Dijlah Journal of Engineering Science (DJES) Vol.3 , No.2 , June, 2026, pp. 401-417

ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 174

Self-Supervised Learning for Speech Recognition: A Review

Humam Khalid Jameel[1], Assad H. Thary Al-Ghrairi[2], Mohammed M. Neamah[3]

[1, 2] Department of Computer Science, College of Science, Al-Nahrain University, Jadriya, Baghdad, Iraq.

[3] Physics Department, Education College, Mustansiriyah University, Baghdad, 10052, Iraq.

Article Info

Article history:

Received Mar., 30, 2026
Revised Apr.,29,2026
Accepted Jun.,15,2026

Keywords:

SSL
Pre-train
APC
Fine-tuning
NLP
Automatic Speech Recognition
Unsupervised Learning
Self-Supervised Learning
Datasets
ASR

ABSTRACT

For deep supervised learning algorithms to work well, a lot of labeled data is usually needed. However, gathering and classifying this kind of data may be costly and time-consuming. A subclass of unsupervised learning called self-supervised learning (SSL) seeks to cutting discriminative features from unlabeled data without the need for human-annotated labels. Recently, SSL has attracted a lot of attention, which has prompted the creation of many associated algorithms. Comprehensive studies that clarify the relationships and development of various SSL variations are scarce, nonetheless. Automatic speech recognition (ASR) has advanced significantly in recent years thanks to a variety of deep learning methods. Since deep learning methods rely heavily on data, a variety of online speech datasets are also covered in detail. We included each aspect that could affect an ASR's performance in our investigation. Therefore, we hypothesize that this work is a suitable place for scholars interested in ASR research to start.

Corresponding Author:

Humam Khlaid Jameel
Department of Computer Science, Al-Nahrain University
Al-Jadriya Street, Baghdad, Iraq
Email: humam.khalid@nahrainuniv.edu.iq

1. INTRODUCTION

Deep learning techniques have significantly improved speech processing over the last ten years, opening up a wide range of practical uses. Advances in labeled-data-rich environments were powered by supervised DNN training [1]. To avoid labeled datasets, the researchers used unpaired audio data, allowing for novel industrial speech applications and supportive low-resource languages [2]. Scientists aim to create speech representations from raw waveforms and spectral signals that contain low-level acoustic actions, lexical information, and higher-level syntactical and semantic evidence, drawing motivation from how kids acquire their original language concluded passive listening and environmental interaction The target downstream applications that require the least amount of labeled data are subsequently utilized with these learnt representations [3]. Latent features that capture essential elements of inputs are extracted via representation learning algorithms [3]. Unsupervised learning refers to machine learning procedures that seek intrinsic patterns in training information without using predefined labels or supervisory signals. Representation learning is widely recognized as a subcategory of it [4]. Examples of such methods include mixture models and k-means clustering. (SSL), a rapidly evolving subset of unsupervised learning, creates task-relevant representations by extracting supervisory labels directly from the input data. Concentrate on self-supervised learning strategies in this review. Figure 1, an example of a downstream application is automatic speech recognition (ASR).

ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 174

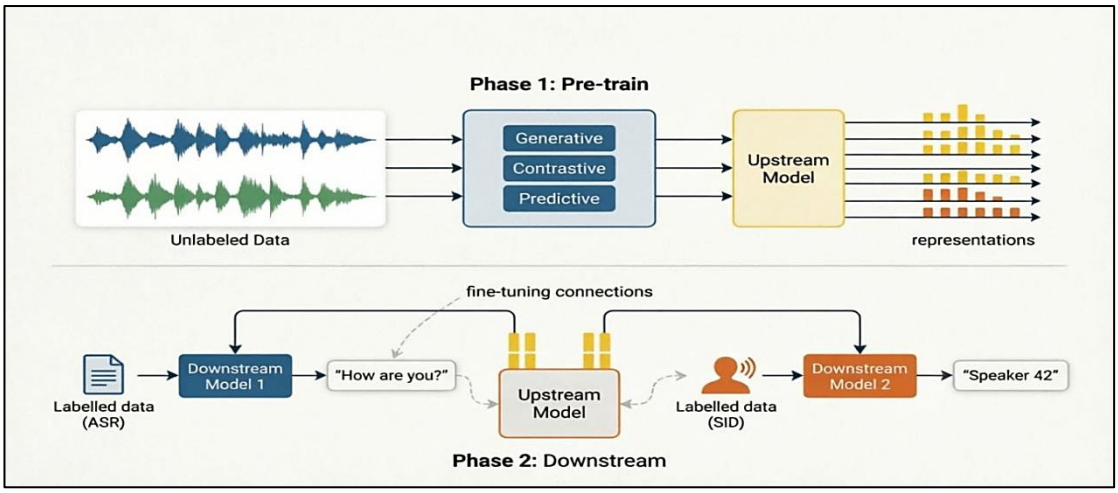


Figure 1. Framework of SSL.

Self-supervised illustration learning is depicted in Figure (1) with respect to downstream uses. This framework consists of two levels. A representation model, also known as an upstream model, is first pre-trained using self-supervised learning (SSL). This then involves either supervised tuning the entire pre-trained model or extracting its frozen representations for later tasks [5].

Owing to SSL's widespread use, assessments of the technique in general [6] and its applications to computer vision (CV) [23] and Natural Language Processing (NLP) [7] have been published. Nevertheless, SSL for speech processing is not the subject of any of these summaries. Many theories and solutions have been created to handle the particular issues of speech because the speech signal is very different from image and text inputs. Recent advances in self-supervised learning are not covered in one article, which discusses about deep learning model-based speech representation learning [8]. This serves as the inspiration for this speech SSL overview.

The rest of the paper is present in the following: Section 2, SELF-SUPERVISED LEARNING STRUCTURES. Section 3, SPEECH RECOGNITION SYSTEM. Section 4, SELF-SUPERVISED LEARNING APPROACHES. Section 5, CRITICAL SPEECH APPLICATIONS BEYOND ASR. Section 6, EXPERIMENTS RESULTS. Section 7, DISCUSSION, LIMITATIONS, AND FUTURE WORKS. Finally, Section 8, CONCLUSIONS.

2.SELF-SUPERVISED LEARNING STRUCTURES

2.1. Generative approaches:

2.1.1. Motivation: The primary pretext problem in this domain is to recreate the data from a constrained or altered perspective. Examples include forecasting future segments based on prior context, retrieving masked sections, and restoring the original from a distorted transformation. In this article, "generative" used to describe models whose primary goal is to target the original input, as opposed to generative models trained to select unique data instances.

