ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447758

REVISITING SELF-SUPERVISED LEARNING OF SPEECH REPRESENTATION FROM A MUTUAL INFORMATION PERSPECTIVE

Alexander H. Liu[*1] *Sung-Lin Yeh*[*2] *James R. Glass*[1]

[1]MIT CSAIL, [2]University of Edinburgh, Informatics

ABSTRACT

Existing studies on self-supervised speech representation learning have focused on developing new training methods and applying pre-trained models for different applications. However, the quality of these models is often measured by the performance of different downstream tasks. How well the representations access the information of interest is less studied. In this work, we take a closer look into existing self-supervised methods of speech from an information-theoretic perspective. We aim to develop metrics using mutual information to help practical problems such as model design and selection. We use linear probes to estimate the mutual information between the target information and learned representations, showing another insight into the accessibility to the target information from speech representations. Further, we explore the potential of evaluating representations in a self-supervised fashion, where we estimate the mutual information between different parts of the data without using any labels. Finally, we show that both supervised and unsupervised measures echo the performance of the models on layer-wise linear probing and speech recognition.

Index Terms— Self-supervised speech representation learning, representation analysis, information theory

1. INTRODUCTION

Estimating the amount of information encoded in learned representations has been an important research topic in speech representation learning. A good estimation can not only offer a better view of designing training objectives, especially under a self-supervised paradigm but also help select models for the applications of interest. To measure to what extent the representations reveal specific information, several studies have adopted a phonetic-related linear probing protocol [1, 8]. Also, a collection of downstream tasks has been proposed to evaluate learned representations, including phone classification and speech recognition [9]. The accuracy obtained from a task is then believed to reflect the accessibility of representations to certain information.

Although the aforementioned approaches for measuring representations have been widely used, there are certain limitations. For example, a probing task is not formally measuring the "information" inherent in the representations but their accuracy on a task. Another limitation is that the probing tasks all rely on labeled data. Further, contextual speech representations learned from self-supervised models are actually trained to predict the context such as the future or masked frames [4, 5, 1, 2, 8]. The current probing approaches, however, ask the classifier to do same-frame prediction. The mismatch between the training objectives and the evaluations makes it unclear whether the current measurements have properly reflected the information representations encoded.

In this paper, we present an information-theoretic approach to assess the information contained in representations [10, 11]. We use mutual information (MI) to measure the relationship between representations and their targets, such as phonetic labels. To evaluate how well representations capture the context, we propose that effective representations should exhibit higher MI between different parts of the input due to self-supervised training. To test this, we divide the input into two parts and estimate the MI between representations derived from different parts. This offers an unsupervised alternative to measure the learned representations.

Through extensive experiments, our findings reveal a strong correlation between unsupervised measures and supervised ones in phonetic-related probing. This correlation suggests the potential for probing representations without labeled data. Furthermore, we observe that models exhibiting higher MI in an unsupervised measure also exhibit superior performance in downstream speech recognition.

2. MEASURING SELF-SUPERVISED MODELS WITH MUTUAL INFORMATION

2.1. A Mutual Information Perspective of Self-supervised Methods on Speech

Prior works have drawn the connections between self-supervised training objectives and maximizing mutual information (MI) [12, 13, 8, 14]. Following a similar vein, we consider self-supervised approaches as maximizing the MI between different parts of the input by dividing the input into two views, defined as X_a and X_b. Specifically, we focus on the MI between X_a and target variables Z_{target} derived from X_b.

[*] Equal contribution.

Table 1: Summary of prior self-supervised speech representation methods in two different categories: Autoregressive Predictive Coding (APC) and Masked Langauge Modeling (MLM). VQ stands for vector quantization; EMA stands for exponential moving average.

