the Conformer XXL model in [40]. The waveform encoder is identical for all the three configurations, which is composed of seven 512-channel layers with strides [5, 2, 2, 2, 2, 2] and kernel widths [10, 3, 3, 3, 3, 2, 2]. The BERT encoder consists of many identical transformer blocks, whose parameters along with the parameter of the subsequent projection layer are specified in Table I.

		BASE	LARGE	X-LARGE
CNN Encoder	strides	5, 2, 2, 2, 2, 2, 2		
	kernel width channel	10, 3, 3, 3, 3, 2, 2 512		
Transformer	layer	12	24	48
	embedding dim.	768	1024	1280
	inner FFN dim.	3072	4096	5120
	layerdrop prob	0.05	0	0
	attention heads	8	16	16
Projection	dim.	256	768	1024
Num. of Params		95M	317M	964M

TABLE I: Model architecture summary for BASE, LARGE, and X-LARGE HuBERT models

The convolutional waveform encoder generates a feature sequence at a 20ms framerate for audio sampled at 16kHz (CNN encoder down-sampling factor is 320x). The audio encoded features are then randomly masked as described in Section II-B. The BERT encoder takes as input the masked sequence and outputs a feature sequence $[o_1, \dots, o_T]$. The distribution over codewords is parameterized with

$$p_f^{(k)}(c | \tilde{X}, t) = \frac{\exp(\text{sim}(A^{(k)} o_t, e_c) / \tau)}{\sum_{c'=1}^C \exp(\text{sim}(A^{(k)} o_t, e_{c'}) / \tau)}, \quad (3)$$

where A is the projection matrix, e_c is the embedding for codeword c , $\text{sim}(\cdot, \cdot)$ computes the cosine similarity between two vectors, and τ scales the logit, which is set to 0.1. When cluster ensembles are used, one projection matrix $A^{(k)}$ is applied for each clustering model k .

After HuBERT pre-training, We use the connectionist temporal classification (CTC) [41] loss for ASR fine-tuning of the whole model weights except the convolutional audio encoder, which remains frozen. The projection layer(s) is removed and replaced with a randomly initialized softmax layer. The CTC target vocabulary includes 26 English characters, a space token, an apostrophe, and a special CTC blank symbol.

III. RELATED WORK

We discuss recent studies on self-supervised speech representation learning by grouping them by training objective. The earliest line of work learns representations by postulating a generative model for speech with latent variables, which are assumed to capture the relevant phonetic information. Training of these models amounts to likelihood maximization. Different latent structures have been applied to encode the

prior assumption, such as continuous [29], discrete [31], [42], or sequential [28], [30], [32], [33], [43].

Prediction-based self-supervised learning has gathered increasing interests recently, where a model is tasked to predict the content of the unseen regions [4], [44]–[50] or to contrast the target unseen frame with randomly sampled ones [1]–[3], [6]. Some models combine both the predictive and the contrastive losses [5], [51]. These objectives can usually be interpreted as mutual information maximization [52]. Other objectives do not belong to these categories, for example, [53].

This work is most related to DiscreteBERT [51]: both HuBERT and DiscreteBERT predict discrete targets of masked regions. However, there are several crucial differences. First, instead of taking quantized units as input, HuBERT takes raw waveforms as input to pass as much information as possible to the transformer layers, which was shown to be important in [6]. Furthermore, in the experiment section, we show that our model, with simple k-means targets, can achieve better performance than DiscreteBERT that uses vq-wav2vec [5] learned units. Second, we also present many techniques to improve teacher quality instead of using a single fixed teacher as done in DiscreteBERT.

HuBERT is also related to wav2vec 2.0 [6]. However, the latter employs a contrastive loss that requires careful design of where to sample negative frames from, an auxiliary diversity loss to encourage the discrete unit usage, and demands a proper Gumbel-softmax temperature annealing schedule. In addition, it only explores quantizing the waveform encoder output, which may not be the best feature for quantization due to the limited capacity of the convolutional encoder, as suggested by our ablation studies in Figure 2. Concretely, our proposed method adopts a more direct predictive loss by separating the acoustic unit discovery step from the masked prediction representation learning phase and achieves the state-of-the-art results that match or outperform wav2vec 2.0 on different fine-tuning scales.

Finally, the idea of iterative refinement target labels is similar to iterative pseudo labeling for semi-supervised ASR [12], [54], which leverages an improving student model to generate better pseudo-labels for the next iteration of training. The HuBERT approach can be seen as extending this method to the self-supervised setup with a masked prediction loss.

IV. EXPERIMENTAL DETAILS

A. Data

For unsupervised pre-training, we use the full 960 hours of LibriSpeech audio [24] or 60,000 hours of Libri-light [25] audio, both of which are derived from the LibriVox project that contains English recordings of copyright-free audiobooks by volunteers from the Internet. For supervised fine-tuning, five different partitions are considered: Libri-light 10-minute, 1-hour, 10-hour splits and LibriSpeech 100-hour (train-clean-100) and 960-hour (train-clean-100, train-clean-360, train-other-500 combined) splits. The three Libri-light splits are subsets of the LibriSpeech training split, and each of them contain half of the audio from train-clean-* and the other from train-other-500.