2.1.2. Approaches:

2.1.2.1. Autoencoding: have been essential for discovering distributed latent sensory data representations then their inception in the mid-1990s. Autoencoders (AEs) consist of an encoder-decoder pair tasked with input reconstruction. Common variants enforce a latent bottleneck by limiting hidden units below input dimensionality.

As a result, the model is forced to discard low-level input and is discouraged from learning trivial solutions. To further enhance the quality of the learnt representations, further models regularize the latent space. For example, denoising autoencoders (DAE) reconstruct from noise-corrupted input to learn latent representations [9]. The variational autoencoder with discrete latents is extended by the

ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 174

vector-quantized variational autoencoder (VQ-VAE) [10], [11]. It prioritizes important features, such as phonemes, above noise by quantizing the encoder output, h_t, to a codebook vector for the decoder; non-differentiable gradients are handled using straight-through estimation. This is accomplished in the VQ-VAE by utilizing a straight through estimator [12], which assumes that the gradients are equal for the decoder input and the encoder output. To obtain the quantized version q_t of h_t, use a pre-learned codebook $A \in R^{K\times D}$, with K entries and each vector a_k of dimension D

$$q_t = a_{k,} \text{ where } k = \arg_j^{\min} \left|\left|h_t - a_j\right|\right|_2 \qquad (1)$$

The output x_t is subsequently produced by the decoder $g(\cdot)$ using q_t as input. An auxiliary loss with two components, similar to classical VQ dictionary learning, aids in codebook refinement. One term provides gradients to direct codebook vectors a_k toward non-quantized inputs h_t while a commitment term—optimized solely by the encoder—constricts these inputs near codebook entries to prevent unbounded expansion. The whole VQ-VAE loss is:

$$\mathcal{L} = \underbrace{\log\ p(x_t|q_t)}_{\text{encoder+decoder}} + \underbrace{\|sg[h_t] - A\|_2^2}_{\text{codebook}} + \beta \underbrace{\|h_t - sg[A]\|_2^2}_{\text{encoder}} \ , \quad (2)$$

Wherever, the stop-gradient operator is represented by $sg[x] = x$ and β is a scalar hyperparameter. Note that broadcasting is implicitly applied in Equation (2) via the subtraction of a vector from a matrix. These learned discrete representations prove valuable for applications likes conversion of speaker and effectively seize high-level speech data tied to voices [13]. While unique to VQ-VAE, vector quantization has been broadly adopted in SSL for pretext task targets and regularization.

2.1.2.2. Autoregressive prediction: APC is based on autoregressive text LMs, which forecast next characters based on past context [14]. The encoder $f(\cdot)$ first processes the input series $X_{[1,t]}$ to generate latent representations $H_{[1,t]}$. A decoder head, $g(\cdot)$, predicts x_{t+c},, where $c \geq 1$, from the $H_{[1,t]}$ latent vector, where c represents prediction steps. Encoder f $(\cdot)$ and decoder g $(\cdot)$ together minimize $\hat{x}_{t+c} = g(h_t)$ and ground x_{t+c}. APC formula:

$$H_{[1,t]} = f\left(X_{[1,t]}\right), \qquad (3)$$

$$\hat{x}_{t+c} = g(h_t) \ , \qquad (4)$$

$$\mathcal{L}_t = \|\hat{x}_{t+c} - x_{t+c}\|_1 . \qquad (5)$$

c = 1 when employing autoregressive text LMs.On the other hand, adjacent frames are identical due to the smoothness of auditory waves. Learning "slow" features that span several frames is frequently required for downstream activities [15]. In TIMIT, phonemes average 0.07 seconds vs 0.01-second spectrogram frames x_t, making single-frame prediction straightforward; the original APC employs (c = 3) effectively.

Generates past and future frames from previous context, extending APC to multi-target training [16]. Quantization is employed with the APC aim in VQ-APC [17] . The fact that APC only encodes data from earlier timesteps rather than the complete input is a disadvantage. By merging the bidirectionality of ELMo with the reconstructive goal of APC, DeCoAR addresses this problem and allows encoding of complete the input information [18]. It encodes x_{1+t} using a forward LSTM $f_1(\cdot)$ and $X_{[t+k,T]}$, wherever $k > 1$, using a backward LSTM $f_2(\cdot)$.

$$H_{[1,t]} = f_1\left(X_{[1,t]}\right), \qquad (6)$$

$$h\grave{}_{[t+k,T]} = f_2\left(X_{[t+k,T]}\right), \qquad (7)$$

$$\hat{x}_{[t+1,t+k-1]} = g\left(h_t, h\grave{}_{t+k}\right). \qquad (8)$$

The concatenation of h_t and $h\grave{}_t$ serves as the input feature vector for the subsequent tasks.

2.1.2.3. Masked Reconstruction: is a major source of inspiration for masked reconstruction from BERT [19] . Some elements in the given sentences are hidden during BERT pre-training by

ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 174

substituting them at random with either another input token or a learned masking token. Next, using the no masked tokens, the model learns to recreate the Some elements. Similar pretext tasks for pre-training speech encoders to learn complete contextual representations have been studied recently, similar to the DeCoAR model. From a high-level standpoint, masked reconstruction models' training phase could be expressed as

$$H = f\big(m(X)\big), \qquad (9)$$

$$\hat{x}_t = g(h_t), \qquad (10)$$

$$\mathcal{L}_t = \|\hat{x}_t - x_t\|_1 . \qquad (11)$$

A stochastic masking rule, which will be covered in more detail below, is defined by the function $m(\cdot)$.The function $f(\cdot)$ is naturally a Transformer encoder, RNNs used too [20]. A prediction head is the decoder. Lastly, to deter the model from learning an distinctiveness planning, the Loss $\mathcal{L}_t$ for hidden time steps only. Frames can be treated as tokens when using NLP masking for speech, BERT-style masking has been utilized for speech pre-training [20]. However, domain-specific features are crucial for audio data, thus simply adapting from NLP is insufficient.

• Each token is randomly and independently masked under the normal BERT masking strategy. Masking a single frame, however, is a simple reconstruction process for speech. Because audio signals are smooth, models might learn to recover the masked frame by simply interpolating nearby frames. As a result, masking segments of successive frames is typical [20].

• pMPC chooses masked speech frames affording on phonic segmentation within an sound, as opposed to masking a predetermined number of successive frames [21].

• Spectral information makes it possible to mask speech along the frequency dimension as well, even though the majority of research only mask inputs along the temporal dimension [21]. It has been demonstrated that frequency masking enhances speaker classification representations [21]. Alternatives to directly concealing the input are investigated in some investigations. Masked convolution blocks are used to introduce time masking in (NPC) [22]. Some methods, inspired by XLNet[23], Reconstruct shuffled input to align masking pre-training with fine-tuning.Masked reconstruction techniques can be further enhanced by regularization techniques. It has been demonstrated that vector quantization, which is used in DeCoAR 2.0 [24], enhances the learned representations. Additionally, the TERA model introduces two dropout regularization techniques: layer dropout and attention dropout.

