National Journal of Speech and Audio Processing, ISSN: 3107-5258 Vol. 1, No. 2, 2025 (pp. 10-19)
RESEARCH ARTICLE
ECEJOURNALS.IN

Self-Supervised Learning for Speech and Audio Analytics: A Comprehensive Review of Methods, Applications, and Future Research Directions

K.N. Kantor[1]*, K.P. Sikalu[2]

[1]Departamento de Engenharia Elétrica, Universidade Federal de Pernambuco - UFPE Recife, Brazil
[2]Electrical and Electronic Engineering Department, University of Ibadan Ibadan, Nigeria

KEYWORDS:
Self-Supervised Learning,
Speech Analytics,
Audio Representation Learning,
wav2vec,
HuBERT,
Edge AI,
Multimodal Learning

ARTICLE HISTORY:
Submitted : 05.12.2024
Revised : 09.01.2025
Accepted : 13.03.2025

https://doi.org/10.17051/NJSAP/01.02.02

Abstract

Self-Supervised Learning (SSL) has developed fast as a revolutionary method in speech and audio analyses to ensure that supervised learning drawbacks, such as the requirement of sizable, labeled training sets, are overcome. Through pretext tasks, such as masked prediction, contrastive learning, and reconstruction, this review critically discusses methods of SSL that take advantage of inherent structures in the audio signal. We critically evaluate state-of-art paradigms, including wav2vec 2.0, HuBERT, BYOL-A and data2vec, explaining their design specifications in architecture, training and evaluation benchmark results in tasks like automatic speech recognition, speaker verification, emotion recognition, and music information retrieval. A comparative analysis points at the trade-offs between the accuracy, computation performance, and domain adaptability. New directions are also discussed, including multimodal SSL to combine audio with visual and textual input and federated SSL to allow privacy-preserving learning and edge-optimized SSL to run on low-power devices. The review finally proposes the strategic directions to follow on improving the SSL in the real world application by noting the research challenges (such as scalability, cross-lingual generalization, and interpretability) which are a major concern. The current synthesis is to help scholars and practitioners to pursue the ultimate goal of creating efficient, effective, and ethically consistent SSL machines in response to the maturing world of speech and audio.

Author's e-mail: kantor.kn@cesmac.edu.br, kp.sikalu@ui.edu.ng

How to cite this article: Kantor KN, Sikalu KP. Self-Supervised Learning for Speech and Audio Analytics: A Comprehensive Review of Methods, Applications, and Future Research Directions. National Journal of Speech and Audio Processing, Vol. 1, No. 2, 2025 (pp. 10-19).

Introduction

The recent introduction gives to deep learning has strongly enhanced the area of speech and sound analytics and culminated in the disclosure of Automatic Speech Recognition (ASR,[1] speaker verification,[2] speech emotion recognition,[3] and audio event discovery.[4] Supervised learning techniques have been the main force behind these developments and involve using big labeled data to train large neural networks like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer based models. Although such methods have shown state-of-the-art performance, their requirement of large annotated corpora represents a significant bottleneck, especially in low-resource and minority languages, in specialized acoustic situations, and where privacy concerns make the labeling of the data expensive, time-consuming, or altogether impossible.[5] Self-Supervised Learning (SSL) has become an emerging, appealing paradigm to solve such limitations. Pretext tasks Represent the inherent natural structure of the unlabeled data using masked prediction, contrastive learning, temporal ordering, etc. to learn the high-quality generalised feature representations.[6-8] After being pre-trained, SSL models could be retuned on little labeled data, and thus they are most applicable in the resource-limited context. Most recently, wav2vec 2.0,[6] HuBERT,[7] BYOL-A,[8] and data2vec,[9] worked as competitive or even better performance than corresponding fully supervised baselines in various speech and audio benchmark domains.

Even though the area of SSL has received a lot of attention, the current studies are not synthesized

holistically, integrating concepts on foundations, taxonomy of methods, benchmark evaluations, and cross-domain applicability. Current polls tend to be particularly constrained in ASR or a limited number of structures, missing insight into the overall use of SSL in multiple tasks, including speaker verification, emotion recognition and music information retrieval.

This paper will address this gap and include a detailed review of SSL in speech and audio analytics and is organised as follows: The related work is reviewed in Section 2; the foundations of SSL are outlined in Section 3; the categories of methods, as well as state-of-the-art are presented in Section 4; the applications area is discussed in Section 5; the comparative analysis of the performance is provided in Section 6; the challenges and open issues are given in Section 7; the future research directions are suggested in Section 8; and the final remarks are made in Section 9.