B. Unsupervised Unit Discovery

To demonstrate the effectiveness of the proposed method on utilizing low-quality cluster assignments, we consider the k-means algorithm [55] for acoustic unit discovery by default. It is one of the most naive unit discovery models that can be treated as modeling an isotropic Gaussian with the same scalar variance for each acoustic unit. To generate labels for the first iteration HuBERT training over the 960 hour LibriSpeech training set, we run k-means clustering with 100 clusters on 39-dimensional MFCC features, which are 13 coefficients with the first and the second-order derivatives.

To generate better targets for the subsequent iterations, we run k-means clustering with 500 clusters on the latent features extracted from the HuBERT model pre-trained in the previous iteration (not fine-tuned) at some intermediate transformer layer. Since the feature dimension at the transformer output is much higher than the MFCC features (768-D for HuBERT BASE), we cannot afford to load the entire 960 hour training split to the memory. So instead, we randomly sample 10% of the data for fitting the k-means model.

The MiniBatchKMeans algorithm implemented in the `scikit-learn` [56] package is used for clustering, which fits a mini-batch of samples at a time.² We set the mini-batch size to be 10,000 frames. k-means++ [57] with 20 random starts is used for better initialization.

C. Pre-Training

We train the BASE model for two iterations on the 960 hours of LibriSpeech audio on 32 GPUs, with a batch size of at most 87.5 seconds of audio per GPU. The first iteration is trained for 250k steps, while the second iteration is trained for 400k steps using labels generated by clustering the 6-th transformer layer output of the first iteration model. Training for 100k steps takes about 9.5 hours.

Next we train HuBERT LARGE and X-LARGE for one iteration on 60,000 hours of Libri-light audio on 128 and 256 GPUs, respectively, for 400k steps. The batch sizes are reduced to 56.25 and 22.5 seconds of audio per GPU due to memory constraints. Instead of restarting the iterative process from clustering MFCC features, we extract features from the 9-th transformer layer of the second iteration BASE HuBERT for clustering and use those labels for training these two models. Hence, these two models can also be seen as the third iteration models.

For all HuBERT configurations, mask span is set to $l = 10$, and $p = 8\%$ of the waveform encoder output frames are randomly selected as mask start if not otherwise mentioned. Adam [58] optimizer is used with $\beta = (0.9, 0.98)$, and the learning rate ramps up linearly from 0 to the peak learning rate for the first 8% of the training steps, and then decays linearly back to zero. The peak learning rates are $5e-4/1.5e-3/3e-3$ for BASE/LARGE/X-LARGE models.

D. Supervised Fine-Tuning and Decoding

We fine-tune each model on 8 GPUs on the labeled splits described in Section IV-A. The batch sizes per GPU are at

²It still requires loading the entire dataset to the memory first.

most 200/80/40 seconds of audio for BASE/LARGE/X-LARGE models. During fine-tuning, the convolutional waveform audio encoder parameters are fixed. Like wav2vec 2.0, we introduce a *freeze-step* hyperparameter to control how many fine-tuning steps the transformer parameters are fixed, and only the new softmax matrix is trained. We sweep over peak learning rate ([1e-5, 1e-4]), learning rate schedule (percentage of steps for linear ramp-up and decay), number of fine-tuning steps, freeze step, and waveform encoder output masking probability for each model size and fine-tuning split combination using the word error rate (WER) on the dev-other subset as a criterion for model selection.

We use the wav2letter++ [59] beam search decoder wrapped in Fairseq [60] for language model-fused decoding, which optimizes:

$$\log p_{CTC}(Y | X) + w_1 \log P_{LM}(Y) + w_2 |Y|, \quad (4)$$

where Y is the predicted text, $|Y|$ is the length of the text, and w_1 and w_2 denote the language model weight and word score. The decoding hyperparameters are searched with Ax, a Bayesian optimization toolkit.³ In this work, we consider both n -gram and transformer language models trained on the official Librispeech language modeling data.

E. Metrics of Target Quality

For analysis, we derive frame-level forced-aligned phonetic transcripts using a hybrid ASR system to measure the correlation between the k-means cluster assignments and the actual phonetic units. Given aligned frame-level phonetic labels $[y_1, \dots, y_T]$ and k-means labels $[z_1, \dots, z_T]$, the joint distribution between the two variables $p_{yz}(i, j)$ can be estimated by counting the occurrences:

$$p_{yz}(i, j) = \frac{\sum_{t=1}^T [y_t = i \wedge z_t = j]}{T}, \quad (5)$$

where i denotes the i -th phoneme class and j denotes the j -th k-means label class. The marginal probabilities are computed as $p_z(j) = \sum_i p_{yz}(i, j)$ and $p_y(i) = \sum_j p_{yz}(i, j)$.