	Views X_a/X_b	Choice of Z_{target}	Choice of p_ψ	$\mathrm{H}(Z_{\text{target}})$	connection to $\mathrm{H}(Z_{\text{target}}\vert f_\theta(X_a))$
APC family					
APC [1] / VQ-APC [2]	past / future	spectrogram	identity matrix	intractable, fixed	regression
Co-training [3]		discrete	VQ	estimable[‡]	cross-entropy
MLM family					
wav2vec 2.0 [4]	masked[†] / unmasked	discrete	VQ	estimable[‡]	contrastive loss
HuBERT [5] / WavLM [6]		discrete	k-means assignment	estimable[‡], fixed	cross-entropy
DinoSR [7]		discrete	EMA(f_θ) + VQ	estimable[‡]	cross-entropy

[†] WavLM additionally introduced noise to generate X_a. [‡] We consider $\mathrm{H}(Z_{\text{target}})$ estimable by sampling Z_{target} from $p_\psi(Z_{\text{target}}|X_b)$ to compute the empirical entropy.

Formally,

$$\mathrm{I}(X_a; Z_{\text{target}}) = \mathrm{H}(Z_{\text{target}}) - \mathrm{H}(Z_{\text{target}}|X_a), \tag{1}$$

where a network f_θ is employed to model $p(Z_{\text{target}}|X_a)$. Besides the different views of data, the choice of target variable Z_{target} is perhaps the most significant difference between methods. Some methods [1, 5, 6] have propose to use predefined transformation $p_\psi(X_b) = Z_{\text{target}}$ to derive target variable; other methods [4, 7] have proposed to learn it jointly during training by introducing $p_\psi(Z_{\text{target}}|X_b)$. Table 1 provides some concrete examples. Note that we only list open-sourced models in the table, there are more prior works [15, 16, 17] in the field that are not covered.

Theoretically, measuring MI through Eq. 1 allows us to compare different SSL methods in speech, but in practice, it is infeasible due to some limitations. For example, methods like APC [1] and VQ-APC [2] in Table 1, Eq. 1 is intractable due to the unknown distributions in the equation [18]. Even in the case where Eq. 1 can be approximated (e.g., HuBERT [5]), it is hard to fairly compare it to other methods due to the different definitions of p_ψ and Z_{target}. This motivates us to bound MI in a tractable way that is invariant to the choice of Z_{target} and p_ψ such that we can compare these self-supervised methods.

2.2. Bounding MI with labeled data

After the SSL stage, the representation Z of data X can be extracted with the pre-trained model f_θ, i.e., $Z = f_\theta(X)$. An intuitive way to compare the quality of speech representations Z from different models is by examining the mutual information between Z and specific *target* Y using labeled data (e.g., the underlying phone at the corresponding time), namely, $\mathrm{I}(Z;Y) = \mathrm{H}(Y) - \mathrm{H}(Y|Z)$. While the entropy of target $\mathrm{H}(Y)$ is a constant (depending solely on the choice of target) that can be estimated with an empirical distribution, the metric itself is still intractable since we do not know the relation between Z and Y. Nevertheless, the target mutual information $\mathrm{I}(Z;Y)$ can be lower-bounded through upper-bounding the conditional entropy $\mathrm{H}(Y|Z)$ with an auxiliary prediction model $q_\phi(y|z)$. More precisely,

$$\begin{aligned}
\mathrm{I}(Z;Y) &= \mathrm{H}(Y) - \mathbb{E}_{(y,z)\sim p(Y,Z)}\Big[-\log p(y|z)\Big] && (2)\\
&= \mathrm{H}(Y) - \mathbb{E}_p\Big[-\log q_\phi(y|z) - \log \frac{p(y|z)}{q_\phi(y|z)}\Big] && (3)\\
&\geq \mathrm{H}(Y) - \mathbb{E}_p\Big[-\log q_\phi(y|z)\Big], && (4)
\end{aligned}$$

where $\mathbb{E}_p\big[\log \frac{p(y|z)}{q_\phi(y|z)}\big] = \mathbb{E}_Z D_{\mathrm{KL}}(p(y|z)||q_\phi) > 0$ leads to Eq. 4. In other words, we can estimate the lower bound of the desired mutual information by training the auxiliary prediction model $q_\phi(y|z)$ to approximate $p(y|z)$. Note that this corresponds to *probing tasks* in the literature [1, 9] since the last term in Eq. 4 is cross-entropy loss. For example, in linear probing tasks, $q_\phi(y|z)$ is modeled by a linear layer.

