Interspeech 2024
1-5 September 2024, Kos, Greece

Balanced-Wav2Vec: Enhancing Stability and Robustness of Representation Learning Through Sample Reweighting Techniques

Mun-Hak Lee[1], *Jae-Hong Lee*[1], *DoHee Kim*[2], *Ye-Eun Ko*[1], *Joon-Hyuk Chang*[1*]

Hanyang University, Seoul, Republic of Korea,
Department of Electronics Engineering[1], Department of Artificial Intelligence Application[2]
{lullaby0804, ljh931jh, dohe0342, yee772, jchang}@hanyang.ac.kr

Abstract

Mode collapse refers to the phenomenon where a representation model fits only a subset of modes in the feature space. Today, numerous self-supervised learning algorithms, including Wav2Vec 2.0, encounter the problem of reduced expressiveness due to mode collapse or dimension collapse. In this study, we experimentally verify that the highly skewed codebook distribution of the Wav2Vec 2.0 exacerbates the mode collapse problem. Based on this empirical finding, we propose the balanced-infoNCE loss, which suppresses the emergence of over-represented modes. We show that the Wav2Vec 2.0 model trained with balanced-infoNCE loss maintains high codebook entropy and converges stably. Furthermore, through finetuning experiments on a multilingual dataset for the ASR task, we demonstrate that balanced-Wav2Vec 2.0 models exhibit superior generalization performance.

Index Terms: self-supervised learning, Wav2Vec 2.0, mode collapse, diversity loss, speech recognition

1. Introduction

Deep neural networks (DNNs) employ many parameters to model complex functions. Consequently, DNNs trained with a small labeled dataset tend to overfit the training set. Overfitting leads to a decrease in generalization performance, making the performance of DNNs significantly dependent on the size of the labeled data. Self-supervised learning (SSL) effectively reduces this dependency by utilizing unlabeled data for pretraining. Pretraining with SSL algorithms enables achieving performance comparable to models trained on large datasets with supervised learning using only a small amount of labeled data for finetuning [1, 2]. Inspired by this success, various SSL algorithms have been explored in the speech domain. Research in speech domain SSL has included methods based on contrastive learning [3–6], methods predicting discrete hidden unit sequences from distorted speech features [7, 8], and methods predicting continuous representations [2, 9, 10]. As research on SSL algorithms has deepened, there has been active discussion about effective representation learning methods [11–13]. [12] highlights two primary conditions as crucial factors for the performance of SSL algorithms. The first is *alignment*, meaning that slight perturbations in the input space do not lead to significant dynamic changes in the output space (or feature space). Therefore, a good representation model should map positive pairs to proximate points in the feature space [12]. The second is *diversity* or *uniformity*, where uniformity refers to feature vectors being evenly distributed across the feature space, and feature diversity is used in a similar sense [13].

*Corresponding author.

Representation models that fail to achieve sufficient diversity encounter problems such as mode collapse or dimension collapse. Mode collapse is a phenomenon where a representation model fits only a subset of modes in the feature space, thereby diminishing the model's expressiveness, computational efficiency, and generalization performance [1]. Various methods have been developed to prevent mode collapse in representation models and to enhance feature diversity [5, 6]. This study aims to identify the causes behind the occurrence of mode collapse and to propose solutions to alleviate them. The contributions of this study are summarized as follows:

1. We validate that high skewness in the codebook distribution contributes to mode collapse [14].
2. Based on this empirical finding, we propose the balanced-infoNCE loss, a method designed to suppress the emergence of over-represented modes.
3. Through pretraining tasks conducted with the LibriSpeech dataset and automatic speech recognition (ASR) fine-tuning tasks performed with out-of-domain data (Korean, Spanish, and German), we demonstrate that the balanced-infoNCE loss enhances the stability and robustness of the Wav2Vec 2.0 model.