For each phone class i , we further compute the most likely target label as:

$$z^*(i) = \arg \max_j p_{yz}(i, j). \quad (6)$$

Likewise, for each k-means class j , we compute the most likely phone label as:

$$y^*(j) = \arg \max_i p_{yz}(i, j). \quad (7)$$

Three metrics are considered:

1) **phone purity** (Phn Pur.):

$$\mathbb{E}_{p_z(j)} [p_{y|z}(y^*(j) | j)], \quad (8)$$

where $p_{y|z}(i | j) = p_{yz}(i, j)/p_z(j)$ denotes the conditional probability of phone given a k-means label. This metric measures the average phone purity within one class, which can be interpreted as the frame-level phone accuracy if we transcribe each k-means class with its

most likely phone label. When comparing different sets of target labels with the same number of units, higher purity indicates better quality. However, this metric is less meaningful when comparing two sets with different numbers of units: in the extreme case where each frame is assigned a unique target label, the phone purity would be 100%.

2) **cluster purity** (Cls Pur.):

$$\mathbb{E}_{p_y(i)} [p_{z|y}(z^*(i) | i)], \quad (9)$$

where $p_{z|y}(j | i) = p_{yz}(i, j)/p_y(i)$ denotes the conditional probability of a k-means label given phone label. Cluster purity is the counterpart of phone purity, whose value would typically decrease when the number of units increases. When comparing target labels with the same number of units, higher cluster purity also indicates a better quality, as frames of the same phone are more likely labeled as the same k-means label class.

3) **phone-normalized mutual information** (PNMI):

$$\frac{I(y; z)}{H(y)} = \frac{\sum_i \sum_j p_{yz}(i, j) \log \frac{p_{yz}(i, j)}{p_y(i)p_z(j)}}{\sum_i p_y(i) \log p_y(i)} \quad (10)$$

$$= \frac{H(y) - H(y | z)}{H(y)} \quad (11)$$

$$= 1 - \frac{H(y | z)}{H(y)}. \quad (12)$$

PNMI is an information-theoretic metric that measures the percentage of uncertainty about the phone label y eliminated after observing the k-means label z . Higher PNMI also indicates better k-means clustering quality.

V. RESULTS

A. Main Results: Low- and High-Resource Setups

Table II presents results for the low-resource setup, where pre-trained models are fine-tuned on 10 minutes, 1 hour, 10 hours, or 100 hours of labeled data. We include comparison with semi-supervised (iterative pseudo labeling (IPL) [12], slimIPL [54], noisy student [61]) and self-supervised approaches (DeCoAR 2.0 [50], DiscreteBERT [51], wav2vec 2.0 [6]) in the literature. Increasing the amount of unlabeled data and increasing the model size improve performance, demonstrating the scalability of the proposed HuBERT self-supervised pre-training method. In the ultra-low resource setup with just 10 minutes of labeled data, the HuBERT LARGE model can achieve a WER of 4.7% on the test-clean set and 7.6% on the test-other set, which is 0.1% and 0.6% WER lower, respectively than the state-of-the-art wav2vec 2.0 LARGE model. By further scaling up the model size to 1B parameters, the HuBERT X-LARGE model can further reduce the WER to 4.6% and 6.8% on test-clean and test-other. The superiority of HuBERT persists across setups with different amounts of labeled data, with the only exceptions being fine-tuning on 100 hours of labeled data, where HuBERT LARGE is 0.1% WER higher than wav2vec 2.0 LARGE on test-clean, and HuBERT BASE is 0.1% WER higher than wav2vec 2.0