By using labeled data and Eq. 4, we are able to establish a lower bound of mutual information $\mathrm{I}(Z;Y)$ that can be used as an intuitive metric to measure the quality of representation. However, there are several downsides making this metric less ideal for selecting SSL models such as the need for labeled data and the narrow viewpoint from the choice of target. These properties somewhat contradict the spirit of the SSL paradigm for learning a general model that can be applied to different tasks with minimum supervision.

2.3. Bounding MI with unlabeled data

In light of how self-supervised methods are designed, we propose to measure the mutual information between different views of the input instead of using labeled data. For simplicity, here we use $Z_a = f_\theta(X_a)$ to denote the representation extracted from one view and $Z_b = f_\theta(X_b)$ from the other.

Table 2: Results of MLM methods on LibriSpeech test-other. For downstream tasks, frozen feature results are taken from Speech processing Universal PERformance Benchmark [9] and the 10-hour fine-tuning results are the decoding result with language model reported by each prior work.

	mutual information lower-bound				downstream tasks		
	$\mathrm{I}(Z;Y)$ (Eq. 4)		$\mathrm{I}(Z_a;Z_b)$ (Eq. 8)		frozen feature		fine-tune
	logistic	MLP	logistic	MLP	PER	WER	WER
BASE models (12-layer)							
wav2vec 2	3.50	3.52	1.13	1.08	5.74	6.43	9.5
HuBERT	3.73	3.75	2.04	2.04	5.41	6.42	9.4
WavLM	3.80	3.82	2.05	2.05	4.84	6.21	9.2
DinoSR	3.83	3.89	2.10	2.15	3.21	4.71	7.6
LARGE models (24-layer)							
wav2vec 2	3.81	3.87	1.67	1.63	4.75	3.75	4.9
HuBERT	3.85	3.90	2.08	2.05	3.53	3.62	4.6
WavLM	3.87	3.94	2.40	2.36	3.06	3.44	4.6

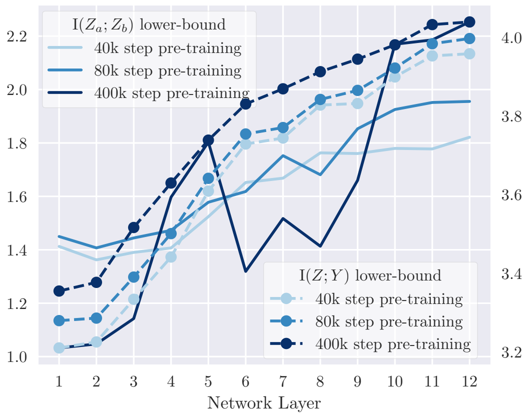

Table 3: Lower-bound dynamic w.r.t. different layers and pre-training steps of DinoSR. Y-axis on the left corresponded to $\mathrm{I}(Z_a;Z_b)$ lower-bound (bit); right corresponded to $\mathrm{I}(Z;Y)$ lower-bound (bit).