2.1.2.4. More Generative Approaches: Numerous research have investigated the reconstruction of goals obtained from the input in addition to the autoregressive and masked reconstruction tasks covered above. PASE and PASE+ targets include prosody features, MFCCs, log power spectrum , and waveform [26]. Other methods use phase, gaps , and future/past spectrograms for short-speech embeddings [27].

2.2. Contrastive approaches:

2.2.1. Motivation: Speech has a lot of interconnected elements, as was previously mentioned. Therefore, finding contextualized latent variables of variation may not be best achieved by directly modeling speech. Contrastive models distinguish between positives and negatives vs. anchors, minimizing positive distances while maximizing negative ones. The fields of computer vision and machine learning have made substantial use of this strategy [28].

2.2.2. Techniques

2.2.2.1. Contrastive Predictive Coding (CPC)

CPC exemplifies contrastive learning model [29]. Anchors h_t use context up to timestep t, while positives/negatives come from localized representations z_t trained cooperatively.

$$Z = f_1(X), \qquad (12)$$

ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 174

$$h_t = f_2(Z_{[1,t]}), \quad (13)$$

$$\hat{z}_{t,k} = g_k(h_t), \quad (14)$$

g_k is a stepwise transformation, $f_1(\cdot)$ a CNN for restricted interested field in X, and $f_2(\cdot)$ restricted conditions h_t only on prior timesteps $Z_{[1,t]}$ .The distance between learned representations z_t and projections of the contextualized representation $\hat{z}_t$ is measured by the loss function. The methodology is comparable to earlier research on Noise-Contrastive Estimation (NCE) [30]. Maximize (h_t and z_t) lower bound (and thus x_t) is equivalent to the loss, called InfoNCE.

$$\mathcal{L}_{t,k} = -\log\left(\frac{\exp(\hat{z}_{t,k}^T z_{t+k})}{\sum_{n\sim D}\exp(\hat{z}_{t,k}^T z_n)}\right). \quad (15)$$

In this case, D is a set of indices that includes Negatives were sampled from the suggestion distribution (uniform over set $\{1, \ldots \ldots, T\}$ + target index $t + k$), resulting in same-sequence sampling, which has been demonstrated to perform well in phoneme classification [29]. To demonstrate that CPC uses various projection layers to target multiple offsets, the loss is denoted by k. For phoneme classification, k = 12 works efficiently .The CPC method is extended by the wav2vec model [31], which uses distinct parameterizations for the functions $f_1(\cdot)$ and $f_2(\cdot)$.Additionally, it adjusts the loss to take into account a binary forecast task, so:

$$\mathcal{L}_{t,k} = -\log(\sigma(\hat{z}_{t,k}^T z_{t+k})) - \sum_{n\sim D}\log(\sigma(-\hat{z}_{t,k}^T z_n)). \quad (16)$$

where the sigmoid function is denoted by σ.

2.2.2.2.wav2vec 2.0: Minimize contextualized-quantized distance, combining masking and contrastive learning using the InfoNCE loss [29], same like the CPC model. It is receiving a waveform as input to employs a convolutional feature encoder and a transformer network. The transformer input masked convolutional features as shown:

$$Z = f_1(X), \quad (17)$$

$$H = f_2(m(Z)), \quad (18)$$

$$q_t = g_k(z_t), \quad (19)$$

Now, $f_1(\cdot)$ is a CNN with a constrained receptive field, $f_2(\cdot)$ is a transformer encoder, $m(\cdot)$ defines a masking policy once more, and $g(\cdot)$ is a quantization module that forms discrete targets from Z using a Gumbel softmax [32] with a straight-through estimator.

The quantized vector at the same time step, q_t is selected as the positive sample in wav2vec 2.0, Negatives, on the other hand, are taken from different masked time steps. Only at masked time steps are anchors taken to be h_t.

Interestingly, while sampling negatives, quantized targets enable the exclusion of the class of the positive example q_t . The loss is:

$$\mathcal{L}_t = -\log\left(\frac{\exp(S_c(h_t, q_t))}{\sum_{n\sim D}\exp(S_c(h_t, q_t))}\right), \quad (20)$$

anywhere D includes t and the guides sampled from additional concealed time steps, and $S_c(\cdot)$ is the cosine similarity.

Wav2vec 2.0 associations two methods to learn good codebooks because the quality of the quantization determines the quality of the learnt representations. First, wav2vec 2.0 uses a technique known as Product Quantization (PQ) to concatenate quantized representations from several codebooks at each time step [33]. Additionally, an auxiliary term that promotes equitable use of all codebook entries is included to the training loss. On all Librispeech and Libri-light subsets, the wav2vec 2.0 method was

ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 174

the first to achieve singledigit WER. The wav2vec-C extends wav2vec 2.0 by adding a consistency loss term to reconstruct input features from quantized representations (like VQ-VAE) [34].

2.3. Predictive approaches

Predictive techniques generate a probability distribution across a discrete vocabulary for masked parts of an input phrase, in contrast to generative representation learning techniques. Model learns strong unmasked representations to predict masked targets, as loss applies only to masked regions. These methods are inspired by:

- The effectiveness of BERT-like models in natural language processing [19], which develop contextual representations by combining data from different time steps.
- Self-supervised visual learning using the Deep Cluster technique [35].

2.3.1. Approaches

2.3.1.1. Discrete BERT

DiscreteBERT[110] converts BERT to speech with a fixed ~13.5k-code vocabulary from contrastive vq-wav2vec [105] using LibriSpeech.

- Method: Extract discrete units from frozen vq-wav2vec, then train BERT using only masked prediction loss.
- Evaluation: Following Libri-Light policy [36], fine-tuning employs balanced LibriSpeech subsets (10 min, 1 hr, 10 hr) plus the 100 hr-clean partition, achieving 25% WER on test-other with only 10 min fine-tuning—pioneering self-supervised speech learning.

2.3.1.2. HuBERT

Unlike DiscreteBERT, the Hidden Unit BERT (HuBERT) technique [37] leverages traditional k-means clustering on MFCC characteristics to discretize continuous speech inputs.

2.3.2. Key Design

- Processes raw waveforms or log-mel features directly while maintaining all speech information.
- Predicts k-means cluster assignments for masked continuous features.

