

On the Disentanglement and Robustness of Self-Supervised Speech Representations

¹Yanjue Song, ²Doyeon Kim, ¹Nilesh Madhu, and ²Hong-Goo Kang

¹IDLab, Ghent University - imec, Ghent, Belgium

² Dept. of Electrical and Electronic Eng., Yonsei University, Seoul, Korea

Abstract—This paper conducts an analysis of latent embeddings generated by a range of pre-trained, self-supervised learning (SSL) models. Departing from conventional practices that predominantly focus on examining these embeddings within the realm of speech recognition tasks, our study investigates the characteristics associated with speakers and their behavior under the influence of input distortions. We establish a controlled setting with varying background noise levels and different room impulse response conditions to assess the robustness of these embeddings. We measure speaker-related information by utilizing repetitive sentences spoken by multiple speakers. The results demonstrate that the robustness of pre-trained SSL models is influenced by the type and severity of distortion, whereas the inclusion of speaker information is determined by the specific pre-training approach employed. This distinct perspective offers valuable insights into the versatility and limitations of SSL models.

Index Terms—Self-Supervised learning, Speech representation analysis

I. INTRODUCTION

Deep learning has had a significant influence on the domain of speech processing. However, the availability of a massive amount of labelled data is essential for attaining top-tier model performance. The scarcity of labeled data pairs for training has prompted researchers to develop various self-supervised learning (SSL) methods. These methods eliminate the requirement for explicit target speech data during the model training process.

Based on the training methodology, self-supervised models can be categorized [1] into generative [2]–[6], contrastive [7]–[9], and predictive [10]–[12] models. Generative models are designed to predict future information by relying on preceding data. On the other hand, contrastive models utilize anchor representations as part of their training process to distinguish positive samples from negative samples, which are obtained within the model, while predictive models compute the target with a completely separate model, which is more similar to teacher-student training. These pre-trained models, which are trained on massive data, offer the benefit of adaptability to downstream tasks. This adaptability involves the utilization of the model’s hidden states or fine-tuning it for the specific downstream task. The characteristics of the hidden states in self-supervised learning models have led researchers to incorporate them into large language models. In [13], [14], the w2v-BERT model is employed to extract semantic information from input speech.

There is no doubt that SSL models enhance the performance of downstream tasks when training the back-end

model with SSL representations, especially for tasks with limited annotated data. Yet, it remains to be answered what information is preserved in the latent embeddings and how these models perform in complex acoustic environments for real-world applications. Some investigation into these queries has been carried out in associated applications. The evaluation of SSL models through keyword spotting or speaker verification [6], [11] offers some insights into the information present in these embeddings. With the help of a speech synthesis system, the information disentanglement properties of speech representations from three distinct SSL models are investigated in [15] through the assessment of various tasks. Furthermore, SUPERB [16] is proposed as a universal benchmark framework to ensure a fair comparison of different SSL models. It provides the standard datasets and metrics for the training and evaluation of compact back-end models across a range of downstream tasks. Since the pre-trained SSL model remains fixed and only a small back-end is trained, the scores obtained reflect the quality of the embeddings considering the given task. It should be noted that all of these evaluations are conducted using clean speech data. When enhancing or separating speech, it is crucial to consider potential interference caused by noise or cross-talk. The framework of SUPERB is expanded even further in [17], showing the effectiveness of SSL models in speech enhancement and separation tasks.

Our research aims to improve the efficiency of selecting an appropriate SSL model and accelerating system design by introducing a novel analysis of SSL representations that addresses two key aspects. Firstly, our analysis will show what kind of information is preserved within different SSL representations. Once this information is identified, one can select the pre-trained model that aligns with the requirements of the downstream tasks, or alternatively, integrate additional components into the system to compensate for any missing information. While there have been some investigations about this topic, these studies have primarily focused on different layers of a pre-fixed SSL model [15], or have been limited to specific aspects according to particular system designs [18]. The lack of a comprehensive understanding of the information preserved in embeddings often leads to the optimization of the entire system through an iterative process of trial and error. Considering the large amount of available SSL models and the multiple hidden layers within each model, it would be very inefficient, if not impractical, to explore all possible combinations exhaustively. Consequently, an indication on the

information preserved by the latent embeddings, which can be inferred from the embeddings themselves, would be helpful.