2. Background

2.1. Model Structure of Wav2Vec 2.0

In this section, we introduce the structure and training strategies of the Wav2Vec 2.0 model [6]. The Wav2Vec 2.0 model takes raw waveform $X \in \mathcal{X}$ as input and produces a length T latent speech representation $Z = [z_1, ..., z_T] \in \mathcal{Z}$, quantized representation $Q = [q_1, ..., q_T] \in \mathcal{Q}$, and contextual representation $C = [c_1, ..., c_T] \in \mathcal{C}$. The Wav2Vec 2.0 structure is divided into three main modules. The *feature encoder* $f : \mathcal{X} \rightarrow \mathcal{Z}$, consisting of a convolutional neural network, performs subsampling to reduce the length of the raw waveform and extracts local features. The *quantization module* $h : \mathcal{Z} \rightarrow \mathcal{Q}$ uses the gumbel softmax function to quantize the latent speech representation [5, 15]. Lastly, the *context network* $g : \mathcal{Z} \rightarrow \mathcal{C}$, built with transformers, generates contextual representation considering the sequential dependency of the latent speech representation.

2.2. Training Strategies of Wav2Vec 2.0

The Wav2Vec 2.0 model is trained using a mask prediction method that employs contrastive loss, updating its parameters to minimize the following objective function:

$$\mathcal{L}_{\mathrm{W2V2}} = \mathcal{L}_{\mathrm{info}} + \alpha_1 \mathcal{L}_{\mathrm{div}} + \alpha_2 \mathcal{L}_{\mathrm{L2}} \quad (1)$$

10.21437/Interspeech.2024-1875

where $\mathcal{L}_{\text{info}}$ represents the contrastive loss (infoNCE loss), $\mathcal{L}_{\text{div}}$ represents auxiliary diversity loss, $\mathcal{L}_{\text{L2}}$ represents L2 loss[1], and (α_1, α_2) are tuning factors. The objective function of Wav2Vec 2.0 is designed to maximize the mutual information of the positive pair (q_t, c_t) using the infoNCE loss and to increase codebook diversity using the diversity loss [3, 6]. In this section, we provide a detailed introduction to the objective function of Wav2Vec 2.0 and analyze how each component contributes to improving codebook diversity.

2.2.1. *infoNCE Loss*

Due to the continuous nature of speech signals and the correlation between adjacent samples, they are more susceptible to mode collapse [9, 16]. The infoNCE loss, proposed by [3], effectively improves in-utterance diversity by sampling both negative and positive samples within the same utterance:

$$\mathcal{L}_{\text{info}}(q_t, c_t) = -\log \frac{\exp(\text{sim}(q_t, c_t)/\kappa)}{\sum_{\tilde{q}\in Q_t} \exp(\text{sim}(\tilde{q}, c_t)/\kappa)} \tag{2}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and κ is a tuning factor. Q_t comprises one positive quantized vector q_t and K negative quantized vectors. The negative quantized vectors are uniformly sampled from the masked time steps within the utterance.

2.2.2. *Diversity Loss and Others*

Achieving sufficient diversity with contrastive loss alone requires a large number of negative samples, which reduces the computational efficiency of the model [17]. Diversity loss represents the negative entropy of the codebook distribution, and [6] proposed using diversity loss in conjunction with contrastive loss to enhance codebook diversity:

$$\mathcal{L}_{\text{div}} = \frac{1}{GV} \sum_{i=1}^{G} \sum_{j=1}^{V} \hat{p}_{i,j} \log \hat{p}_{i,j} \tag{3}$$

where G denotes the number of groups, and V denotes the number of entries in a codebook. Adjusting the value of α_1 can increase the codebook entropy. However, as the value of α_1 increases, the dominance of diversity loss in the overall loss landscape of Wav2Vec 2.0 can lead to reduced stability [6].

In addition to the methods mentioned, L2 loss, quantization, and multiple variable groups have been utilized to enhance the stability and diversity of the Wav2Vec 2.0 model. L2 loss projects the latent speech representation onto the unit hypersphere, thereby improving the stability of the training [12]. Wav2Vec 2.0 employs quantization modules and multiple variable groups, akin to the vq-Wav2Vec model, to prevent mode collapse [5]. In the following section, we present a new training approach that effectively enhances the codebook diversity of the Wav2Vec 2.0 model without hindering the convergence of the infoNCE loss.