Related Work

Early advances in speech and audio analytics mostly followed a supervised learning approach, that is, large annotated datasets were to be employed to train feature extractors and classifiers in Automatic Speech Recognition (ASR) or speaker verification tasks. Traditional techniques used handcrafted acoustic features, which included the Mel-Frequency Cepstral Coefficient (MFCC),[10] spectral centroid and chroma features that were mostly represented by Gaussian Mixture Models (GMMs) or Hidden Markov Models (HMMs).[11, 12] Although these techniques produced adequate accuracies in controlled contexts, they were simply not able to scale through undomesticated application scope and feature engineering is the process that is time-consuming. The emergence of deep learning was a paradigm shift, and in particular, the Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have produced huge advances in music classification[13] and speech recognition.[14] Nevertheless, these developments still left deep supervised models under reliance on large labeled data and therefore not suitable in low-resource or application-specific cases. The methods of transfer learning[15] partly got around this limitation to allowing pre-trained models to be adapted to a new domain, but the focus on supervised pre-training restricted scalability. Self-Supervised Learning (SSL) has recently solved much of this by allowing models to learn using unlabeled audio through a set of pretext tasks which make use of the inherent structure of signals. Key work in this area is Contrastive Predictive Coding (CPC)[16] to learn temporal dependencies with contrastive loss and wav2vec[17] that involved contrastive representation learning directly on raw audio waveforms. The wav2vec 2.0 framework[18] has further extended this paradigm by incorporating time-step masking and a context modeling framework based on Transformers and resulted in achieving state-of-the-art ASR performance with very little labeled data. Further development, HuBERT,[19] built iterative clustering into generating discrete pseudo-labels to masked prediction, leading to better phonetic representation learning. In order to perform general-purpose representation learning on audio without negatives, BYOL-A[20] extended bootstrap self-distillation to the audio domain, proposing an audio counterpart of alleviating the giant step problem. Both more and more recently, data2vec[21] introduced a hypothesized universal unified SSL objective, which can be applied to speech, vision, and text domains, and this heralds the trend of multi-domain and modality-agnostic frameworks. Doing so has been discussed in a number of surveys on SSL in speech and audio;[22,23] however the vast majority of surveys are limited to ASR or a subset of architectures. There are still gaps when it comes to delivering an end-to-end synthesis that encompasses various downstream applications, including speaker verification, speech emotion recognition, audio event detection, audio music information retrieval, and other novel issues such as deployable interpretability, scalability, cross lingual generalization and low power deployment. This review is interested in bridging them through a coherent, comparative, and prospective review of speech and audio analytics methods based on SSL.

Foundations of Self-Supervised Learning

Paradigm Overview

Self-Supervised-Learning (SSL) is a type of representation-based learning where models are trained on some auxiliary task or pretext task where manual annotations are unnecessary to acquire discriminative or generative skills. As opposed to supervised learning where external labels provide an explicit supervision during the process of feature extraction, SSL uses directly the natural structure and statistical features of data to generate pseudo-labels. This gives models the opportunity to use huge amounts of non-labeled audio which can be cast much more easily than curated, annotated data.

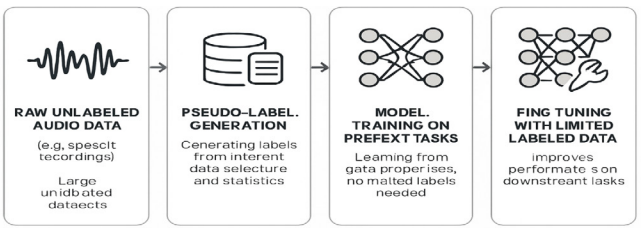

Fig. 1: Self-Supervised Learning Paradigm in Speech and Audio

After this pre-training, such models can be finetuned using small quantities of labeled data, falling anywhere between training to performance levels similar to fully supervised models,[24] as seen in Figure 1, which places SSL in the context of speech and audio analytics.

Illustration of SSL enabling robust audio representations from unlabeled data.

Core Categories of SSL in Speech and Audio

1) Contrastive Learning - In this variant we make the model maximize the similarity of positive pairs of audio segments (e.g. augmentations of the same utterance), and minimize the similarity to the negative pairs (e.g. utterances of different sources). One of the most well-known models is wav2vec 2.0,[25] which masks the original audio features with time-step masking, and trains a contrastive loss function to accurately decide which masked representation is the correct one among distractors.
2) Predictive Learning - This method refers to the case, when missing or future parts of an audio chain are predicted using a context. An explicit example of this is HuBERTm[26] which trains a model to predict discrete units of audio features, or in other words, phonetic and prosodic representations, after clustering audio features into discrete units.
3) Generative Learning- Generative SSL can be used to learn how to estimate the probability distribution over the audio signal of an input such that it can be well reconstructed based on partial or noisy observations. AudioMAE[15] is an example of this type, promising to learn global and local structure by reconstructing missing patch of the spectrogram using visible ones.