Secondly, there is a gap between the SSL training data and the real acoustic environment encountered during the inference stage. Distortions such as noise and reverberation are not always included in the SSL pre-training, but they are common in numerous audio applications. These distortions differ from the interference introduced by masking methods or pseudo prediction methods, which are commonly employed in unsupervised training schemes. The mismatch between the training tasks and the application scenarios in the real-world application may lead to performance degradation of SSL models. SSL models have been reported to exhibit strong performance in adverse environments, such as separating [17] or enhancing speech [15], [17]. Nevertheless, it remains unclear to what extent the latent embeddings are affected by audio interference. In this work, we introduce an analysis of the robustness of SSL representations, which should be considered when selecting SSL models for real-world applications. In addition, the level of robustness could also serve as an indicator of how much fine-tuning is necessary for the system to operate as intended.

II. EXPERIMENTS

A. Methodology

We select four well-known SSL models in the following analysis, each with a diverse training scheme, namely: HuBERT [10], TERA [6], wav2vec 2.0 [8], and wavLM [11]. To investigate the content retained within self-supervised speech representations, we utilize the TIMIT dataset [19], given its comprehensive annotations at both phoneme and word levels. By employing sentences shared among different speakers, it becomes straightforward to create a controlled dataset to examine the information contained within the embeddings. If the embedding effectively preserves a specific type of information (such as phoneme information), then the representations containing similar information (extracted from the same phoneme, for instance) should exhibit proximity to each other. Thereby, the hyperspace should be separable based on the preserved information. Inspired by this intuition, we extract the latent embeddings from clean TIMIT data and assess how closely their distribution correlates with various labels, which reveals the primary information encapsulated in the embeddings.

To illustrate how this analysis can shed light on the information preserved in SSL representations, we visualize all the averaged embeddings of the phoneme ‘ao’ from TIMIT sentence ‘sa1’, spoken by all speakers in the training set, using t-SNE [20] in Figure 1. The selection of this particular phoneme for illustration is due to its high frequency in the sentences, and its occurrence in multiple words across the sentence. The low-dimensional projection provided by t-SNE aims to retain local similarities as effectively as possible, making it a valuable tool for visualizing the distribution of latent embeddings in high-dimensional space. In the plot, every sample point corresponds to the average embedding of the consecutive frames of a single phoneme, and the same

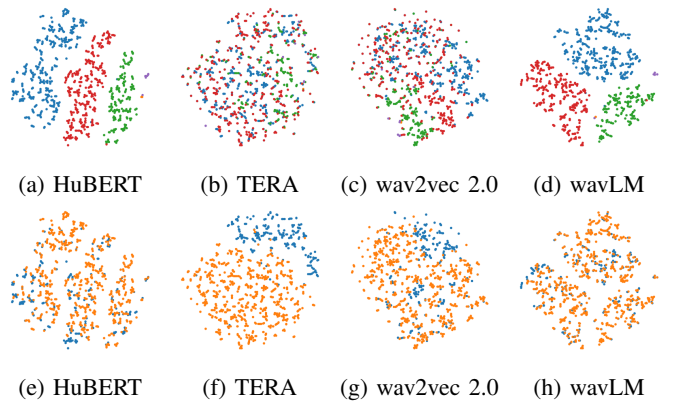


Fig. 1: The embedding distributions of all ‘ao’ sounds in ‘sa1’ from TIMIT training set, visualized by t-SNE. Each column presents the embeddings extracted by one SSL model. The plots in the first row are labeled by the words to which the phoneme belongs, and the second row by the speaker genders.

distribution is represented using two distinct labels: the word associated with the phoneme (top row), and the speaker’s ID (bottom row). It is clear that certain models predominantly keep contextual information (forming a cluster according to the word information), while others are more influenced by the speaker information (with minimal overlap between male and female speakers). It is important to acknowledge that there is an inherent loss of information when projecting embeddings into a lower-dimensional space for visualization. Therefore, the absence of clear clustering based on one type of label does not necessarily imply the complete loss of that information.