Formally, we consider

$$\mathrm{I}(Z_a;Z_b) = \mathrm{H}(Z_a) - \mathrm{H}(Z_a|Z_b). \tag{5}$$

Similar to Eq. 2, the mutual information is intractable since the underlying distributions of the variables are unknown. To overcome the issue, we introduce a clustering function f_{cluster} to quantize the representation Z_b. We can then approximate $\mathrm{H}(Z_b)$ with the empirical distribution and estimate a lower bound of the mutual information with

$$\begin{aligned}
\mathrm{I}(Z_a;Z_b) &\geq \mathrm{I}(Z_a;f_{\text{cluster}}(Z_b)) && (6)\\
&= \mathrm{H}(f_{\text{cluster}}(Z_b)) - \mathrm{H}(f_{\text{cluster}}(Z_b)|Z_a) && (7)\\
&\geq \mathrm{H}(f_{\text{cluster}}(Z_b)) - \mathbb{E}_p\Big[-\log q_\phi(f_{\text{cluster}}(Z_b)|Z_a)\Big], && (8)
\end{aligned}$$

where Eq. 6 follows from the data-processing inequality and Eq. 8 can be derived in a similar way to Eq. 4. The key advantage of this approach is the lower bound can be estimated regardless of the choice of Z_{target}, making cross-method comparisons and checkpoint selections possible without labeled data as we show later in our experiments.

3. EXPERIMENTS

Setup. We use k-means as the clustering function f_{cluster} to quantize the representation space with default 50 clusters. For q_ϕ, we test both logistic regression and multi-layer perceptron (MLP) with 3 layers, ReLU activation, and dropout. Since our goal is to evaluate pre-trained self-supervised models, both f_{cluster} and q_ϕ are trained on the clean dev set of LibriSpeech [19] and used to estimate MI on test-clean/test-other for APC/MLM models respectively. By default, we use k-means with 50 clusters for a max of 100 iterations, then optimize Eq 8 with a learning rate of 0.1 for 10 epochs. For $\mathrm{I}(Z;Y)$, we use force-aligned [20] phone sequences as the target Y. For MLM models, we follow the same setup but adopt the noisy subsets.

For APC models, views Z_a and Z_b in Eq. 1 are generated by applying a time shift identical to the pre-training stage (60ms; 3 frames) to the representations. For MLM models, the masked view is generated by masking the last 30 frames of every 40-frame (i.e., 75% masking ratio), a more detailed discussion is provided later with results on different masking ratios. Representations are extracted from the last layer unless otherwise specified. All numbers reported are averaged over 5 runs with different random seeds, we find the variance across different runs negligible (<4e-4) for all cases.

Bounding the MLM family. We begin with results in Table 2 on the MLM family that take masked/unmasked speech as the different views. We compare both supervised metric and unsupervised metric against the downstream performance of each model. For downstream performance, we consider the speech recognition performance from SUPERB [9]. Unsurprisingly, there is a strong connection between the supervised metric and downstream performance. Models with higher $\mathrm{I}(Z;Y)$ provide representations with higher accessibility of phonetic information and benefit phonetic-related tasks. On the other hand, a similar pattern can be observed with the unsupervised metric (despite not using any labeled data) where the increasing lower bound of $\mathrm{I}(Z_a;Z_b)$ reflects stronger downstream performance. This key result suggests that it is possible to evaluate self-supervised speech representations in a self-supervised manner.

Impact of bounding conditional entropy with q_ϕ. Another observation from Table 2 worth mentioning is the choice of q_ϕ actually have small impact to our estimated lower bounds. The results are consistent between the two choices with

Table 4: Results of APC family on LibriSpeech test-clean. MLP is used as q_θ for bounding conditional entropy. LB stands for low-bound. The layer-wise PER are evaluated following the protocol of [1, 8, 3]. All models are trained and evaluated with the same time-shift of 60ms.