³<https://github.com/facebook/Ax>

Model	Unlabeled Data	LM	dev-clean	dev-other	test-clean	test-other
10-min labeled						
DiscreteBERT [51]	LS-960	4-gram	15.7	24.1	16.3	25.2
wav2vec 2.0 BASE [6]	LS-960	4-gram	8.9	15.7	9.1	15.6
wav2vec 2.0 LARGE [6]	LL-60k	4-gram	6.3	9.8	6.6	10.3
wav2vec 2.0 LARGE [6]	LL-60k	Transformer	4.6	7.9	4.8	8.2
HUBERT BASE	LS-960	4-gram	9.1	15.0	9.7	15.3
HUBERT LARGE	LL-60k	4-gram	6.1	9.4	6.6	10.1
HUBERT LARGE	LL-60k	Transformer	4.3	7.0	4.7	7.6
HUBERT X-LARGE	LL-60k	Transformer	4.4	6.1	4.6	6.8
1-hour labeled						
DeCoAR 2.0 [50]	LS-960	4-gram	-	-	13.8	29.1
DiscreteBERT [51]	LS-960	4-gram	8.5	16.4	9.0	17.6
wav2vec 2.0 BASE [6]	LS-960	4-gram	5.0	10.8	5.5	11.3
wav2vec 2.0 LARGE [6]	LL-60k	Transformer	2.9	5.4	2.9	5.8
HUBERT BASE	LS-960	4-gram	5.6	10.9	6.1	11.3
HUBERT LARGE	LL-60k	Transformer	2.6	4.9	2.9	5.4
HUBERT X-LARGE	LL-60k	Transformer	2.6	4.2	2.8	4.8
10-hour labeled						
SlimIPL [54]	LS-960	4-gram + Transformer	5.3	7.9	5.5	9.0
DeCoAR 2.0 [50]	LS-960	4-gram	-	-	5.4	13.3
DiscreteBERT [51]	LS-960	4-gram	5.3	13.2	5.9	14.1
wav2vec 2.0 BASE [6]	LS-960	4-gram	3.8	9.1	4.3	9.5
wav2vec 2.0 LARGE [6]	LL-60k	Transformer	2.4	4.8	2.6	4.9
HUBERT BASE	LS-960	4-gram	3.9	9.0	4.3	9.4
HUBERT LARGE	LL-60k	Transformer	2.2	4.3	2.4	4.6
HUBERT X-LARGE	LL-60k	Transformer	2.1	3.6	2.3	4.0
100-hour labeled						
IPL [12]	LL-60k	4-gram + Transformer	3.19	6.14	3.72	7.11
SlimIPL [54]	LS-860	4-gram + Transformer	2.2	4.6	2.7	5.2
Noisy Student [61]	LS-860	LSTM	3.9	8.8	4.2	8.6
DeCoAR 2.0 [50]	LS-960	4-gram	-	-	5.0	12.1
DiscreteBERT [51]	LS-960	4-gram	4.0	10.9	4.5	12.1
wav2vec 2.0 BASE [6]	LS-960	4-gram	2.7	7.9	3.4	8.0
wav2vec 2.0 LARGE [6]	LL-60k	Transformer	1.9	4.0	2.0	4.0
HUBERT BASE	LS-960	4-gram	2.7	7.8	3.4	8.1
HUBERT LARGE	LL-60k	Transformer	1.8	3.7	2.1	3.9
HUBERT X-LARGE	LL-60k	Transformer	1.7	3.0	1.9	3.5

TABLE II: Results and comparison with the literature on low resource setups (10-min, 1-hour, 10-hour, and 100-hour of labeled data).

BASE on test-other. In addition, HuBERT also outperforms DiscreteBERT by a large margin in all setups, while both are trained with a virtually identical objective - masked prediction of discovered units. The considerable performance gap suggests two things. First, using waveform as the input to the model is crucial for avoiding loss of information during quantization. Second, while vq-wav2vec [5], the units that DiscreteBERT uses for training, may discover better units than k-means clustering of MFCC features, the proposed iterative refinement benefits from the improving HuBERT model and learn better units eventually. We will verify these statements in the ablation study sections.

We report results of fine-tuning HuBERT models on the full 960 hours of Librispeech data and compare with the literature in Table III. Prior studies using additional unpaired speech are classified into:

- 1) self-training: first train an ASR on labeled data to annotate unlabeled speech, and then combine both golden and ASR-annotated text-speech pairs for supervised training.
- 2) pre-training: first use unlabeled speech for pre-training a

model, and then fine-tune the model on labeled data with a supervised training objective.

- 3) pre-training + self-training: first pre-train and fine-tune a model, and then use it to annotate unlabeled speech for self-training combined with supervised data.

HuBERT outperforms the state-of-the-art supervised and self-training methods and is on par with the two best pre-training results in the literature; both are based on wav2vec 2.0 contrastive learning. In contrast, it lags behind methods combining pre-training with self-training. However, as observed in [63] and [40], we expect that HuBERT can achieve comparable or better performance after combining with self-training, since the pre-trained HuBERT model is on par or better than the pre-trained model those two methods use for pseudo labeling.

B. Analysis: K-Means Stability

To better understand why masked prediction of discovered units is effective, we conduct a series of analyses and ablation studies. We start with probing the stability of the k-means clustering algorithm concerning different numbers of clusters and

Model	Unlabeled Data	LM	dev-clean	dev-other	test-clean	test-other
Supervised						
Conformer L [62]	-	LSTM	-	-	1.9	3.9
Self-Training						
IPL [12]	LL-60k	4-gram + Transformer	1.85	3.26	2.10	4.01
Noisy Student [61]	LV-60k	LSTM	1.6	3.4	1.7	3.4
Pre-Training						
wav2vec 2.0 LARGE [6]	LL-60k	Transformer	1.6	3.0	1.8	3.3
pre-trained Conformer XXL [40]	LL-60k	LSTM	1.5	3.0	1.5	3.1
Pre-Training + Self-Training						
wav2vec 2.0 + self-training [63]	LL-60k	Transformer	1.1	2.7	1.5	3.1
pre-trained Conformer XXL + Noisy Student [40]	LL-60k	LSTM	1.3	2.6	1.4	2.6
This work (Pre-Training)						
HUBERT LARGE	LL-60k	Transformer	1.5	3.0	1.9	3.3
HUBERT X-LARGE	LL-60k	Transformer	1.5	2.5	1.8	2.9

TABLE III: Comparison with the literature on high resource setups using all 960 hours of labeled LibriSpeech data.

different sizes of its training data. Two features are considered: 39-dimensional MFCC features and 768-dimensional output from the 6-th transformer layer of the first iteration HuBERT-BASE model. These two features are used to produce cluster assignments for the first and the second iteration HUBERT training, respectively.