In the robustness analysis, the inclusion of background noise and reverberation is considered to simulate the real-world recording conditions. The robustness of pre-trained SSL models is evaluated using the Valentini dataset [21] for speech, DEMAND noise dataset [22], and the MIT Impulse Response Survey dataset [23] for simulating distortion. We used one male and one female speaker, along with five different background noise types from various environments: living room, office, cafeteria, square, and bus noise, each at signal-to-noise ratio (SNR) levels of [-7, 0, 5, 10, 15] dB. With respect to the distortion, we utilize three distinct datasets: a noisy test set, a reverberation (reverb) test set, and a combined test set containing both noise and reverberation (all).

B. Metrics

To gauge the accessibility of specific information within the embeddings, we choose the training score of a logistic regression model. It indicates how easily the hyperspace can be separated in a linear manner when the embeddings are labeled based on that particular piece of information.

To measure the degree of embedding distortion resulting from audio interference, we employ two metrics to quantify this effect. The first metric relies on the mean square error (MSE) between the clean and the distorted embeddings. In addition, we propose to normalize the embeddings by removing the mean and variance of the clean set before

TABLE I: Logistic regression model training accuracy. Embeddings are extracted from TIMIT training set and averaged at phoneme level.

Data source	Target	Accuracy (%)			
		HuBERT	TERA	wav2vec2.0	wavLM
sa1	Phoneme	93.2	86.8	89.1	92.7
	Word	99.2	94.5	95.6	99.0
sx	Sentence	98.7	73.8	93.0	92.9
	Speaker	90.0	94.5	94.7	53.0

calculating the MSE. This normalization approach allows for a direct comparison of the robustness of different models, even when they embed the same audio into different latent spaces. Referring to the N -dimension embedding of one frame from the distorted signal as \mathbf{e}_s and its clean reference as \mathbf{e}_x , the normalized MSE between the two can be computed as follows:

$$d(\mathbf{e}_s, \mathbf{e}_x) = \frac{1}{N} \left(\frac{\mathbf{e}_s - \mathbf{e}_x}{\sigma} \right)^T \cdot \left(\frac{\mathbf{e}_s - \mathbf{e}_x}{\sigma} \right), \quad (1)$$

where σ is the variance of the *clean* data set embeddings.

The second measure we employ is the cosine similarity (CS) distance, which calculates the similarity between the clean and distorted embeddings in a polar coordinate system. With \mathbf{e}_s and \mathbf{e}_x represent the distorted embedding and its clean reference of one frame, respectively, the cosine similarity $c(\mathbf{e}_s, \mathbf{e}_x)$ between them is defined as follows:

$$c(\mathbf{e}_s, \mathbf{e}_x) = \frac{\mathbf{e}_s \cdot \mathbf{e}_x}{\|\mathbf{e}_s\| \cdot \|\mathbf{e}_x\|}, \quad (2)$$

where $\|\cdot\|$ denotes the L2 norm.

III. RESULTS

A. Preserved Information

We analyze the type of information extracted by the SSL models when the input is clean. The embedding is aggregated by averaging the embeddings of the same phoneme in consecutive frames. Since word, sentence, and speaker information is annotated in TIMIT, we systematically investigate these different types of information one by one. In Table I, we provide the training scores, which represent the average prediction accuracy on the clean TIMIT training dataset.

The first two rows of the table compare the significance of contextual information (predicting words) and phonetic information (predicting phonemes) within the embeddings. Since each speaker utters the same sentence ('sa1'), the cross-frame information (stemming from the transformer architecture) remains consistent for the same word. Therefore, the training scores, which signify the challenge of distinguishing phonemes or words in the latent space, are solely affected by the information preserved by the model. In all four models tested, classifying words is found to be a more straightforward task compared to classifying phonemes. This observation indicates that at the final hidden layer, all models tend to preserve a higher degree of contextual information compared to phonetic information. HuBERT exhibits the highest performance on both tasks.

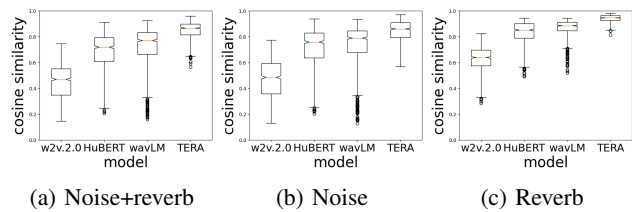


Fig. 2: Cosine similarity between the clean and distorted embeddings from pre-trained SSL models. Three types types of distortions (noise+reverb, noise, reverb) are simulated.