3. Proposed Methods

3.1. Problem Setup

The stochastic gradient descent method is utilized to find model parameters that minimize $\mathcal{L}_{\text{W2V2}}$. During this process, the model is trained to minimize the empirical risk $\hat{R}(\mathcal{D})$, computed using a training set $\mathcal{D} = \{(X_i)\}_{i=1}^{M}$ on a mini-batch basis. Here, T_i denotes the length of the quantized vector sequence Q_i derived from X_i, and M is the size of the mini-batch. We reveal that during the calculation of empirical risk, over-represented modes are assigned higher weights, experimentally demonstrating that it causes a decrease in codebook diversity. The empirical risk computed on a mini-batch basis can be expressed as an expectation over individual codebook entries within $\mathcal{V}$:

$$\begin{aligned}\hat{R}(\mathcal{D}) &= \mathbb{E}_{\mathcal{D}}[\mathcal{L}_{\text{info}}(q_{i,t}, c_{i,t})] \\ &= \mathbb{E}_{v\in\mathcal{V}}\mathbb{E}_{(i,t)\in\mathcal{D}_v}[\mathcal{L}_{\text{info}}(q_{i,t}, c_{i,t})] \\ &= \sum_{v=1}^{V} p(v) \sum_{(i,t)\in\mathcal{D}_v} p(i,t|q_{i,t}=v)\mathcal{L}_{\text{info}}(q_{i,t}, c_{i,t})\end{aligned} \tag{4}$$

where $\mathcal{D}_v = \cup_{i=1}^{M}\{(i,t)|q_{i,t} = v, \forall t \in [T_i]\}$, $[T_i] = \{1, 2, ..., T_i\}$, and $[M] = \{1, 2, ..., M\}$. In a typical training scenario for Wav2Vec 2.0, $\hat{p}(v)$ and $\hat{p}(i,t|q_{i,t} = v)$ are estimated at the mini-batch level through a frequentist approach.

$$\hat{p}(v) = \frac{N_v}{N} \quad \text{and} \quad \hat{p}(i,t|q_{i,t}=v) = \frac{1}{N_v} \tag{5}$$

where $N_v = \sum_{i\in[M]}\sum_{t\in[T_i]} \mathbf{1}(q_{i,t} = v)$, $N = \sum_{i\in[M]} T_i$, and $\mathbf{1}(\cdot)$ denotes indicator function. By substituting $\hat{p}(v)$ and $\hat{p}(i,t|q_{i,t}=v)$ into Eq. (4), it can be expressed as follows:

$$\begin{aligned}\hat{R}(\mathcal{D}) &= \sum_{v=1}^{V} \frac{N_v}{N} \sum_{(i,t)\in\mathcal{D}_v} \frac{1}{N_v}\mathcal{L}_{\text{info}}(q_{i,t}, c_{i,t}) \\ &= \frac{1}{N} \sum_{(i,t)\in\mathcal{D}} \mathcal{L}_{\text{info}}(q_{i,t}, c_{i,t})\end{aligned} \tag{6}$$

The original Wav2Vec updates its model parameters at each iteration to minimize the estimated empirical risk, as shown in Eq. (6). However, this estimation method has the following limitations:

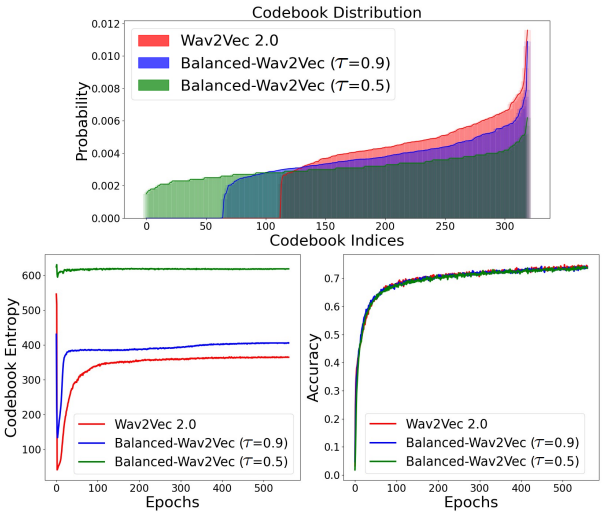

Figure 1: *Top: Codebook distribution estimated using LibriSpeech validation set. Bottom left: Codebook entropy estimated using LibriSpeech validation set for each epoch. Bottom right: Validation set accuracy for each epoch.*

[1]L2 loss was not specified in the original Wav2Vec 2.0 paper but it can be utilized to improve model stability [6, 12].