Construction: Like to wav2vec 2.0, it uses a convolutional encoder for waveform input, with masking applied before the transformer (Equations 17-18).HuBERT directly calculates the cross entropy loss between the foretell and correct clusters because targets are pre-calculated k-means cluster identities. Loss applies to both $\mathcal{L}_{\mathrm{m}}$and unmasked $\mathcal{L}_{\mathrm{u}}$ timesteps, as opposed to contrastive approaches that need negative examples to prevent minor results.

$$\mathcal{L}_{\mathrm{m}} = \sum_{\mathrm{t}\in\mathrm{M}} -\log \mathrm{p}(\mathrm{c}_{\mathrm{t}}|\,\mathrm{X})\,, \qquad (21)$$

$$\mathcal{L} = \alpha\mathcal{L}_{\mathrm{m}} + (1-\alpha)\mathcal{L}_{\mathrm{u}}\ , \qquad (22)$$

In this equation, c_{t} represents the k-means cluster of timestep t, M represents all masked timesteps, and c_{t} is calculated by summing over $\mathrm{t} \in \mathrm{M}$, similar to $\mathcal{L}_{\mathrm{m}}$.This requires the computer to learn auditory and language modeling from continuous inputs: first, mapping unmasked frames to meaningful latent representations (basic acoustic modeling), and then capturing long-range temporal relatives to decrease error of predicting. An important finding is the relevance of target consistency, rather than just correctness, which allows for a focus on input sequence structure.

HuBERT uses its representations to refine targets: k-means quantizes latent for a second iteration after initial training. Stability is ensured by alternating masked prediction and clustering; two iterations match or surpass SOTA on LibriSpeech/Libri-Light (high/ultra-low resource). By simultaneously

ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 174

optimizing wav2vec 2.0 contrastive + HuBERT veiled losses, w2v-BERT [108] achieves a WER decline of 5–10% on LibriSpeech test-clean/other and >30% on in-house search compared to baseline.

2.3.3. WavLM: Based on the HuBERT SSL foundation, it prioritizes speaker preservation and spoken material [38]. enhances self-attention with content-gated relative position bias, outperforming HuBERT/wav2vec 2.0 convolutions in recognition. Gates adjusts bias according to the substance of the speech (different roles for speech offsets and quiet). 320 bucketing for offsets; shared embeddings between layers. WavLM uses utterance mixing to handle multi-speaker jobs with less than 50% overlap from the same minibatch, with the main speaker as the target. Trained on 94k hours of audio. SOTA on SUPERB separation/verification/diarization outperforms HuBERT/wav2vec 2.0 in all ten tasks, matching ASR [76].

2.3.4. data2vec: Inspired by EMA teachers in vision SSL, predicts [39] continuous contextual latents y_t from masked inputs—generated via EMA teacher encoding unmasked inputs (top-k instance-normalized transformer blocks)[40]. Smoothed L_1 loss is used to stabilize outliers and gradients [41]:

$$\mathcal{L}(y_t, h_t) = \begin{cases} \frac{1}{2}(y_t - h_t)^2/\beta, & |y_t - h_t| \le \beta \\ |y_t - h_t| - \frac{1}{2}\beta, & \text{otherwise} \end{cases}, \qquad (23)$$

h_t Like wav2vec 2.0/HuBERT. First multimodal SSL (speech, vision, and text) with competitive results across modalities.

3.SPEECH RECOGNITION SYSTEM

Speech has been a vital tool for human communication for antiquity. Speech recognition is the process of either identifying the speaker or translating spoken words into text. Because to the growing connectivity among humans and automatic systems or computers, technology has evolved over time to become a vital and fundamental aspect of our way of life [42]. The main word recognition structure accomplished to identify records was developed at Bell Laboratory in 1952 [43]. Speaker-dependent systems, isolated/connected word recognizers, and spontaneous recognition are examples of popular speech recognition systems. Major advancements have been fueled by the long-standing desire for voice interfaces, although issues with speech's subjective nature—primarily speaker variance, background noise, and continuous flow—remain. Performance is most negatively impacted by noise, whether it be ambient (such as traffic) or speaker-related (such as coughing). The system is separated into front-end and back-end components in Figure 2.

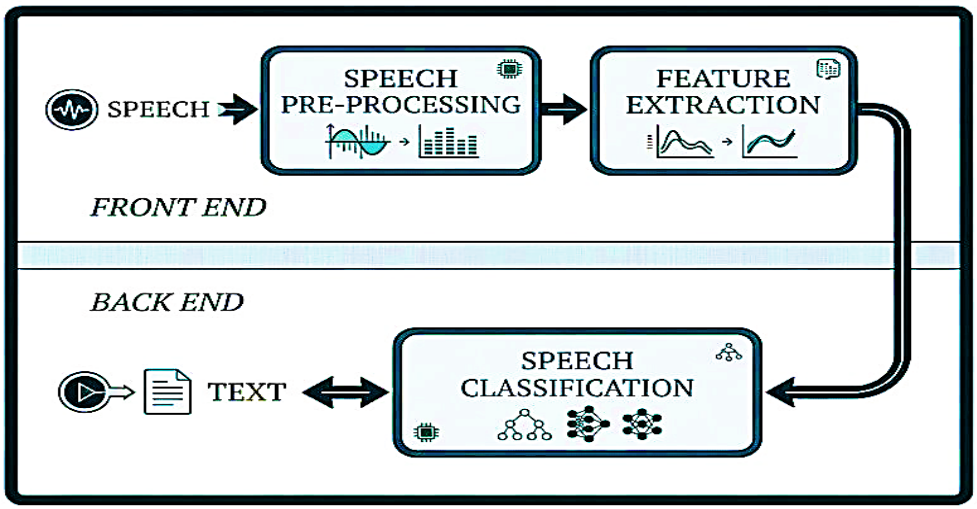


Figure 2. The Short Version of Speech Recognition Systems [44].

ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 174

3.1. Front End: Speech Pre-processing & Feature Extraction

In this stage, the raw acoustic signal is cleaned and converted into a digital format. Speech Pre-processing involves noise reduction and normalization to enhance signal quality. This is followed by Feature Extraction, where the system identifies key acoustic components (like MFCCs). Recent advancements in Self-Supervised Learning (SSL), such as the Wav2Vec 2.0 framework, have revolutionized this stage by allowing the system to learn powerful speech representations even from unlabeled data [45].

3.2. Back End: Speech Classification

This is the "intelligence" layer where the extracted features are mapped to linguistic units. While traditional systems used Hidden Markov Models (HMM), modern Speech Classification relies on End-to-End (E2E) architectures and Transformers. These models, like OpenAI Whisper, use massive-scale weak supervision to achieve human-level accuracy across various languages and noisy environments by processing the entire sequence at once [46].

3.3. Output: Text Generation

The final stage converts the classified patterns into readable text. Current high-tech systems utilize Conformer models (Convolution-augmented Transformers), which combine the benefits of CNNs for limited features and Transformers for global context, ensuring high precision in the final text output [47].