	$\mathrm{I}(Z;Y)$ LB Layer 1/2/3	$\mathrm{I}(Z_a;Z_b)$ LB Layer 1/2/3	PER Layer 1/2/3
APC	3.0 / 3.6 / 3.3	2.3 / 2.8 / 2.3	25.3 / 23.8 / 31.4
VQ-APC	3.4 / 3.6 / 3.1	2.8 / 3.0 / 2.9	25.4 / 22.7 / 28.4
Co-training	3.3 / 3.6 / 3.2	2.7 / 3.0 / 2.5	27.1 / 21.0 / 27.1

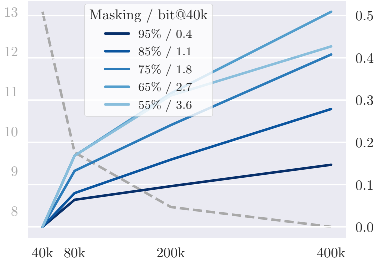

Table 5: Varying masking ration for bounding $\mathrm{I}(Z_a;Z_b)$ on DinoSR. X-axis: pre-training step. Y-axis: the improvement of lower-bound compared to 40k steps (bit; right) and the fine-tuning performance (gray dashed line; WER; left).

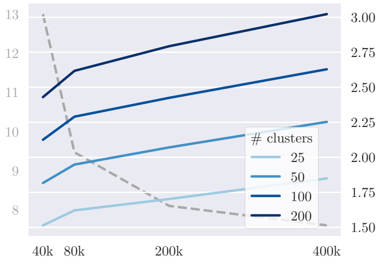

Table 6: Varying f_{cluster} cluster size for bounding $\mathrm{I}(Z_a;Z_b)$ on DinoSR. X-axis: pre-training steps. Y-axis: $\mathrm{I}(Z_a;Z_b)$ lower-bound (bit; right) and the fine-tuning performance (gray dashed line; WER; left).

slightly better estimation obtained via MLP in most cases. While only two different options are tested in this paper, we note that exploring better options of q_ϕ is worthwhile in practice since they provide tighter lower bound [10].

Bounding the APC family. In addition to MLM, we experiment with models trained on future prediction pre-training, in which the past and future are selected as different views. As shown in Table 4, the trends of supervised and unsupervised metrics do not completely align with the results of phoneme classification. Nonetheless, the layer with the highest MI consistently corresponds to the best PER of the model. This indicates the proposed metrics can be potentially applied to layer selection for phonetic-related tasks. Note that even though the number of clusters is fixed, unsupervised metric is not comparable to that of the MLM methods due to different model configurations.

Layer-wise analysis. We discover the middle layer of the APC family, which is farthest from the surface feature, achieves the highest bound against both labeled data and future view. More importantly, a consistent trend can be found across all three columns within each model, suggesting that the unsupervised metric can be used to select the target layer for feature extraction. A similar study on MLM models is carried out in Fig 3. Conversely, we find the trend of the lower bound of $\mathrm{I}(Z;Y)$ and $\mathrm{I}(Z_a;Z_b)$ to match at the early and last layers of the model. This shows that different self-supervised learning methods result in different representation patterns as suggested by existing study [21, 22]. Nevertheless, we note that using the last layer with $\mathrm{I}(Z_a;Z_b)$ lower bound is a robust option for comparing different models at the same size as it consistently matches the downstream performance trend as shown in Table 2 and Table 4.

Robustness for checkpoint selection. In Figure 5 and Figure 6, we showcase the robustness of $\mathrm{I}(Z_a;Z_b)$ lower bound by using it to evaluate DinoSR at different pre-training steps given that checkpoint selection is sometimes a hard problem for self-supervised methods in practice. To simulate the training scenario, this part is conducted on the validation set only by splitting it into half for fitting f_{cluster} and MLP q_ϕ, using the remaining part to estimate MI.

For MLM-based methods, the masking ratio is an important hyper-parameter that controls the view X_a used for training. Prior works have applied different masking ratios varying between 65% to 80%[1]. While lower masked ratios result in higher absolute MI estimation, we find $\mathrm{I}(Z_a;Z_b)$ lower bound robust to the choice of masking ratio, providing a consistent pattern as the training continues.