For k-means clustering, we consider $K = \{100, 500\}$ clusters fitted on $\{1, 10, 100\}$ hours of speech sampled from the LibriSpeech training split. Each combination of the hyperparameters and the features are trained for 10 trials, and the mean and standard deviation of the supervised PNMI metric on the development set (combining dev-clean and dev-other from LibriSpeech) is reported in Table IV. The results show that the k-means clustering is reasonably stable given the small standard deviations across different hyperparameters and features. Furthermore, increasing the amount of data used for fitting k-means models improves PNMI in general, but the gain is only as much as 0.012, suggesting the feasibility of using k-means for unit discovery even with limited CPU memory relative to the feature matrix size. Lastly, the PNMI score is much higher when clustering on HuBERT features than clustering on MFCC features, and the gap is even larger with 500 clusters, indicating that iterative refinement significantly improves the clustering quality.

feature	C	PNMI (mean \pm std) with K-means Training Size =		
		1h	10h	100h
MFCC	100	0.251 \pm 0.001	0.253 \pm 0.001	0.253 \pm 0.001
	500	0.283 \pm 0.001	0.285 \pm 0.000	0.287 \pm 0.001
BASE-it1-L6	100	0.563 \pm 0.012	0.561 \pm 0.012	0.575 \pm 0.008
	500	0.680 \pm 0.005	0.684 \pm 0.003	0.686 \pm 0.004

TABLE IV: Stability of K-means as an unsupervised unit discovery algorithm with respect to different features, numbers of clusters, and training data sizes. PNMI stands for phone-normalized mutual information.

C. Analysis: Clustering Quality Across Layers and Iterations

We next study how each layer of the HuBERT model from each iteration performs when used for clustering to generate

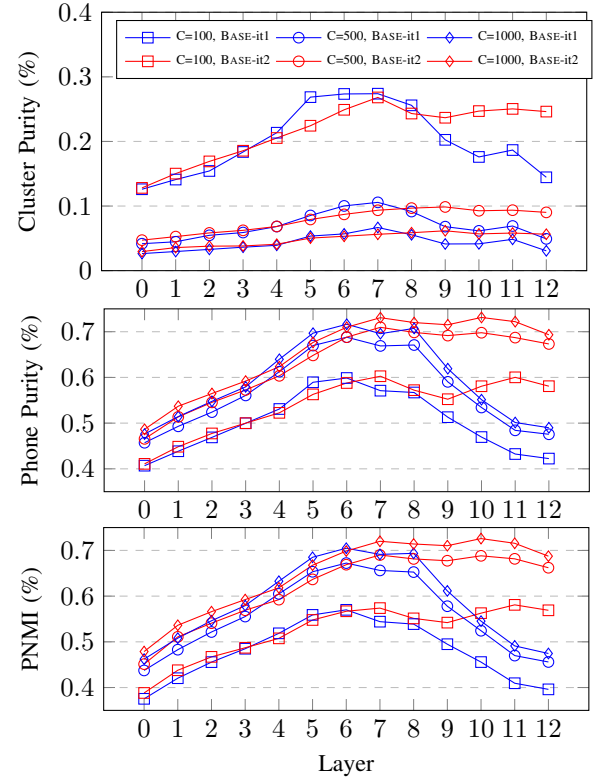


Fig. 2: Quality of the cluster assignments obtained by running k-means clustering on features extracted from each transformer layer of the first and the second iteration BASE HuBERT models.

training targets. The two BASE HuBERT models from the first two iterations as described in Section IV-C are considered, which are referred to as BASE-it1 and BASE-it2, respectively. There are 26 features representing 12 transformer layers plus the input to the first transformer layer (denoted as “Layer 0”) from the two HuBERT models. For each feature, we fit three k-means models ($K = \{100, 500, 1000\}$ clusters) on a 100 hour subset randomly sampled from the LibriSpeech training data.

teacher	C	PNMI	dev-other WER (%)		
			$\alpha = 1.0$	$\alpha = 0.5$	$\alpha = 0.0$
Chenone (supervised top-line)	8976	0.809	10.38	9.16	9.79
K-means on MFCC	50	0.227	18.68	31.07	94.60
	100	0.243	17.86	29.57	96.37
	500	0.276	18.40	33.42	97.66
K-means on BASE-it1-layer6	500	0.637	11.91	13.47	23.29
K-means on BASE-it2-layer9	500	0.704	10.75	11.59	13.79

TABLE V: The effect of the training objective and clustering quality on performance. C refers to the number of units, and α is the weight for masked frames.