1. The presence of over-represented modes results in $\hat{p}(v)$ being highly skewed (see Figure 1, top).
2. $\hat{p}(v)$ is estimated using a limited size of mini-batches, where the correlation between in-utterance samples is high, resulting in a large variance of the $\hat{p}(v)$.

The high variance of $\hat{p}(v)$ undermines the stability of model training, and the high skewness shifts more probability mass towards over-represented modes in the codebook distribution. We hypothesize that the high skewness of $\hat{p}(v)$ intensifies mode collapse through iterative parameter updates. Based on this assumption, we propose smoothing $\hat{p}(v)$ to prevent mode collapse and improve the stability of the representation models.

3.2. Balanced-infoNCE Loss

We utilize a smoothing factor $0 \leq \tau \leq 1$ to smooth the codebook distribution $\hat{p}(v)$ estimated on a mini-batch basis, aiming to suppress the emergence of over-represented modes.

$$\hat{p}(v;\tau) = (\frac{N_v}{N})^{\tau} \tag{7}$$

By substituting $\hat{p}(v;\tau)$ and $\hat{p}(i,t|q_{i,t}=v) = \frac{1}{N_v}$ into Eq. (4), we can recalculate the empirical risk as follows:

$$\begin{aligned}\hat{R}(\mathcal{D}) =& \sum_{v=1}^{V}(\frac{N_v}{N})^{\tau} \sum_{(i,t)\in\mathcal{D}_v} \frac{1}{N_v}\mathcal{L}_{\text{info}}(q_{i,t},c_{i,t}) \\ =& \frac{1}{N}\sum_{(i,t)\in\mathcal{D}}(\frac{N_v}{N})^{\tau-1}\mathcal{L}_{\text{info}}(q_{i,t},c_{i,t}) \\ =& \frac{1}{N}\sum_{(i,t)\in\mathcal{D}}\mathcal{L}_{\text{B}}\end{aligned} \tag{8}$$

The balanced-infoNCE loss $\mathcal{L}_{\text{B}}$ is equivalent to the infoNCE loss $\mathcal{L}_{\text{info}}$ multiplied by the sample weight $(\frac{N_v}{N})^{\tau-1}$ [18, 19]. When $\tau = 1$, the balanced-infoNCE loss is identical to the original infoNCE loss. Conversely, when $\tau = 0$, the sample weight decreases inversely with the frequency of samples sharing classes within the mini-batch, thereby assigning a smaller weight to over-represented modes. For simplicity, we set the number of groups to one in this section; when there are more than two groups, the average sample weight of the groups is used.

3.3. Interpretation

The balanced-infoNCE loss is methodologically similar to sample reweighting approaches and focal loss in that it assigns higher weights to sparsely occurring modes [18, 20]. Additionally, balanced-infoNCE can be interpreted as a Bayesian approach that assumes a noninformative prior (uniform distribution) for the codebook distribution. Given our lack of prior knowledge about the codebook distribution, setting it to a uniform distribution according to the *principle of insufficient reason*[2] is a reasonable choice [22]. The diversity loss in Eq. (3) and the maximum entropy regularization proposed in [23] share a similar motivation to ours.

[2]Principle of insufficient reason: When specific domain information is absent, it is preferable to use an uninformative prior or noninformative prior based on the least amount of assumptions. A representative example of an uninformative prior is the maximum entropy prior [21, 22].

4. Experiments

We performed finetuning for the ASR task and used word error rate (WER) as the evaluation metric. In experiments using the LibriSpeech dataset, we employed a method of fusing an external 4-gram-based language model (LM). We used beam-search decoding with a beam size set to 1,500. For experiments using out-of-domain datasets, no external LM was used, and decoding was performed using the Viterbi search algorithm.