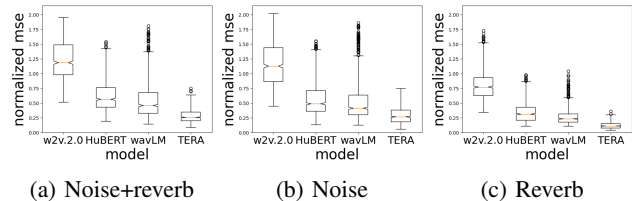


Fig. 3: Standardized MSE between the clean and distorted embeddings from pre-trained SSL models. Three types types of distortions (noise+reverb, noise, reverb) are simulated.

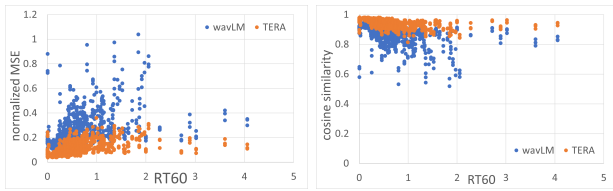
The following two rows involve a more comprehensive set of sentences, specifically all the phonetically-compact sentences ('sx') in the training dataset, which comprises a total of 330 unique sentences uttered by 462 speakers (5 sentences per speaker). This analysis provides a more extensive perspective on contextual information (i.e. sentence classification). Given the greater diversity of sentences, the training scores for speaker classification serve as an indicator of whether speaker information is preserved within the embeddings. The results clearly reveal a substantial contrast in the predominant information extracted by different models. It is challenging to deduce long-term contextual information from TERA representations, as evidenced by the low accuracy in sentence ID prediction from phonemes. However, TERA performs as the second-best model in terms of speaker information preservation, with only a slight 0.2% lower performance compared to wav2vec 2.0 in speaker classification. On the contrary, wavLM, which ranks as the second-best in preserving contextual information with a 92.9% accuracy in sentence classification, significantly diminishes speaker information in its last hidden layer, achieving only 53.0% accuracy in speaker classification.

B. Distortion

To assess robustness with respect to various distortion types, we calculated the standardized mean squared error and cosine similarity for each pre-trained SSL model, as depicted in Figure 2 and Figure 3. Notably, the introduction of noise has a more detrimental impact on the robustness of the last hidden states in comparison to the distortion caused by reverberation. The further decline in robustness on the noise+reverb test set indicates that a combination of distortions exacerbates the quality of the embeddings. Among the pre-trained SSL models, TERA consistently exhibits the highest robustness across all test sets and measurements, with wavLM coming in

TABLE II: Analysis of pretrained SSL model according to the SNR (dB) of noise distortion. ↓ means MSE is lower the better and ↑ means CS is higher the better.

Model		-7	0	5	10	15
wavLM	MSE ↓	0.967	0.593	0.430	0.352	0.295
	CS ↑	0.521	0.701	0.778	0.816	0.847
TERA	MSE ↓	0.452	0.334	0.265	0.212	0.166
	CS ↑	0.746	0.818	0.859	0.889	0.915



(a) Standardized MSE (b) Cosine similarity

Fig. 4: The influence of T60 on speech representations. Yellow is TERA and blue is wavLM.

second. The superior performance of TERA can be attributed to the augmentation techniques employed during training, which involved a range of masking and dropout methods. For a more in-depth analysis, we explored the results with consideration of the signal-to-noise ratio (SNR) and the room impulse response’s RT60 parameter, as presented in Table II and Figure 4, focusing on top two models that exhibited similar performance: wavLM and TERA. In terms of additive noise, the distinction between the clean and distorted embeddings becomes more prominent when SNRs decrease. However, the differences observed in the context of reverberation were not as substantial, as illustrated in Figure 4. Interestingly, when evaluating the performance of TERA and wavLM on the reverberation test set, both models exhibited comparable mean performance. However, it is worth noting that TERA demonstrated significantly lower variance among samples compared to wavLM, highlighting its outstanding robustness compared to other pre-trained models.