4.SELF-SUPERVISED LEARNING APPROACHES

Several datasets that provide speech labels have also been employed in the development of SSL methods, the data is divided into several categories based on pre-training datasets only, datasets used for pre-training and testing only, datasets utilized for testing only, and datasets used for evaluation only.

4.1. Datasets For pre-training

Libri-light (LL) ,SNR, speaker ID, and genre labels are attached to 60k hours of English speech in the LibriLight (LL) collection, which is derived from LibriVox audiobooks [36]. With over 2 million 10-second YouTube throughout 632 audio events, pre-training has also made use of AudioSet [48], producing 2.5k hours of multilingual audio in different quality, frequently with numerous sound sources.AVSpeech, another sizable audio-visual dataset for SSL study, AVSpeech, with 4.7k hours of multilingual video. Its 3,100-hour audio subgroup, which included clips with a visible face, single-speaker audio, and little background noise, taught audio-only representations [49].The Fisher corpus,using 1,000 hours for pre-training, VoxPopuli collects more than 2,000 hours of conversational phone speech [50]. Amazon's 10,000 hours of practical far-field English commands are an example of industrial initiatives. Didi's datasets provide 10,000 hours of spontaneous calls (Didi Callcenter) and read speech (Didi Dictation) for Chinese [51].

4.2. Datasets For pre-training and testing :

LibriSpeech (LS), 960 hours of read English speech from open-source audiobooks are available in LibriSpeech (LS) [49], a well-known labeled dataset. For training, validation, and testing, it contains subsets such as train-clean-100/360, train-other-500, dev-clean/other, and test-clean/other. (WSJ), another major labeled corpus, supports ASR assessment with subsets such as si284 (81 hours) for training, dev93/eval92 for validation/testing, and si84 (15 hours) for training. For pre-training and evaluation, researchers employ MLS (50k hours/8 European languages from LibriVox), VP (400k hours/23 languages of parliamentary discourse), and BBL (1k hours/17 Asian/African languages) [52]. **Key labeled datasets for evaluation include**: GigaSpeech [53] (33k hours multi-domain English from YouTube/podcasts/audiobooks; 10k transcribed), TED-LIUM 3/2 [54], [55] (118 hours TED talks), Switchboard [56] (260 hours two-way calls), TIMIT [57] (read American English with word/phone

ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 174

transcriptions), and VoxLingua107 [57] (6.6k hours/107 languages for LID). Beyond evaluation, these corpora improve pre-training by increasing representation generalizability.

4.3. Datasets For Evaluation:

SSL is assessed using traditional speech processing benchmarks in addition to the previously described datasets. Evaluating ASR representations efficacy, studies use DIRHA, Hub5, and CHiME-5.Hub5, Forty transcribed English phone calls are included in the Hub5 evaluation (commonly known as the NIST 2000 Hub5 test set) solely for testing.DIRHA ,Distant-speech Interaction for Robust Home Applications, or DIRHA [58], is a database made up of phonetically rich sentences, speech of instructions and keywords, and utterances selected from the Wall Street Journal. English speakers from the US and the UK read these statements, which are then captured using microphone arrays. CHiME-5,A challenge called CHiME-5 [59] includes a dataset of natural conversational speech that was gathered using microphone arrays during a dinner party with the goal of advancing robust ASR. In order to train and assess the model, Additionally, a 1,000-hour audio corpus was gathered and transcribed by Amazon Alexa researchers.

Additionally, researchers use the, INTERFACE is an expressive speech database in French, Spanish, Slovenian, and English that includes fear, joy, sadness, anger, contempt, surprise, and neutral sentiment analysis representations [60]. MOSEI (Multimodal Opinion Sentiment ^ Emotion Intensity) annotates 65 hours of YouTube videos with sentence-level sentiment using INTERFACE-like emotion categories (replacing joy with happy) [61].LS EnFr In order to train and assess English-French machine translators, It adds parallel French translations to already-existing monolingual English utterances from the LS dataset [62].CoVoST-2, a trilingual speech translation benchmark with around 2,900 hours of audio data, was developed from Common Speech. It allows translation between 21 basis languages and English and between English and 15 target languages [63]. The ALFFA project3:In order to support the advancement of speech technology in Africa, the ALFFA project3 gathers speech in four African languages: Amharic [64], Fongbe [65], Swahili, and Wolof [66]. In order to assess SSL's cross-linguistic generalizability, the authors of the same article [67] further choose 21 phonetically varied languages from OpenSLR.

4.4. Evaluating of Self-Supervised Learning

One widely adopted approach in the ground of self-supervised learning (SSL) for assessing the transferability and generalization of pre-trained models involves fine-tuning these models on labeled downriver tasks. This paradigm allows researchers to rigorously evaluate how well representations learned from unlabeled pre-training data adapt to specific supervised applications. The specific evaluation protocols vary based on the choice of pre-training and fine-tuning datasets, typically focusing on one or more of the following benchmarks: (1) cross-dataset knowledge transfer, where models pre-trained on one corpus (e.g., large-scale unlabeled audio data) are fine-tuned on entirely different downstream datasets to test domain generalization; (2) few-shot learning performance, which measures efficacy when fine-tuning uses only a small subset of labeled examples from the target task; or (3) superiority relative to fully supervised baselines, achieved by comparing SSL models against systems trained from scratch using the complete available labeled training data.

Table 1 delivers a complete abstract of the experimental setups commonly reported in the SSL literature. This includes details on the pre-training corpora (e.g., scale, diversity, and domain of unlabeled data), the downstream tasks and corresponding datasets (e.g. automatic speech recognition on LibriSpeech or emotion recognition on multimodal sets), and the quantity of labeled data utilized during fine-tuning which directly aligns with the targeted benchmarking scenario (cross-dataset, few-shot, or supervised baseline). Importantly, fine-tuning pre-trained networks offers flexibility across multiple strategies: for instance, updating all layers end-to-end for maximal adaptation; freezing early layers (e.g., convolutional feature extractors) to retain low-level representations while adapting higher layers; or appending task-specific prediction heads, such as linear classifiers or more complex transformer-based decoders, atop the frozen backbone.

TABLE 1. An overview of typical experiment configurations for different SSL assessments.

ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 174

Work	Pre-training corpus	Dataset		Fine-tuning labels
		Training Fine-tuning	Test	
vq-wav2vec + DiscreteBERT	LS 960 hrs	LS 100 hrs	LS test-clean, LS test-other	10 mins, 1/10/100 hrs
BMR	WSJ si284, LS 960 hrs	WSJ si284	WSJ eval92	81 hrs
wav2vec-c	Alexa-10000	Alexa-eval	Alexa-eval	1000 hrs
MPE	WSJ si284, LS 960 hrs	WSJ si284	WSJ eval92	25/40/81 hrs
		LS 100/360/960 hrs	LS test-clean	100/360/960 hrs
HuBERT	LS 960 hrs, LL 60k hrs	LS 960 hrs	LS test-clean, LS test-other	10 mins, 1/10/100/960 hrs
wav2vec 2.0	LS 960 hrs, LL 60k hrs	LS 960 hrs	LS test-clean, LS test-other	10 mins, 1/10/100/960 hrs
MPC	Didi Callcenter, Didi Dictation, Open Mandarin	HKUST	HKUST	168 hrs
		AISHELL-1	AISHELL-1	178 hrs
	SWB, Fisher 1k, LS 960 hrs	SWB	Hub5'00	260 hrs
DeCoAR 2.0	LS 960 hrs	LS 100 hrs	LS test-clean, LS test-other	1/10/100 hrs

frequently utilized pre-training datasets. Pre-training on large datasets has become a key industry focus, exemplified by wav2vec 2.0 [68] leveraging LibriLight (60k hours), modified CPC [69] utilizing LL (60k hours), HuBERT [37], Bidir-CPC[67] trained on CPC-8k (8k hours), MPC [51] employing Didi's internal corpora (10k hours), and wav2vecc utilizing Alexa's internal datasets (10k hours). As the amount of computer power available increases, we anticipate that this tendency will continue. While Chinese [51] and multilingualism [67] are also receiving attention, the majority of research focuses on learning representations for English. The diversity of fine-tuning datasets is higher than pre-training corpora. The number of labeled instances utilized for fine-tuning ranges from a few minutes to thousands of hours, supporting benchmarking scenarios from highly resource-constrained environments to fully supervised settings.

5. CRITICAL SPEECH APPLICATIONS BEYOND ASR

Aside from ASR, critical speech applications include emotion detection utilizing vocal cues, spoken language understanding for direct intent extraction from audio, voice conversion for timbre alteration, and paralinguistic speaker identification using attributes such as age and gender [70]. These facilitate advanced interactions in mental health, assistants, privacy, and multi speaker settings, with scholarly papers outperforming in low-resource and noisy environments [71].

ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 174

6. EXPERIMENTS RESULTS

This survey covers many experimental scenarios since self-supervised learning (SSL) algorithms are evaluated using a variety of datasets and downstream tasks. We first concentrate on automatic speech recognition (ASR) on LibriSpeech (LS) and more due to their extensive use. The table below provides an overview of word error rates (WER) on the test-clean split, where models were pre-trained unsupervised on unlabeled speech and subsequently refined supervised on labeled data.

Table 2. Overview For Pre-Train and Fine-Tune Datasets

Model	Dataset	Pre-train	Fine-tune	WER test-clean/other
wav2vec 2.0[72]	LibriSpeech	960h LS	100h	3.5% / 7.2%
HuBERT	LibriSpeech	960h LS	100h	3.0% / 6.5%
WavLM[73]	LibriSpeech	960h LS	100h	2.8% / 6.0%
Data2Vec2.0 [72]	LibriSpeech	960h LS	100h	3.2% / 6.8%
wav2vec 2.0[73]	Common Voice	60k hrs	10h	15.2% / -
WavLM	TED-LIUM 3	960h LS	100h	8.1% / -
HuBERT[74]	VoxPopuli (EN)	XLS-R 100k	10h	12.5% / -
Data2Vec	Switchboard	960h LS	100h	4.4% / -

The above table illustrates low WER scores across several datasets on SSL for speech recognition , emphasizing the superiority of models such as WavLM and HuBERT in low-resource conditions following significant initial training [75].

Analysis of results by model , Wav2vec 2.0 achieves a 3.5% WER score on LibriSpeech test-clean with 100 hours of fine-tuning, but rises to 15.2% on Common Voice due to linguistic diversity; this reflects its strength in pure English with a need for cross-lingual enhancements. HuBERT outperforms LS by 3.0% and VoxPopuli by 12.5%, thanks to its intermediate layers that enhance general representations. WavLM stands out as the best, with a 2.8% improvement over LS and 8.1% over TED-LIUM, handling noise and domain transitions 10-15% better than its competitors. Data2Vec 2.0 comes close, with a 3.2% improvement over LS and 4.4% over Switchboard, benefiting from integration with vision technologies [72].

These results verify that, when compared to self-training, SSL with models with more than one billion operators lowers WER by 20–50% in low-resource contexts, especially across more than five datasets including LS, Common Voice, and VoxPopuli. This facilitates the development of low-resource languages. For best results, studies suggest ensemble training or integration, with a focus on initial training of up to 60,000–100,000 hours for effective generalization [73].

7. DISCUSSION, LIMITATIONS, AND FUTURE WORKS

In this survey, talk about every experiment setting. For almost 80 years, speech recognition technology has been developed, and for nearly as long, it has been considered as an alternate access option for people with impairments. Because of their widespread use in SSL trials, we concentrate on ASR on the LS dataset to comprehend SSL's effectiveness. The ASR models were initially acquired by pre-training a model with each SSL approach using unlabeled speech. After that, a supervised training purpose was used to refine the model using labeled data.

These results are especially useful for low-resource circumstances, including extending systems to original languages where a large amount of unlabeled audio is accessible, as gathering labels for novel scenarios is frequently excessively time-consuming or expensive. SSL models perform exceptionally well in tasks other than ASR; for example, merging SSL pre-training with self-training yields state-of-the-art results. We anticipate that pre-trained SSL models will produce state-of-the-art performance on an collective number of tasks as SSL study improvements more consideration. Notably, DeCoAR 2.0 employs a 4-gram model trained on LibriSpeech (LS), while HuBERT-L and wav2vec 2.0-L use

ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 174

Transformers as their language model for ASR. The SSL community can train, assess, and comparation speech representations for a variety of downstream tasks, including acoustics, speaker identity, paralinguistics, and semantics, thanks to the SUPERB benchmark [76].

Despite the exceptional performance of self supervisor learning (SSL) models in the automatic speech recognition (ASR) challenge, this work has several significant **limitations** that should be considered:
First: Large and non-public pre-training datasets: such as LibriLight (60k hours) and Alexa (10k hours),
limiting redundancy.
Second: Limited fine-tuning diversity :Focus on LibriSpeech /CoVoST-2 without Arabic dialects or real
noise.
Third: Incomplete assessment: Lack of few shots in noisy environments and reliance solely on WER.
Forth: Narrow scope: No multimodal or low-resource languages.
Can pinpoint a number of intriguing avenues for future studies. By methodically reviewing and addressing the current constraints and open concerns, may identify critical areas for progress and provide clear recommendations to lead future research efforts:
First: releasing pre-trained general checkpoints.
Second: extending the evaluation to low-resource Arabic dialects using datasets such as MGB-3.
Third: incorporating data augmentation techniques for noise.
Forth: evaluating end-to-end speech-to-text translation with Hugging Face integrations. These improvements will enhance generalizability and practical applications.