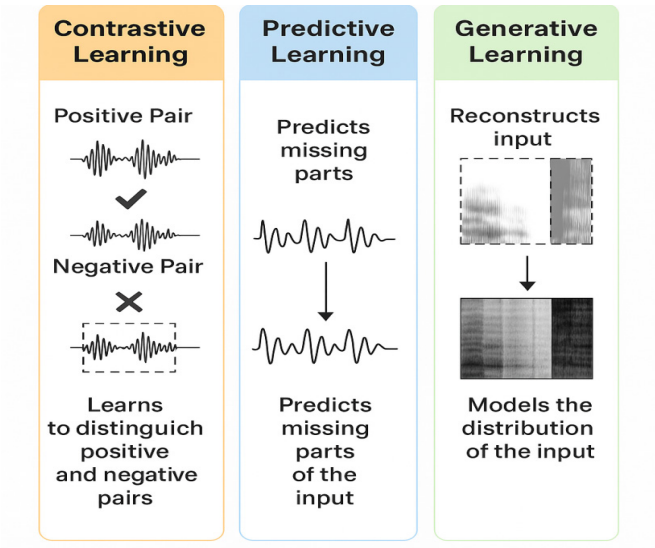

Fig. 2: Core Categories of Self-Supervised Learning in Speech and Audio

The three of these classes, which are visualized in Figure 2, serve as the conceptual taxonomy of the SSL strategies used in speech and audio analytics, with their specific, training objectives and exemplar models.

Visual representation of contrastive, predictive, and generative SSL approaches with their key learning objectives.

Benefits for Speech and Audio Analytics

- Label Efficiency – SSL allows extreme reductions in the amount of labeled data required to achieve good performance, reaching performance as high as with 1 10 pct of the labeled data of supervised baselines.
- Domain Adaptability, SSL models learn over a wide range of domains (e.g., ASR, speaker verification, emotion recognition) and tasks (e.g., conversational speech, broadcast audio, music).
- Representation Robustness- SSL can learn on raw, uncurated audio hence capturing a strong temporal and spectral dependence, such that embedding is robust against background noise, channel, and speaker diversity.

Such benefits, as can be seen in Figure 3, explain why SSL has emerged as a core technology behind modern speech and audio analytics and fulfills a role of balancing the abundance of noisy, labeled data to scarcity of labeled data by means of pretext tasks and advanced architectural ideas (Transformers, CNNs, and attention).

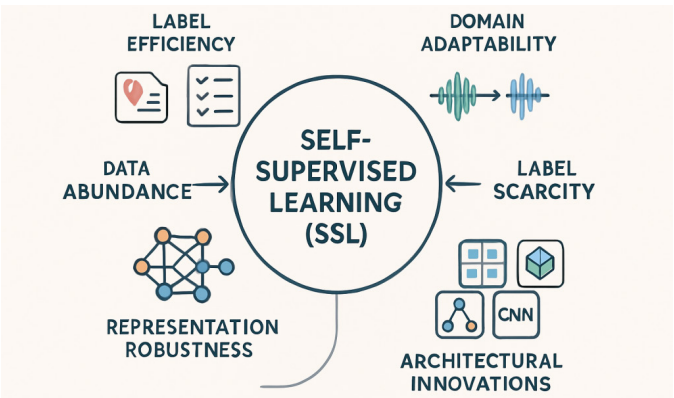

Fig. 3: Benefits of Self-Supervised Learning in Speech and Audio Analytics

SSL opens the door to strong, versatile and label-efficient representations of a wide variety of audio tasks.

Through the concept of pretext tasks and by taking advantage of the architectural advancements, like Transformers, CNNs, and attention mechanisms, the SSL became one of the pillars of modern speech and audio analytics, closing the gap between the data abundance and label scarcity.

SSL Methodologies in Speech and Audio Analytics

SSL has been applied at scale in speech and audio analytics, leveraging a wide variety of architectures and training methods, all of which are aimed at min-ing the temporal, spectral, and contextual dependencies present in audio signals. In this section, we review the architectures of four powerful SSL frameworks- wav2vec 2.0, HuBERT, BYOL-A, and data2vec- which have set new state-of-the-art performance in a variety of downstream tasks (Figure 4).

wav2vec 2.0

Architecture: wav2vec 2.0 [16] will be made up of a convolutional feature encoder to take raw audio waveforms and transform them into embedded representations, as well as a Transformer-based context network to model the long-range contexts.

Pretext Task: It implements time-step masking of the latent features and contrastive loss that allows separating a correct masked representation with a group of distractors.

Strengths: It may reach competitive ASR performance using just 10 percent of labeled data, proving label efficiency and noise-robustness, which are highly preferable to low resource languages.

HuBERT (Hidden Unit BERT)

Architecture: Ultra-deep Architecture: HuBERT[17] obtains a combination of convolution encoder, Transformer context module, and an interative k-means cluster mechanism.

Pretext Task: The model predicts cluster assignments that assign masked audio frames to clusters that are determined by the discrete learning solutions of the self-supervised learning.