IV. CONCLUSIONS

In this paper, we proposed to evaluate the quality of latent embeddings from SSL models directly through the embeddings themselves. This approach can furnish valuable insights for the selection of SSL models for specific downstream tasks, especially in the presence of noise and reverberation. We conducted an investigation on the latent embeddings from the last hidden layer of four well-known pre-trained models that were trained by various training schemes. Our analysis, based on embeddings from annotated clean speech, reveals that all four examined pre-trained SSL models tend to prioritize contextual information over phonetic information. The preservation of long-term contextual information and speaker information is contingent on the training scheme employed in SSL. For the practical application of SSL models in complex acoustic environments, we conducted a comparison of the robustness of the selected models’ embeddings. When the input audio is distorted by noise or reverberation, the embeddings from TERA are least affected in terms of both the standardized

MSE and cosine similarity. The quantitative results further highlight that additive noise has a more significant impact on the embeddings compared to reverberation.

REFERENCES

- [1] A. Mohamed, H. Lee, L. Borgholt *et al.*, “Self-supervised speech representation learning: A review,” *IEEE J. Sel. Top. Signal Process.*, 2022.
- [2] A. Van Den Oord, O. Vinyals, and K. k., “Neural discrete representation learning,” *NeurIPS*, vol. 30, 2017.
- [3] S. Pascual, M. Ravanelli, J. Serra *et al.*, “Learning problem-agnostic speech representations from multiple self-supervised tasks,” *arXiv:1904.03416*, 2019.
- [4] M. Ravanelli, J. Zhong, S. Pascual *et al.*, “Multi-task self-supervised learning for robust speech recognition,” in *ICASSP*, 2020, pp. 6989–6993.
- [5] A. Liu, S. Yang, P. Chi *et al.*, “Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders,” in *ICASSP*, 2020, pp. 6419–6423.
- [6] A. Liu, S. Li, and H. Lee, “Tera: Self-supervised learning of transformer encoder representation for speech,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2351–2366, 2021.
- [7] A. Baevski, S. Schneider, and M. Auli, “vq-wav2vec: Self-supervised learning of discrete speech representations,” in *ICLR*, 2020.
- [8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *NeurIPS*, vol. 33, pp. 12 449–12 460, 2020.
- [9] Y. Chung, Y. Zhang, W. Han *et al.*, “W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training,” in *ASRU*, 2021, pp. 244–250.
- [10] W. Hsu, B. Bolte, Y. Tsai *et al.*, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3451–3460, 2021.
- [11] S. Chen, C. Wang, Z. Chen *et al.*, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [12] A. Baevski, W. Hsu, Q. Xu *et al.*, “Data2vec: A general framework for self-supervised learning in speech, vision and language,” in *ICML*, 2022, pp. 1298–1312.
- [13] Z. Borsos, R. Marinier, D. Vincent *et al.*, “Audiolm: a language modeling approach to audio generation,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, 2023.
- [14] A. Agostinelli, T. Denk, Z. Borsos *et al.*, “Musiclm: Generating music from text,” *arXiv:2301.11325*, 2023.
- [15] K. Hung, S. Fu, H. Tseng *et al.*, “Boosting Self-Supervised Embeddings for Speech Enhancement,” in *Proc. Interspeech 2022*, 2022, pp. 186–190.
- [16] S. Yang, P. Chi, Y. Chuang *et al.*, “SUPERB: Speech Processing Universal PERFORMANCE Benchmark,” in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.
- [17] Z. Huang, S. Watanabe, S. Yang *et al.*, “Investigating self-supervised learning for speech enhancement and separation,” in *ICASSP*, 2022, pp. 6837–6841.
- [18] A. Polyak, Y. Adi, J. Copet *et al.*, “Speech Resynthesis from Discrete Disentangled Self-Supervised Representations,” in *Proc. Interspeech 2021*, 2021, pp. 3615–3619.
- [19] J. Garofolo, L. Lamel, W. Fisher *et al.*, “The DARPA TIMIT acoustic-phonetic continuous speech corpus,” *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.
- [20] L. Van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, no. 11, 2008.
- [21] Valentini-Botinhao, C. and others, “Noisy speech database for training speech enhancement algorithms and its models,” 2017.
- [22] J. Thiemann, N. Ito, and E. Vincent, “DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments,” in *Proc. Meetings Acoust.*, 2013, pp. 1–6.
- [23] J. Traer and J. McDermott, “Statistics of natural reverberation enable perceptual separation of sound and space,” *PNAS*, vol. 113, no. 48, pp. E7856–E7865, 2016.