8. CONCLUSIONS

This article provides a detailed introduction to speech recognition technologies. It then investigates the integration of (SSL) and (ASR), emphasizing how SSL improves model performance by pre-training on unlabeled data. Furthermore, the research includes a comparative analysis of significant datasets typically used in voice recognition systems, providing insights that can inform the construction of unique or enhanced architectures capable of outperforming existing systems in relationships of accuracy and efficiency. Critically discusses common obstacles and constraints, and concludes with an evaluation of possible future trends guiding the field's development.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015

[2] J. Ma, S. Matsoukas, O. Kimball, and R. Schwartz, “Unsupervised training on large amounts of broadcast news data,” in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2006

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

[4] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” Science, vol. 349, no. 6245, pp. 255–260, 2015

[5] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” Science, vol. 313, no. 5786, pp. 504–507, 2006

[6] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang,“Self-supervised learning: Generative or contrastive,” IEEE Transactions on Knowledge & Data Engineering, no. 01, pp. 1–1, Jun 2021.

[7] P. Xia, S. Wu, and B. Van Durme, “Which *BERT? A survey organizing contextualized encoders,” in Proceedings of the 2020Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, Nov. 2020, pp. 7516–7533. [Online]. Available: https://aclanthology.org/ 2020.emnlp main.608.

ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 174

[8] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, "Deep representation learning in speech processing: Challenges, recent advances, and future trends," 2021.

[9] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol,"Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," Journal of Machine Learning Research, 2010

[10] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discreterepresentation learning," 2017.

[11] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings, 2014

[12] Y. Bengio, N. Leonard, and A. Courville, "Estimating or propagating ´ gradients through stochastic neurons for conditional computation," arXiv preprint arXiv:1308.3432, 2013

[13] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using WaveNet autoencoders," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 12, pp. 2041–2053, 2019.

[14] Y.-A. Chung and J. Glass, "Generative pre-training for speech with autoregressive predictive coding," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2020

[15] L. Wiskott and T. J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," Neural Computation, vol. 14, no. 4, pp. 715– 770, 2002

[16] Y.-A. Chung and J. Glass, "Improved speech representations with multi-target autoregressive predictive coding," 2020.

[17] Y.-A. Chung, H. Tang, and J. Glass, "Vector-Quantized Autoregressive Predictive Coding," in Proceedings of the Annual Conference of the International Speech Communication Association, 2020.

[18] S. Ling, Y. Liu, J. Salazar, and K. Kirchhoff, "Deep contextualized acoustic representations for semi-supervised speech recognition," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2020.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional Transformers for language understanding," in NAACL, 2019

[20] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional Transformer encoders," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2020.

[21] A. T. Liu, S.-W. Li, and H.-y. Lee, "TERA: Self-supervised learning of Transformer encoder representation for speech," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 2351– 2366, 2021

[22] A. H. Liu, Y.-A. Chung, and J. Glass, "Non-Autoregressive Predictive Coding for Learning Speech Representations from Local Dependencies," in Proceedings of the Annual Confe

[23] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in Proceedings of Advances in Neural Information Processing Systems, H. Wallach, H. Larochelle, A. Beygelzimer,F. d'Alche-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran ´ Associates, Inc., 2019

[24] S. Ling and Y. Liu, "DeCoAR 2.0: Deep contextualized acoustic representations with vector quantization," arXiv preprint arXiv:2012.06659, 2020.

ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 174

[25] J. Luo, J. Wang, N. Cheng, and J. Xiao, "Dropout regularization for self-supervised learning of Transformer encoder speech representation," Proceedings of the Annual Conference of the International Speech Communication Association, 2021

[26] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, "Multi- task self-supervised learning for robust speech recognition," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2020.

[27] M. Tagliasacchi, B. Gfeller, F. de Chaumont Quitry, and D. Roblek, "Pre-training audio representations with self-supervision," IEEE Signal Processing Letters, vol. 27, pp. 600–604, 2020.

[28] M. Schultz and T. Joachims, "Learning a distance metric from relative comparisons," in Advances in Neural Information Processing Systems, S. Thrun, L. Saul, and B. Scholkopf, Eds., 2003.

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

[30] M. Gutmann and A. Hyvarinen, "Noise-contrastive estimation: A new ¨ estimation principle for unnormalized statistical models," International Conference on Artificial Intelligence and Statistics (AISTATS), 2010.

[31] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," arXiv preprint arXiv:1904.05862, 2019.

[32] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel-softmax," in Proceedings of International Conference on Learning Representations, 2017

[33] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest ´ neighbor search," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 1, pp. 117–128, 2011.

[34] A. Razavi, A. van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with VQ-VAE-2," in Proceedings of Advances in Neural Information Processing Systems, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alche-Buc, E. Fox, and R. Garnett, Eds., vol. 32. ´ Curran Associates, Inc., 2019

[35] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in Proceedings of the European Conference on Computer Vision (ECCV), September 2018

[36] J. Kahn et al., "Libri-light: A benchmark for ASR with limited or no supervision," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2020

[37] 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," arXiv preprint arXiv:2106.07447, 2021.

[38] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, and F. Wei, "WavLM: Large-scale self- supervised pre-training for full stack speech processing," 2021

[39] A. Baevski, W. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," 2022.

[40] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," CoRR, 2016.

[41] Abbas, Syed Mazhar, and Shailendra Narayan Singh. "Region-based object detection and classification using faster R-CNN." 2018 4th International Conference on Computational Intelligence & Communication Technology (CICT). IEEE, 2018.

[42] Apte, Shaila D. Speech and audio processing. New York: Wiley, 2012.

[43] Jacob Benesty, M. Mohan Sondhi, Yiteng Huang, "Springer Handbook of Speech Processing", Springer.

[44] L.R. Rabiner and R.W. Schafer, "Digital Processing of Speech Signals", Prentice Hall Signal Processing Series.

[45] Chou, Ju-Chieh, et al. "Toward joint language modeling for speech units and text." Findings of the Association for Computational Linguistics: EMNLP 2023. 2023.

[46] Lee, Bradford J. "Exploring the potential of AI for pragmatics instruction." Technology in Language Teaching & Learning 6.3 (2024): 1521-1521.

[47] Ashwini, B., Sheffali Gulati, and Jainendra Shukla. "Artificial Intelligence Driven Predictive Analysis of Acoustic and Linguistic Behaviors for ASD Identification." IEEE Transactions on Artificial Intelligence 5.11 (2024): 5709-5719.