4.1. Data

We used the LibriSpeech dataset, a public benchmark for English speech recognition [24]. The LibriSpeech dataset consists of 960 hours of training data and four evaluation sets (dev-clean, dev-other, test-clean, and test-other). We conducted pretraining using the 960-hour training set, selected models using the validation sets and performed evaluations on the test set. All LMs used in the experiments with the LibriSpeech dataset were trained using the LibriCorpus dataset [24]. We conducted finetuning experiments with Korean, German, and Spanish to assess performance in out-of-domain conditions. The Korean data in our experiments consists of commands recorded in a vehicle environment, with about 90 hours of training data, and each validation and test set comprising about 5 hours. We segmented Korean words into graphemes for training the CTC model [25]. The Spanish and German datasets were from the public benchmark CommonVoice 5.1 [26]. CommonVoice is composed of sentences from Wikipedia, and we used voice files recorded at 48kHz, re-sampled to 16kHz for our experiments. The Spanish dataset in our experiments comprises about 200 hours of training data, and each validation and test set is about 25 hours. The German dataset consists of about 314 hours of training data, and each validation and test set is about 25 hours. We used letters as recognition units for the CTC model for German, Spanish, and English experiments.

4.2. Models

In all experiments, we utilized the open-source toolkit FairSeq [27]. The Wav2Vec 2.0 base model served as our baseline. During the pretraining phase, we used four A100 GPUs, setting the maximum number of tokens in a mini-batch to 11.2M and the update frequency to two. Other hyperparameters, including the number of groups set to two, codebook dimension at 320, $\alpha_1 = 0.1$, and $\alpha_2 = 0$, were consistent with those reported in [6]. Pretraining was conducted for 593 epochs over the entire LibriSpeech 960h dataset, and for all experiments, the model from the last epoch was used for finetuning. In the finetuning stage, we saved the finetuned model at the end of each epoch and used the model that showed the lowest WER on the validation set for evaluation.

5. Related Works

We introduce studies that have attempted to improve the diversity and stability of representation models in the speech domain. [28] proposed an auxiliary consistency loss that uses a consistency network to reconstruct input features. This auxiliary consistency loss plays a role in improving codebook diversity and preventing the loss of information, thereby enhancing the robustness of representation models. [10] suggested a position randomization method to prevent positional collapse issues and proposed an in-utterance contrastive loss using the output from a teacher network obtained through a moving av-

Table 1: *Evaluation results (WER) of models finetuned with the in-domain dataset (LibriSpeech).*

		test_clean		test_other		dev_clean		dev_other	
train set	Model	viterbi	4-gram	viterbi	4-gram	viterbi	4-gram	viterbi	4-gram
100h	Wav2Vec2.0	5.78	3.22	13.45	8.22	5.63	3.04	13.67	8.07
	Balanced-Wav2Vec ($\tau = 0.9$)	**5.65**	3.24	**12.81**	**7.94**	**5.58**	**2.99**	**13.22**	**7.85**
	Balanced-Wav2Vec ($\tau = 0.5$)	5.68	**3.21**	13.01	8.11	5.61	3.03	13.48	7.90
960h	Wav2Vec2.0	3.32	2.51	8.61	6.16	3.29	2.22	8.85	5.87
	Balanced-Wav2Vec ($\tau = 0.9$)	**3.21**	**2.50**	**8.33**	**5.86**	3.29	2.19	**8.57**	**5.72**
	Balanced-Wav2Vec ($\tau = 0.5$)	3.37	2.53	9.09	6.22	**3.17**	**2.15**	8.73	5.79

Table 2: *Evaluation results (WER) of models finetuned with out-of-domain datasets (Korean, Spanish, and German).*

	Korean		Spanish		German	
	dev	test	dev	test	dev	test
Wav2Vec2.0	22.47	23.07	13.89	15.11	16.26	18.30
Balanced-Wav2Vec ($\tau = 0.9$)	**20.93**	**21.89**	**13.57**	**14.81**	**15.66**	**17.71**
Balanced-Wav2Vec ($\tau = 0.5$)	22.11	22.37	13.81	15.07	15.94	18.00

erage method. [16] utilized a regularization loss that predicts pseudo targets generated by a frozen teacher model to improve the stability and diversity of the representation model. Although the proposed method shares the motivation of improving the diversity and stability of representation models with these papers, it differs in that it does not depend on external modules such as teacher networks or consistency networks and does not use data augmentation or auxiliary regularization losses.