The teacher quality measured in cluster purity, phone purity, and phone normalized mutual information (PNMI) is shown in Figure 2. As a baseline, MFCC achieves (cluster purity, phone purity, PNMI) = (0.099, 0.335, 0.255) for $K = 100$ and (0.031, 0.356, 0.287) for $K = 500$.

Both BASE-it1 and BASE-it2 features result in significantly better clustering quality on all three metrics than MFCC with the same number of clusters. On the other hand, the best BASE-it2 feature is better than the best BASE-it1 on phone purity and PNMI, but slightly worse on cluster purity. Finally, we observe different trends across layers from BASE-it1 and BASE-it2: while BASE-it2 model features generally improve over layers, BASE-it1 has the best features in the middle layers around the 6th layer. Interestingly, the quality of the last few layers degrades dramatically for BASE-it1, potentially because it is trained on target assignments of worse quality, and therefore the last few layers learn to mimic their bad label behavior.

D. Ablation: The Importance of Predicting Masked Frames

We present a series of ablation studies in the following sections to learn how pre-training objective, cluster quality, and hyperparameters affect the performance. The models for ablation studies are pre-trained for 100k steps and fine-tuned on the 10-hour libri-light split using fixed hyperparameters. MFCC-based k-means units with $C=100$ are used if not otherwise mentioned. We report WERs on the dev-other set decoded with the n -gram language model using fixed decoding hyperparameters.

To understand the importance of our proposal to predict the masked frames only, we compare three conditions: 1) predicting masked frames, 2) predicting all frames, and 3) predicting unmasked frames, which can be simulated by setting α to 1.0, 0.5, and 0.0, respectively. We are comparing three k-means models learned from clustering MFCC teachers with 50, 100, 500 clusters, one learned from clustering HuBERT-BASE-it1 6th transformer layer features, and supervised labels obtained from the forced-alignment of character-based HMM models (chenone) [64].

Results shown in Table V indicate that when learning from bad cluster assignments, computing loss only from the masked regions achieves the best performance, while the inclusion of unmasked loss results in significantly higher WERs. However, as the clustering quality improves, the model would suffer less when computing losses on the unmasked frames (BASE-

it1-layer6) or even achieve better performance as the case of chenone.

E. Ablation: The Effect of Cluster Ensembles

To understand the effect of combining multiple k-means models for generating targets, we consider two setups. The first one has k-means models of different numbers of clusters presented in Table V, denoted with KM- $\{50,100,500\}$. The second one has k-means models trained on spliced MFCC features with a window of three; hence, each input feature is represented as a 117-dimensional vector. In this second case, we apply product quantization on the spliced features, where dimensions are split into the coefficients of the zeroth, first, and second-order derivatives, with each 39-dimensional subspace quantized to a codebook of 100 entries. We denote these codebooks with Product k-means- $\{0,1,2\}$ -100, respectively. By comparing the results from Table V and Table VI, it is clear that using an ensemble leads to better performance than what a single k-means clustering can achieve.

F. Ablation: Impact of Hyperparameters

Figure 3 and Table VII studies how hyperparameters affect HuBERT pre-training. It is shown that (1) the portion of frames selected as mask start is optimal at $p = 8\%$; (2) increasing the batch size can significantly improve the performance; (3) training for longer consistently helps for both k-means models with $C=\{50, 100\}$, and the best model achieves a WER of 11.68%. These findings are also consistent with those from BERT-like models [20]. In addition, we include a comparable result from DiscreteBERT [51] in Table VII which applies k-means to quantize the same MFCC features into 13.5k units, used as both the output and the *input* to the BERT model. Besides using continuous speech input rather than discrete units, We hypothesize that HuBERT achieves significantly better performance because its fewer k-means clusters of 100 or 500 help capture broad phonetic concepts without delving into inter/intra-speaker variation.

VI. CONCLUSION

This paper presents HuBERT, a speech representation learning approach that relies on predicting K-means cluster assignments of masked segments of continuous input. On both the Librispeech 960 hours and the 60,000 hours Libri-light pre-training setups, HuBERT matches or outperforms the state-of-the-art systems over all fine-tuning subsets of 10mins, 1h,

teacher	WER
K-means $\{50,100\}$	17.81
K-means $\{50,100,500\}$	17.56
Product K-means-0-100	19.26
Product K-means-1-100	17.64
Product K-means-2-100	18.46
Product K-means- $\{0,1,2\}$ -100	16.73

TABLE VI: Cluster ensembles with k-means and product k-means.

teacher	C	dev-other WER (%)			
		steps=100k	250k	400k	800k
K-means	50	18.68	13.65	12.40	11.82
	100	17.86	12.97	12.32	11.68
[51]	13.5k	26.6			

TABLE VII: Varying the number of HuBERT pre-training steps. p is set to 6.5%.