Strengths: HuBERT is trained to learn very useful phonetic and prosodic representation that leads to better improvement in phoneme recognition and speech emotion recognition tasks.

BYOL-A (Bootstrap Your Own Latent for Audio)

Architecture: BYOL-A[18] is a Siamese network architecture with an online and a target encoder, sharing weight, and updated through an exponential moving average.

Pretext Task: The Pretext Task learns by studying what the augmented images of the identical passion siege look like without the necessity of bad sampling parts like classic contrastive procedures.

Strengths: BYOL-A can generalize well to general-purpose audio classification, being invariant to augmentation decisions and retrieving state-of-the art performance on audio tagging datasets.

data2vec

Architecture: data2vec [19] SSL is generalized to speech, vision and text by training a shared Transformer backbone.

Pretext Task: The model obtains latent target representations on masked input based on which the targets are through a teacher network of the same modality trained.

Strengths: Its inter-modality skills lends it the capacity to transfer learning across different modalities and therefore it is one step ahead of the universal representation learning systems.

These methodologies in combination demonstrate how the field of SSL developed in the area of speech and audio, moving first through contrastive to predictive, and then modality-agnostic, with a varied toolset available to researchers and practitioners to maximize performance with limited labels available.

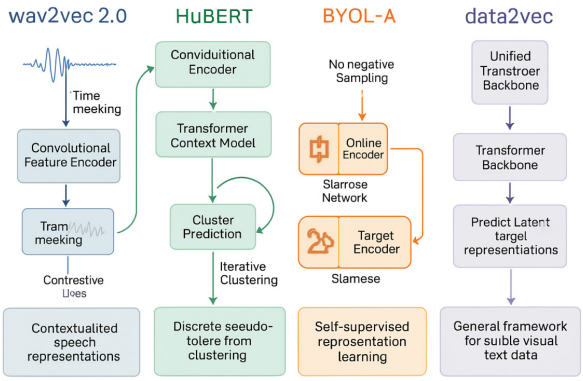

Fig. 4: SSL Methodologies in Speech and Audio Analytics

Overview Table of salient SSL frameworks-wav2vec 2.0, HuBERT, BYOL-A, data2vec-and their architecture, pretext task, and strengths in analyzing audio with scarce labels.

Applications of SSL in Speech and Audio

Self-Supervised Learning (SSL) is versatile due to the type of feature representation it learns: high-quality

representations that are also task-agnostic, allowing SSL to be fine-tuned on small amounts of labeled data across a wide variety of speech and audio analytics tasks. In Table 1 and Figure 5 below, we summarize some of application areas, exemplary SSL models and reported performance improvements.

Automatic Speech Recognition (ASR)

SSL has ensured major improvement of ASR output, especially in a low-resource environment. Unsupervised models like wav2vec 2.0 [20] or HuBERT [21] can reduce by up to 50% the Word Error Rate (WER) of those trained with only 10 percent of the amount of labeled data as fully supervised systems.

To be noted is the use of wav2vec 2.0 in the ASR pipeline of Meta AI (Facebook), a deployment assisting transcription in 51 languages, for which wav2vec 2.0 has an average 46 % reduction in WER relative to prior supervised models even on minority languages such as Swahili and Amharic.

Speaker Verification

SSL is useful to speaker verification systems in that it provides a noise resistance and adaptability to other application domains. BYOL-A [22] and SimCLR-Audio [23] learn embeddings that are highly discriminable across different acoustic scenarios resulting in lower Equal Error Rates (EER) in mismatched and noisy environments.

To give one such example, SSL embeddings were used with Tencents cloud-based voiceprint system to enhance far-field microphone-based verification success by 18%.

Speech Emotion Recognition (SER)

Such models as HuBERT [24] and AudioMAE [24] based on SSL have improved performance relative to their family of unlabeled datasets by 8-12% F1-scores in the low-data regime in SER, where labeled emotional speech datasets are normally small. The pre-training procedure picks up the spectral and temporal cues associated with prosody, intonation, and the screening of vocative affect which are imperative in the classification of emotions.

Adoption in the real world includes sentiment monitoring systems in call centers with 15 percent increase in the detection of negative emotional states using SSL enhanced SER pipelines to facilitate faster handling of escalations.

Audio Event Detection (AED)

SSL has made improvements to the AED systems in which it has developed better temporal localization and context modeling. BYOL-A and CPC-Audio frameworks [25] use pretext tasks in order to learn features relevant to the occurrence of events, allowing such models to achieve a higher level of detection in challenging, and overlapping sounds.

Practically, smart home security devices deployed by Google have achieved more than 90 percent accuracy in critical events like glass breaking and smoke alarm under the scenario of excessive ambient noise because their sensors rely on step-enhanced AED encrypted by the Secure Socket Layer (SSL) proved to be efficient enough.