Finally, as shown in Figure 6, we observe similar trends of $\mathrm{I}(Z_a;Z_b)$ lower bound despite varying the number of clusters. These findings point out a new path for evaluating self-supervised speech models during pre-training as computing $\mathrm{I}(Z_a;Z_b)$ lower bound requires little computations [2].

4. CONCLUSION

In this paper, we revisited self-supervised learning of speech representation from a mutual information point of view. We provided two different MI metrics and showed their lower bounds can be used to evaluate self-supervised models without heavy computations, especially for the $\mathrm{I}(Z_a;Z_b)$ lower bound that is designed in a self-supervised manner. We checked the robustness of these metrics to demonstrate the potential of applying them to different pre-trained models in practice. However, we also note that this work focused on examining the content of speech as we only considered recognition tasks in our experiment. An interesting future direction is to explore the non-content information encoded in speech representations.

[1] In practice, these methods randomly sampled 6.5% to 8% of input frames to apply mask spanning 10 consecutive frames. This results in a lower masking ratio than expected since spans might overlap each other. For our evaluation, the expected masking ratio is precise since masks are not overlapping.

[2] Our method required 2 forward passes on the validation set. As a reference, each estimation with MLP takes less than 5 minutes on CPU for fitting f_{cluster} and q_ϕ (can be further sped up by leveraging GPU), which is considerably short compared to the runtime of self-supervised methods itself.

5. REFERENCES

[1] Yu-An Chung and James Glass, "Generative pre-training for speech with autoregressive predictive coding," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3497–3501.

[2] Yu-An Chung, Hao Tang, and James Glass, "Vector-quantized autoregressive predictive coding," *arXiv preprint arXiv:2005.08392*, 2020.

[3] Sung-Lin Yeh and Hao Tang, "Autoregressive co-training for learning discrete speech representations," *arXiv preprint arXiv:2203.15840*, 2022.

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[5] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman 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.

[6] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong 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.

[7] Alexander H Liu, Heng-Jui Chang, Michael Auli, Wei-Ning Hsu, and James R Glass, "Dinosr: Self-distillation and online clustering for self-supervised speech representation learning," *arXiv preprint arXiv:2305.10005*, 2023.

[8] Gene-Ping Yang, Sung-Lin Yeh, Yu-An Chung, James Glass, and Hao Tang, "Autoregressive predictive coding: A comprehensive study," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1380–1390, 2022.

[9] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al., "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[10] Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell, "Information-theoretic probing for linguistic structure," *arXiv preprint arXiv:2004.03061*, 2020.

[11] Elena Voita and Ivan Titov, "Information-theoretic probing with minimum description length," *arXiv preprint arXiv:2003.12298*, 2020.

[12] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[13] Lingpeng Kong, Cyprien de Masson d'Autume, Wang Ling, Lei Yu, Zihang Dai, and Dani Yogatama, "A mutual information maximization perspective of language representation learning," *arXiv preprint arXiv:1910.08350*, 2019.

[14] Iro Laina, Yuki M Asano, and Andrea Vedaldi, "Measuring the interpretability of unsupervised representations via quantized reverse probing," *arXiv preprint arXiv:2209.03268*, 2022.

[15] Shaoshi Ling and Yuzong Liu, "Decoar 2.0: Deep contextualized acoustic representations with vector quantization," *arXiv preprint arXiv:2012.06659*, 2020.

[16] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu, "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 244–250.

[17] Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu, "Self-supervised learning with random-projection quantizer for speech recognition," in *International Conference on Machine Learning*. PMLR, 2022, pp. 3915–3924.

[18] David McAllester, "Information theoretic co-training," *arXiv preprint arXiv:1802.07572*, 2018.

[19] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[20] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi.," in *Interspeech*, 2017, vol. 2017, pp. 498–502.

[21] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 914–921.

[22] Ankita Pasad, Bowen Shi, and Karen Livescu, "Comparative layer-wise analysis of self-supervised speech models," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.