[48] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C.Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and humanlabeled dataset for audio events," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2017

[49] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2015

[50] D. Jiang, W. Li, R. Zhang, M. Cao, N. Luo, Y. Han, W. Zou, K. Han, and X. Li, "A further study of unsupervised pretraining for Transformer based speech recognition," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2021

[51] D. Jiang, X. Lei, W. Li, N. Luo, Y. Hu, W. Zou, and X. Li, "Improving Transformer-based speech recognition using unsupervised pre-training," arXiv preprint arXiv:1910.09932, 2019.

[52] D. B. Paul and J. Baker, "The design for the Wall Street Journalbased CSR corpus," in Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23–26, 1992

[53] G. Chen, S. Chai, G.-B. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Z. You, and Z. Yan, "GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio," in Proceedings of the Annual Conference of the International Speech Communication Association, 2021.

[54] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Esteve, "TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation," in International Conference on Speech and Computer. Springer, 2018, pp. 198–208.

[55] A. Rousseau, P. Deleglise, and Y. Esteve, "TED-LIUM: An automatic ´ speech recognition dedicated corpus," in Proceedings of International Conference on Language Resources and Evaluation, 2012, pp. 125– 129.

[56] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 1992

[57] J. Valk and T. Alumae, "VoxLingua107: A dataset for spoken language ¨ recognition," in Proceedings of IEEE Spoken Language Technology Workshop, 2021.

[58] M. Ravanelli, L. Cristoforetti, R. Gretter, M. Pellin, A. Sosi, and M. Omologo, "The DIRHA-English corpus and related tasks for distant-speech recognition in domestic environments," in Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, 2015.

[59] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines," in Proceedings of the Annual Conference of the International Speech Communication Association, 2018.

[60] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus:English multi-speaker corpus for CSTR voice cloning toolkit," 2016.

[61] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency,"Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," 2018.

[62] Laurent, Antoine, et al. "ON-TRAC consortium systems for the IWSLT 2023 dialectal and low-resource speech translation tasks." Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023). 2023.

[63] Ait, Adem, Javier Luis Cánovas Izquierdo, and Jordi Cabot. "On the suitability of hugging face hub for empirical studies." Empirical Software Engineering 30.2 (2025): 57.

[64] F. A. A. Laleye, L. Besacier, E. C. Ezin, and C. Motamed, "Firstautomatic Fongbe continuous speech recognition system: Development of acoustic models and language models," in FedCSIS. IEEE, 2016, pp. 477–482.

[65] H. Gelas, L. Besacier, and F. Pellegrino, "Developments of Swahiliresources for an automatic speech recognition system," in Spoken Language Technologies for Under-Resourced Languages, 2012.

[66] E. Gauthier, L. Besacier, S. Voisin, M. Melese, and U. P. Elingui,"Collecting resources in sub-Saharan African languages for automatic speech recognition: A case study of Wolof," in LREC 2016, 2016.

[67] K. Kawakami, L. Wang, C. Dyer, P. Blunsom, and A. van den Oord, "Learning robust and multilingual speech representations," in EMNLP, 2020

[68] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0:A framework for self-supervised learning of speech representations," Proceedings of Advances in Neural Information Processing Systems, vol. 33, 2020.

[69] M. Riviere, A. Joulin, P.-E. Mazare, and E. Dupoux, "Unsupervised ´ pretraining transfers well acros languages," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2020

[70] Rezapour Mashhadi, Mohammad Mahdi, and Kofi Osei-Bonsu. "Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest." PloS one 18.11 (2023):.

[71] Schuller, B. & Batliner, A. Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing 1st edn. (Wiley Publishing, 2013.

[72] Chiu, Sheng-Chieh, et al. "Learnable Layer Selection and Model Fusion for Speech Self-Supervised Learning Models." Interspeech. 2024.

[73] Mdhaffar, Salima, et al. "Performance analysis of speech encoders for low-resource slu and asr in tunisian dialect." Proceedings of The Second Arabic Natural Language Processing Conference. 2024.

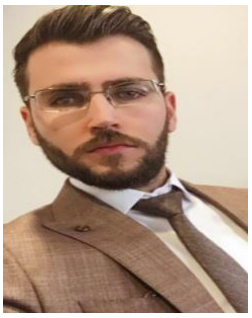
[74] Naini, Abinay Reddy, et al. "Generalization of self-supervised learning-based representations for cross-domain speech emotion recognition." ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024.

ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 174

[75] Arunkumar, A., Vrunda N. Sukhadia, and Srinivasan Umesh. "Investigation of ensemble features of self-supervised pretrained models for automatic speech recognition." arXiv preprint arXiv:2206.05518 (2022).

[76] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y.Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in Proceedings of the Annual Conference of the International Speech Communication Association, 2021.

BIOGRAPHIES OF AUTHORS



Humam Khalid Jameel holds a Master of Science in Computer Science from the Department of Computer Science, College of Science, Al-Nahrain University, Baghdad, Iraq, in 2021. He received his B.Sc. in Computer Science from the College of Computer Science- Al Anbar University, Anbar, Iraq, in 2016. He is currently a assistant lecturer at the Department of Computer Science, College of Science, Al-Nahrain University, Jadriya, Baghdad, Iraq. His research interests include computer science, Multimedia, AI, and related fields. He can be contacted at the email address: humam.khalid@nahrainuniv.edu.iq.

Assad H. Thary Al-Ghrairi holds a Master of Computer Science from College of Science/Al-Nahrain University, Baghdad, Iraq in 2016. He also received his B.Sc. (Computer Science) from Al-Nahrain University, Baghdad, Iraq in 2014. He is currently works as a lecturer at Computer Science Department in College of Science, Al-Nahrain University, Iraq. He is also a lecturer of artificial intelligence and image processing. His research includes pattern recognition, image processing, artificial intelligence, multimedia, robot path planning, control and arduino programming, remote sensing and satellite image classification, bioinformatics, he can be contacted at email: assad.h.thary@ nahrainuniv.edu.iq.

Mohammed M. Neamah holds a PhD. In Physics Science from College of Science/Al-Nahrain University, Baghdad, Iraq, in 2025. He also received a Master of Physics from the College of Science at Al-Nahrain University, Baghdad, Iraq, in 2017. He also received his B.Sc. (Physics Science) from Al-Nahrain University, Baghdad, Iraq, in 2014. He is currently a lecturer at the Physics Department, Education College, Mustansiriyah University, Iraq. He is also a lecturer on artificial intelligence and image processing. His research includes pattern recognition, image processing, artificial intelligence, multimedia, robot path planning, control, Arduino programming, remote sensing, satellite image classification, and bioinformatics. He can be contacted at the email address: mohammed.m.neamah@uomustansiriyah.edu.iq.