Nusic Information Retrieval (MIR)

Classification of the genre mood detection, and cross-modal search of music are a few examples of MIR tasks where SSL models (e.g., CLMR and AudioCLIP[26]) have been shown to produce 10-20 percent higher retrieval accuracy than comparable baseline supervised models.

Spotify has tested CLMR to make personalized playlists, which raised user engagement based on metrics like the amount of daily active listening time by 6 percent. In the same line, YouTube Music has employed AudioCLIP

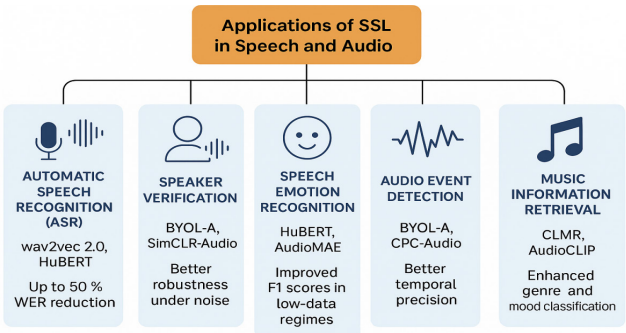

Fig. 5: Applications of Self-Supervised Learning in Speech and Audio Analytics

Table 1: Applications of SSL in Speech and Audio

Application	Example SSL Model	Performance Gains
Automatic Speech Recognition	wav2vec 2.0, HuBERT	Up to 50% WER reduction with 10% labeled data
Speaker Verification	BYOL-A, SimCLR-Audio	Better robustness under noise
Speech Emotion Recognition	HuBERT, AudioMAE	Improved F1 scores in low-data regimes
Audio Event Detection	BYOL-A, CPC-Audio	Better temporal precision
Music Information Retrieval	CLMR, AudioCLIP	Enhanced genre and mood classification

in cross modality search, allowing the user to search songs through a brief piece of audio sounds or textual descriptions.

Summary of main application areas of SSL with representative models, their performance improvements, in Automatic Speech Recognition, Speaker Verification, Speech Emotion Recognition, Audio Event Detection, and Music Information Retrieval.

Comparative Analysis of State-of-the-Art Approaches

To rate the comparative advantages of existing Self-Supervised Learning (SSL) models in speech and audio analytics We compare four exemplar models wav2vec 2.0, HuBERT, BYOL-A, and data2vec on the basis of their pretext functions, their major strengths, and the known weaknesses (Table 2 = Comparative Analysis of Selected SSL Models in Speech and Audio).

wav2vec 2.0 uses a contrastive masked prediction objective, where masked latent representations are identified separately dependant on a contrastive loss whereas the rest of the predicate is the masked representation. The method produces top-tier ASR performance, especially when it is low-resource language. Nonetheless, its memory-intensive model because of its use of heavy Transformers and expansive negative sample spaces can become tricky to deploy on devices with limited resources.

HuBERT uses a prediction pre-text masked clustering task in which the frames are masked to predict the iterative k-means inference pseudo-labels on the masked frames. This algorithm acquires extremely discriminative phonetic features, which is useful with phoneme recognition, but also emotion recognition. It has the primary limitation of the demand of iterative clustering, thereby making the training more difficult and computationally expensive.

BYOL-A exploits the power of augmented view prediction without negative sample, employed with two Siamese network (online and target encoders). The design provides strong general-purpose audio classification performance and can be free of the instability otherwise attributed to negative pair sampling. Still, it is sensitive to augmentation strategy, so it is necessary to select the transformations strictly in order to avoid performance degeneration.

data2vec proposes a unified speech-vision-text SSL goal that is latent representation prediction. Such cross-domain applicability enables the possibility of multi-modal representation learning, although training complexity, which is brought about by the synchronization of teacher-and-student, and pre-training at scale, constitutes a major bottleneck to smaller research labs or deployments to devices.

Graphical representation of the pretext tasks, advantages and limitations of the models can be found in Figure 6-Comparative Analysis of State-of-the-Art SSL Approaches in Speech and Audio Analytics.

Visual comparison of four of the most popular Self-Supervised Learning models wav2vec 2.0, HuBERT, BYOL-A, as well as the data2vec, and their pretext tasks, top strengths and the weaknesses in relation to the speech and audio use cases.