Next, we introduce studies that are methodologically similar to balanced-infoNCE loss. Classification models trained with imbalanced data exhibit lower performance in minor classes. To mitigate this issue, over-sampling methods and sample reweighting methods have been utilized [19]. Sample reweighting methods assign higher weights to samples of minor classes during the loss calculation process [18, 19]. We introduce three papers that combine sample weighting methods with contrastive learning. [29] proposed supervised contrastive learning using labeled data and utilized sample reweighting to eliminate the dependency on the number of samples sharing the same class within a mini-batch. [30,31] suggested assigning different weights to positive and negative samples. [31] demonstrated that giving higher weight to hard-negative samples over easy-negative samples improves the diversity and robustness of sentence embedding models.

6. Results and Discussion

Stability: We trained the model with the hyperparameter τ set to 1.0, 0.9, and 0.5 and presented the results in Figure 1. When $\tau = 1.0$, balanced-infoNCE is identical to the original infoNCE; hence, we refer to it as Wav2Vec 2.0. The bottom left graph in Figure 1 shows the codebook entropy measured for the validation set at each epoch, while the bottom right graph shows the accuracy measured for the validation set at each epoch. Accuracy is defined as follows:

$$\mathrm{ACC}(\mathcal{D}) = \frac{1}{N} \sum_{i=1}^{M} \sum_{t=1}^{T_i} \mathbf{1}(q_{i,t} = \hat{q}_{i,t}), \tag{9}$$

where $\hat{q}_{i,t} = \underset{v \in \mathcal{V}}{\mathrm{argmax}}\ \mathrm{sim}(v, c_{i,t})$. The graphs in Figure 1 demonstrate that models with $\tau < 1.0$ converge more quickly and maintain higher codebook diversity compared to the original Wav2Vec 2.0 model while still preserving a similar level of accuracy. The top graph in Figure 1 depicts the codebook distribution of the trained models, estimated using the validation set. This graph visually illustrates that reducing τ can enhance codebook diversity and mitigate mode collapse. Such experimental results support our hypothesis that the high skewness of $\hat{p}(v)$ could contribute to mode collapse.

Robustness: Recent research in representation learning has discovered that improving a model's diversity leads to enhanced robustness [12, 13, 23, 32]. To verify the robustness of the proposed method, we conducted finetuning on various out-of-domain datasets. We conducted experiments where models pretrained on the LibriSpeech dataset (English) were fine-tuned and evaluated on Korean, German, and Spanish datasets, and the results are shown in Table 2. The experimental outcomes showed relative performance improvements of 0-7% across all datasets, demonstrating that the balanced-infoNCE loss improves the robustness of the Wav2Vec 2.0 model.

Discussion: Quantization allows speech signals to be processed in the same manner as text [5]. We believe that balanced-infoNCE loss will be effective in improving the diversity and stability of multi-modal representation models [33, 34].

7. Conclusions

We proposed a new training strategy to mitigate the mode collapse problem of the Wav2Vec 2.0 model. We assumed that the high variance and skewness of $\hat{p}(v)$ estimated on a mini-batch basis exacerbate mode collapse and undermine training stability. Based on this assumption, we introduced the balanced-infoNCE loss, which smooths $\hat{p}(v)$ at the mini-batch level. We experimentally demonstrated that the proposed method improves codebook diversity and, through finetuning with multilingual datasets, showed that it improves the generalization performance.

8. Acknowledgements

This work was partly supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No.RS-2023-00302424) and the National Supercomputing Center with supercomputing resources including technical support (No.TS-2024-RE-0034).

9. References

[1] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. of the International Conference on Machine Learning* (*ICML*). PMLR, 2020, pp. 1597–1607.

[2] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "BYOL for audio: Self-supervised learning for general-purpose audio representation," in *Proc. of the International Joint Conference on Neural Networks* (*IJCNN*). IEEE, 2021, pp. 1–8.

[3] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arxiv preprint arXiv:1807.03748*, 2018.

[4] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: Unsupervised pre-Training for speech recognition," in *Proc. INTERSPEECH*, 2019.

[5] A. Baevski, S. Schneider, and M. Auli, "VQ-Wav2vec: Self-supervised learning of discrete speech representations," in *Proc. of the International Conference on Learning Representations* (*ICLR*), 2019.