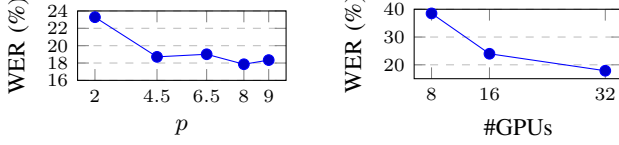


Fig. 3: Varying masking probability p (left) and effective batch size through the number of GPUs (right).

10h, 100h, and 960h. Furthermore, the learned representation quality improves dramatically with iteratively refining K-means cluster assignments using learned latent representations for a previous iteration. Finally, HuBERT scales well to a 1B transformer model showing a relative reduction in WER of up to 13% on the test-other subset. For future work, we plan to improve the HuBERT training procedure to consist of a single phase. Furthermore, given the high quality of its representations, we will consider using HuBERT pre-trained representations for multiple downstream recognition and generation tasks beyond ASR.

REFERENCES

- [1] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.
- [2] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition,” *arXiv preprint arXiv:1904.05862*, 2019.
- [3] E. Kharitonov, M. Rivière, G. Synnaeve, L. Wolf, P.-E. Mazaré, M. Douze, and E. Dupoux, “Data augmenting contrastive learning of speech representations in the time domain,” *arXiv preprint arXiv:2007.00991*, 2020.
- [4] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, “An unsupervised autoregressive model for speech representation learning,” *arXiv preprint arXiv:1904.03240*, 2019.
- [5] A. Baevski, S. Schneider, and M. Auli, “vq-wav2vec: Self-supervised learning of discrete speech representations,” *arXiv preprint arXiv:1910.05453*, 2019.
- [6] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *arXiv preprint arXiv:2006.11477*, 2020.
- [7] G. Zavaliagkos and T. Colthurst, “Utilizing untranscribed training data to improve performance,” in *DARPA Broadcast News Transcription and Understanding Workshop*, 1998.
- [8] J. Ma, S. Matsoukas, O. Kimball, and R. Schwartz, “Unsupervised training on large amounts of broadcast news data,” in *ICASSP*, 2006.
- [9] J. Kahn, A. Lee, and A. Hannun, “Self-training for end-to-end speech recognition,” in *ICASSP*, 2020.
- [10] W.-N. Hsu, A. Lee, G. Synnaeve, and A. Hannun, “Semi-supervised speech recognition via local prior matching,” *arXiv preprint arXiv:2002.10336*, 2020.
- [11] A. Xiao, C. Fuegen, and A. Mohamed, “Contrastive semi-supervised learning for asr,” *arXiv preprint arXiv:2103.05149*, 2021.
- [12] Q. Xu, T. Likhomanenko, J. Kahn, A. Hannun, G. Synnaeve, and R. Collobert, “Iterative pseudo-labeling for speech recognition,” *arXiv preprint arXiv:2005.09267*, 2020.
- [13] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, “Unsupervised learning of visual features by contrasting cluster assignments,” *CoRR*, vol. abs/2006.09882, 2020.
- [14] X. Chen and K. He, “Exploring simple siamese representation learning,” *CoRR*, vol. abs/2011.10566, 2020.
- [15] J. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Á. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, “Bootstrap your own latent: A new approach to self-supervised learning,” *CoRR*, vol. abs/2006.07733, 2020.
- [16] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” *CoRR*, vol. abs/2005.14165, 2020.
- [17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [18] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” *arXiv preprint arXiv:1910.13461*, 2019.
- [19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [20] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, “Electra: Pre-training text encoders as discriminators rather than generators,” *arXiv preprint arXiv:2003.10555*, 2020.
- [21] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” in *NAACL*, 2018.
- [22] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *CVPR*, 2020.
- [23] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, “Deep clustering for unsupervised learning of visual features,” in *ECCV*, 2018.
- [24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *ICASSP*, 2015.
- [25] J. Kahn *et al.*, “Libri-light: A benchmark for asr with limited or no supervision,” in *ICASSP*, 2020.
- [26] C.-y. Lee and J. Glass, “A nonparametric bayesian approach to acoustic model discovery,” in *ACL*, 2012.
- [27] L. Ondel, L. Burget, and J. Černocký, “Variational inference for acoustic unit discovery,” *Procedia Computer Science*, vol. 81, pp. 80–86, 2016.
- [28] J. Ebberts, J. Heymann, L. Drude, T. Glarner, R. Haeb-Umbach, and B. Raj, “Hidden markov model variational autoencoder for acoustic unit discovery,” in *INTERSPEECH*, 2017.
- [29] W.-N. Hsu, Y. Zhang, and J. Glass, “Learning latent representations for speech generation and transformation,” in *INTERSPEECH*, 2017.
- [30] —, “Unsupervised learning of disentangled and interpretable representations from sequential data,” in *NeurIPS*, 2017.
- [31] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, “Unsupervised speech representation learning using wavenet autoencoders,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 12, pp. 2041–2053, 2019.
- [32] S. Khurana, S. R. Joty, A. Ali, and J. Glass, “A factorial deep markov model for unsupervised disentangled representation learning from speech,” in *ICASSP*, 2019.
- [33] S. Khurana, A. Laurent, W.-N. Hsu, J. Chorowski, A. Lancucki, R. Marxer, and J. Glass, “A convolutional deep markov model for unsupervised speech representation learning,” *arXiv preprint arXiv:2006.02547*, 2020.
- [34] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, “Spanbert: Improving pre-training by representing and predicting spans,” *Transactions of the Association for Computational Linguistics*, 2020.
- [35] S. Young, “Large vocabulary continuous speech recognition: A review,” *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 45–57, 1996.
- [36] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn, “Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition,” in *2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP)*. IEEE, 2012, pp. 4277–4280.
- [37] D. Povey, “Discriminative training for large vocabulary speech recognition,” Ph.D. dissertation, University of Cambridge, 2005.
- [38] H. A. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*. Springer Science & Business Media, 2012, vol. 247.
- [39] R. M. Gray and D. L. Neuhoff, “Quantization,” *IEEE transactions on information theory*, vol. 44, no. 6, pp. 2325–2383, 1998.