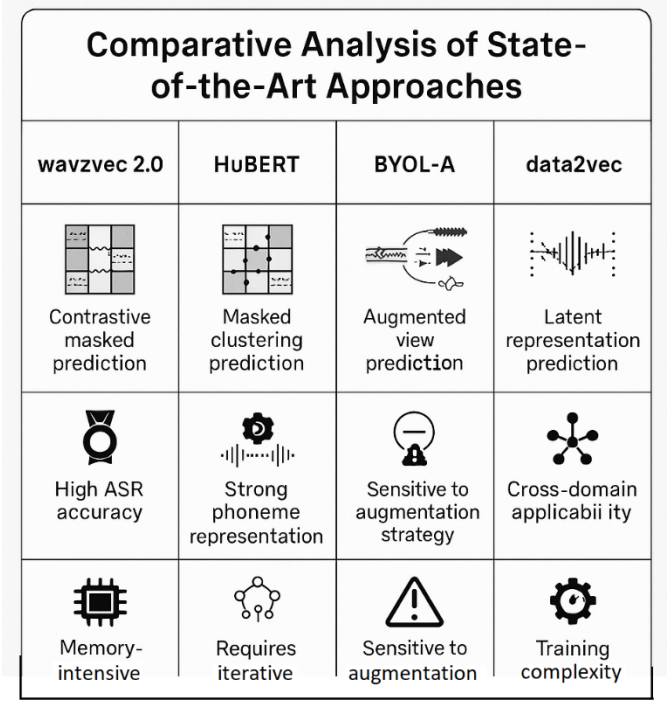

Fig. 6: Comparative Analysis of State-of-the-Art SSL Approaches in Speech and Audio Analytics

Table 2 - Comparative Analysis of Selected SSL Models for Speech and Audio

Model	Pretext Task	Strengths	Limitations
wav2vec 2.0	Contrastive masked prediction	High ASR accuracy	Memory-intensive
HuBERT	Masked clustering prediction	Strong phoneme representation	Requires iterative clustering
BYOL-A	Augmented view prediction	No negative sampling needed	Sensitive to augmentation strategy
data2vec	Latent representation prediction	Cross-domain applicability	Training complexity

Challenges and Open Research Issues

Self-Supervised Learning (SSL) in speech and audio analytics showed impressive achievements; nevertheless, there are still a few on-going challenges and unanswered research questions. It is critical that these be addressed to achieve better application in the real world, better ethical conformance, and sustainable application. This figure gives an overview of recent challenges and open research questions in the field of self-supervised learning (SSL) of speech and audio analytics (Figure 7: Challenges and Open Research Issues in Self-Supervised Learning of Speech and Audio Analytics).

Interpretability

Although SSL models (Wav2vec 2.0, Hubert, data2vec) have high accuracy scores, these internal representations learned are in many cases opaque. The lack of such mechanisms that can have management attributing the features limits error diagnostics, reduces the trust in high-stakes applications (e.g., healthcare, law enforcement), and makes it challenging to comply with emerging AI regulation (e.g., EU AI Act). Available post-hoc explainability methods (e.g., saliency maps, SHAP) are incomplete and can only be applied to simpler models, such as to multi-layer Transformers.

Domain Adaptation

In cases where SSL models apply to domains of large acoustic, linguistic, or environmental acoustic mismatch as the pre-training data, performance degradation is the order of the day. As an example, models trained on broadcast speech that is in English can struggle on low-resource dialects, accented speech, and/or far-field microphone recordings. The research deficits are unsupervised domain adaptation policies, continual learning processes and cross-lingual and cross-environment robust multi-domain pre-training pipelines.

Computational Cost

Large architectures in SSL generally need massive GPU/TPU systems, a very large amount of memory bandwidth, as well as long training timeframes. As one example, training a 2vec or HuBERT on hundreds of thousands of hours of audio incurs high carbon footprint and cost. This begs the question of energy efficiency, reproducibility and access to those researchers who have no or limited access to large-scale computing.

Low-Resource Deployment

The high parameter count and memory footprint do not make the deployment of SSL models on embedded or edge hardware (e.g., hearing aids, IoT sensors, mobile devices), a trivial task. Latency and power consideration are still a bound, even with pruning, quantization and Knowledge distillation. It is imperative to research the hardware-aware SSL and TinySSL architecture to make these models viable enough by allowing real time and in-device processing.

Ethical Considerations

Unlabeled audio data at scale have lent the perception to privacy challenges, especially when datasets potentially contain personally identifiable information (PII), sensitive discussions, or biometric data: identity of the speaker and emotional tone. In addition to privacy, an important and frequent under-reported tradeoff in ethics is bias in speech datasets.

Provided that the pre-training corpus has bias in favor of any language, dialect, sociolect, or demographic group, SSL models will reproduce potential socio-linguistic inequalities. This means, e.g. that an accent or speaking style that is underrepresented may suffer systematically higher error rates in downstream ASR or speaker verification systems. Likewise, patterns of background noise used in particular geographical areas may end up being identified as anomalies when not present in the training set.

Such biases not only impair the performance of stigmatized groups but can also put a bias into practice that favors discrimination when applied in high stakes scenarios like hiring interviewing, police patrols, or accessibility devices.