[6] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. of the Advances in Neural Information Processing Systems* (*NeurIPS*), vol. 33, 2020, pp. 12 449–12 460.

[7] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HUBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[8] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[9] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *Proc. of the International Conference on Machine Learning* (*ICML*). PMLR, 2022, pp. 1298–1312.

[10] W. Huang, Z. Zhang, Y. T. Yeung, X. Jiang, and Q. Liu, "Spiral: Self-supervised perturbation-invariant representation learning for speech pre-training," in *Proc. of the International Conference on Learning Representations* (*ICLR*), 2021.

[11] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[12] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *Proc. of the International Conference on Machine Learning* (*ICML*). PMLR, 2020, pp. 9929–9939.

[13] A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," in *Proc. of the International Conference on Learning Representations* (*ICLR*), 2022.

[14] A. W. W. Eide, E. Solberg, and I. Kåsen, "Sample weighting as an explanation for mode collapse in generative adversarial networks," *arxiv preprint arXiv:2010.02035*, 2020.

[15] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *Proc. of the International Conference on Learning Representations* (*ICLR*), 2016.

[16] L. Cao, J. Wang, B. Yang, D. Su, and D. Yu, "TriNet: Stabilizing self-supervised learning from complete or slow collapse," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2023, pp. 1–5.

[17] L. Jing, P. Vincent, Y. LeCun, and Y. Tian, "Understanding dimensional collapse in contrastive self-supervised learning," in *Proc. of the International Conference on Learning Representations* (*ICLR*), 2022.

[18] T. Fang, N. Lu, G. Niu, and M. Sugiyama, "Rethinking importance weighting for deep learning under distribution shift," in *Proc. of the Advances in Neural Information Processing Systems* (*NeurIPS*), vol. 33, 2020, pp. 11 996–12 007.

[19] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *Journal of Statistical Planning and Inference*, vol. 90, no. 2, pp. 227–244, 2000.

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. of the IEEE International Conference on Computer Vision* (*ICCV*), 2017, pp. 2980–2988.

[21] E. T. Jaynes, "Information theory and statistical mechanics," *Physical Review*, vol. 106, no. 4, p. 620, 1957.

[22] K. P. Murphy, *Probabilistic machine learning: advanced topics*. MIT press, 2023.

[23] X. Liu, Z. Wang, Y.-L. Li, and S. Wang, "Self-supervised learning via maximum entropy coding," in *Proc. of the Advances in Neural Information Processing Systems* (*NeurIPS*), vol. 35, 2022, pp. 34 091–34 105.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2015, pp. 5206–5210.

[25] M.-h. Lee and J. H. Chang, "Korean speech recognition based on grapheme," *Journal of the Acoustical Society of Korea*, vol. 38, no. 5, pp. 601–606, 2019.

[26] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proc. of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

[27] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "Fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of the 2019 Conference of the North*. Proc. of the Association for Computational Linguistics (ACL), 2019.

[28] S. Sadhu, D. He, C.-W. Huang, S. H. Mallidi, M. Wu, A. Rastrow, A. Stolcke, J. Droppo, and R. Maas, "Wav2vec-c: A self-supervised model for speech representation learning," in *Proc. INTERSPEECH*, 2021.

[29] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Proc. of the Advances in Neural Information Processing Systems* (*NeurIPS*), vol. 33, 2020, pp. 18 661–18 673.

[30] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. of the International Conference on Machine Learning* (*ICML*). PMLR, 2020, pp. 1597–1607.

[31] P. Hou and X. Li, "Improving contrastive learning of sentence embeddings with focal infoNCE," in *Findings of the Association for Computational Linguistics: EMNLP*, 2023, pp. 4757–4762.

[32] J.-S. Choi, J.-H. Lee, C.-W. Lee, and J.-H. Chang, "M-CTRL: A continual representation learning framework with slowly improving past pre-Trained model," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2023, pp. 1–5.

[33] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. d. C. Quitry, P. Chen, D. E. Badawy, W. Han, E. Kharitonov *et al.*, "Audiopalm: A large language model that can speak and listen," *arxiv preprint arXiv:2306.12925*, 2023.

[34] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang *et al.*, "SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing." Proc. of the Association for Computational Linguistics (ACL), 2022, pp. 5723–5738.