- [40] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," *arXiv preprint arXiv:2010.10504*, 2020.
- [41] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006.
- [42] A. van den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," in *NeurIPS*, 2017.
- [43] T. Glarner, P. Hanebrink, J. Ebbers, and R. Haeb-Umbach, "Full bayesian hidden markov model variational autoencoder for acoustic unit discovery," in *INTERSPEECH*, 2018.
- [44] Y.-A. Chung and J. Glass, "Generative pre-training for speech with autoregressive predictive coding," in *ICASSP*, 2020.
- [45] —, "Improved speech representations with multi-target autoregressive predictive coding," *arXiv preprint arXiv:2004.05274*, 2020.
- [46] S. Ling, Y. Liu, J. Salazar, and K. Kirchhoff, "Deep contextualized acoustic representations for semi-supervised speech recognition," in *ICASSP*, 2020.
- [47] W. Wang, Q. Tang, and K. Livescu, "Unsupervised pre-training of bidirectional speech encoders via masked reconstruction," in *ICASSP*, 2020.
- [48] A. T. 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 *ICASSP*, 2020.
- [49] P.-H. Chi, P.-H. Chung, T.-H. Wu, C.-C. Hsieh, S.-W. Li, and H.-y. Lee, "Audio albert: A lite bert for self-supervised learning of audio representation," *arXiv preprint arXiv:2005.08575*, 2020.
- [50] S. Ling and Y. Liu, "Decoar 2.0: Deep contextualized acoustic representations with vector quantization," *arXiv preprint arXiv:2012.06659*, 2020.
- [51] A. Baevski, M. Auli, and A. Mohamed, "Effectiveness of self-supervised pre-training for speech recognition," *arXiv preprint arXiv:1911.03912*, 2019.
- [52] Y.-H. H. Tsai, Y. Wu, R. Salakhutdinov, and L.-P. Morency, "Self-supervised learning from a multi-view perspective," *arXiv preprint arXiv:2006.05576*, 2020.
- [53] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio, "Learning problem-agnostic speech representations from multiple self-supervised tasks," in *INTERSPEECH*, 2019.
- [54] T. Likhomanenko, Q. Xu, J. Kahn, G. Synnaeve, and R. Collobert, "slimipl: Language-model-free iterative pseudo-labeling," *arXiv preprint arXiv:2010.11524*, 2020.
- [55] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.
- [56] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, 2011.
- [57] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," Stanford, Tech. Rep., 2006.
- [58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [59] V. Pratap *et al.*, "wav2letter++: The fastest open-source speech recognition system," *arXiv preprint arXiv:1812.07625*, 2018.
- [60] M. Ott *et al.*, "fairseq: A fast, extensible toolkit for sequence modeling," in *NAACL*, 2019.
- [61] D. S. Park, Y. Zhang, Y. Jia, W. Han, C.-C. Chiu, B. Li, Y. Wu, and Q. V. Le, "Improved noisy student training for automatic speech recognition," *arXiv preprint arXiv:2005.09629*, 2020.
- [62] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.
- [63] Q. Xu, A. Baevski, T. Likhomanenko, P. Tomasello, A. Conneau, R. Collobert, G. Synnaeve, and M. Auli, "Self-training and pre-training are complementary for speech recognition," *arXiv preprint arXiv:2010.11430*, 2020.
- [64] D. Le, X. Zhang, W. Zheng, C. Fügen, G. Zweig, and M. L. Seltzer, "From senones to chenones: Tied context-dependent graphemes for hybrid speech recognition," in *ASRU*, 2019.