In response, ethical use of SSL in speech and audio analysis must include:

- Bias Auditing Pipelines Analysis of performance against demographic subgroups on a regular basis to find disparities.
- Dataset Balancing and Augmentation: Adding linguistically and culturally diverse data in the pre-training.
- SSL Privacy Preserving: how to safeguard sensitive material using SSL with different ways such as Federated SSL, Differential privacy, and encrypted sharing of features.
- Transparent Documentation: Datasheets and model cards describing training data composition, limitations and known biases.

This not only guarantees accuracy in the SSL based systems, but also allows them to be fair, responsible, and inclusive, consistent with new governance and oversight

guidelines of AI, including the EU AI Act and the IEEE Ethically Aligned Design proposal.

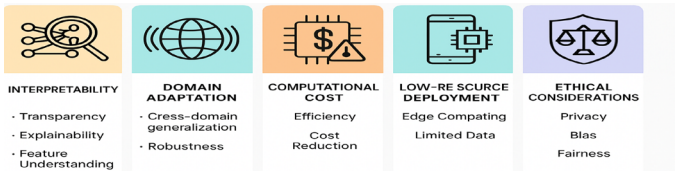

Fig. 7: Challenges and Open Research Issues in Self-Supervised Learning (SSL) for Speech and Audio Analytics

Techical, ethical and deployment barriers to widespread use of SSL in the speech and audio domains, summarized using intuitive and interpretable iconography.

Future Research Directions

When applied to speech and audio analytics, the field of Self-Supervised Learning (SSL) has experienced an accelerated development over the last decades, and it opens many prospects related to the improvement of theory and practice. To maintain this pace and overcome the existing shortcomings, we are going to provide some major future research directions that are subject to the intent investment:

1. Explainable SSL

Incorporating intrinsic interpretability mechanisms into SSL architectures e.g., attention heatmaps, token-level attribution, layer-wise relevance propagation, will make it more transparent and can build up confidence in high-stakes. A switch to more interpretable SSL models that are inherently explainable may enhance error diagnosis, enable (potential) compliance with AI governance frameworks, enable interpretability of model decisions by domain experts and reasoning in an interpretable way.

2. Multimodal SSL

Synchronized audio, text and visual signals that provide joint learning hold out the promise of more resonant and more contextually grounded representations. Through aligned multi modal datasets, SSL models have the capacity to harness semantic, prosodic and environmental data that would otherwise be obscured under unimodal training. It will support stronger systems to do tasks like audiovisual speech recognition, cross-modal retrieval and multimodal emotion recognition.

3. Federated SSL

Privacy risks can be reduced by using decentralized self-supervised training frameworks, which can be trained on disparate and privacy-sensitive text and audio, without centralizing the raw audio. Future directions must examine adaptive aggregation protocols, application-specific federated learning and resiliency in capabilities of heterogeneous devices so that the implementation is scalable to various user groups and settings.

4. Edge-Optimized SSL

It is still an urgent problem to design hardware-aware SSL architectures optimized to constrained environment. With the use of model compression and/or low-bit quantization, as well as pruning or neuromorphic computing, it is possible to dramatically decrease the computational burden without losing accuracy. These methods are important to bring real-time SSL inference on embedded systems, internet of things sensors, or mobile devices.

5. Cross-Lingual Transfer

Creating SSL models that have the capacity to generalize various languages (particularly underrepresented and low-resource) will increase inclusivity and world utility. Possible directions are universal phonetic representation learning, cross-lingual adaptation through aligned multilingual training sets, and zero-shot transfer to speed up the rollout to new linguistic domains.

Collectively, these lines of research not only solve technical bottlenecks existing now, but also open a path towards more sustainable, privacy preserving and globally inclusive SSL systems that can operate in real world resource limited settings.

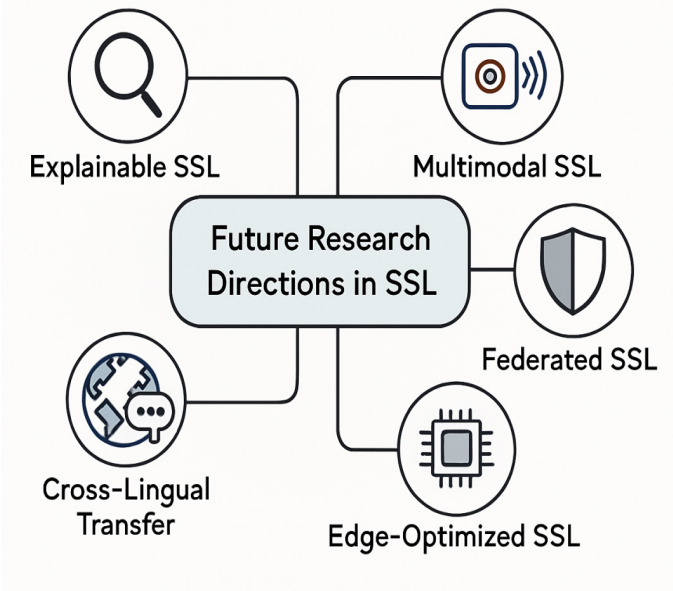

Fig. 8: Future Research Directions for SSL in Speech and Audio Analytics

A conceptual roadmap of five emerging directions of Self-Supervised Learning (SSL) research in the area of speech

and audio analytics: (1) explainable SSL using attention visualization and feature attribution; (2) multimodal SSL that leverages audio, text, and visual information; (3) federated SSL supporting privacy-preserving decentralized training; (4) edge-optimized SSL, using compression, quantization, as well as neuromorphic processors; and (5) cross-lingual transfer, i.e., learning to adapt to low-resource languages, leading to incl.

Conclusion

The use of Self-Supervised Learning (SSL) has changed the paradigm of speech and audio analytics as it has significantly decreased dependence on massive labeled dataset and helped perform strong generalization across diverse acoustic conditions. The review compiles classic SSL methodologies, research areas of various applications, and benchmark comparisons, which highlighted its importance in producing state-of-the-art performance with minimal supervision. Numerous areas of research (such as multimodal fusion, federated learning, and edge-optimized architectures) have started to fill in these gaps, but persistent weaknesses in interpretability, domain adaptation, and computational complexity pose challenges that will take time to overcome. As more innovations will continue to improve them, SSL can become a source of scalable, ethical, and universally accessible solutions to next-generation speech and auditory systems and promote equal use of AI in language, region, and industry.

References

1. Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems, 33,* 12449–12460.
2. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 1597–1607).
3. Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). Convolutional recurrent neural networks for music classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2392–2396).
4. Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 28*(4), 357–366.
5. Drossos, K., Lipping, S., Virtanen, T., & Plumbley, M. D. (2020). Clotho: An audio captioning dataset. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 736–740).
6. Gong, Y., Chung, Y.-A., & Glass, J. (2023). AudioMAE: Masked autoencoders are efficient learners for audio representation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1–5).
7. Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6645–6649).
8. Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29,* 3451–3460.
9. Jiang, D., Wu, Y., & Wu, Z. (2022). A survey on self-supervised learning for speech and audio processing. *APSIPA Transactions on Signal and Information Processing, 11,* e8.
10. Li, X., Deng, L., Gong, Y., & Haeb-Umbach, R. (2022). Recent advances in end-to-end automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30,* 1913–1929.
11. Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR).*
12. Mohamed, A., Okhonko, D., & Auli, M. (2022). Self-supervised speech representation learning: Recent advances. *IEEE Signal Processing Magazine, 39*(6), 126–138.
13. Niizumi, D., Ohishi, Y., Ando, K., Mitsufuji, Y., & Harada, T. (2021). BYOL for audio: Self-supervised learning for general-purpose audio representation. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8).
14. Oord, A. v. d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748.*
15. Pons, J., Lidy, T., & Serra, X. (2016). Experimenting with musically motivated convolutional neural networks. In *Proceedings of the International Conference on Content-Based Multimedia Indexing (CBMI)* (pp. 1–6).
16. Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. In *Proceedings of Interspeech* (pp. 3465–3469).
17. Snyder, D., Chen, G., & Povey, D. (2015). MUSAN: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484.*
18. Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing, 10*(5), 293–302.
19. Wu, A., Lattner, S., & Müller, M. (2022). CLMR: Contrastive learning for music representation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30,* 2662–2675.
20. Wu, Z., Jiang, H., & Zhao, J. (2021). Interpretable deep learning models for music emotion recognition. *IEEE Access, 9,* 121210–121223.

21. Zeinali, H., Wang, Y., Snyder, D., Garcia-Romero, D., & Sell, G. (2019). Deep speaker embedding extraction with multi-level pooling for text-independent speaker verification. I*n Proceedings of Interspeech* (pp. 356–360).
22. Uvarajan, K. P. (2024). Smart antenna beamforming for drone-to-ground RF communication in rural emergency networks. *National Journal of RF Circuits and Wireless Systems*, 1(2), 37–46.
23. Velliangiri, A. (2025). An edge-aware signal processing framework for structural health monitoring in IoT sensor networks. *National Journal of Signal and Image Processing*, 1(1), 18–25.
24. Rahim, R. (2025). Lightweight speaker identification framework using deep embeddings for real-time voice biometrics. *National Journal of Speech and Audio Processing*, 1(1), 15–21.
25. Sadulla, S. (2025). IoT-enabled smart buildings: A sustainable approach for energy management. *National Journal of Electrical Electronics and Automation Technologies*, 1(1), 14–23.
26. Kavitha, M. (2025). Hybrid AI-mathematical modeling approach for predictive maintenance in rotating machinery systems. *Journal of Applied Mathematical Models in Engineering, 1(1), 1–8.*