A Comprehensive Review of Speech Signal Processing, Voice Activity Detection, and Automatic Speech Recognition: Techniques, Challenges, and Future Directions

(IJCSNT) International Journal of Communication Systems and Network Technologies
ISSN Number: 2053-6283
Volume 14, Issue No. 03, 193–212

https://journals.mirlabs.in/index.php/ijcsnt
DOI-10.18486/ijcsnt/14.3.017
IJCSNT

Sapna Rajput[1], Stuti Pandey[2], Tanushka Mishra[3]

Abstract

Speech signal processing has become one of the most critical enablers of modern human–computer interaction, underpinning technologies ranging from voice assistants and telephony to clinical diagnostics and security systems. This paper presents a comprehensive and systematic review of the field, spanning three tightly coupled domains: (i) fundamental speech signal analysis and preprocessing, (ii) voice activity detection (VAD), and (iii) automatic speech recognition (ASR). We begin by examining the physical and mathematical foundations of speech production and perception, followed by a detailed treatment of classical and contemporary feature extraction methods, including Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP), filter-bank features, and learned representations produced by deep neural networks. For VAD, we critically compare energy-based, statistical, model-driven, and end-to-end deep learning approaches with respect to accuracy, latency, and robustness under diverse noise conditions. In the ASR domain, we trace the evolution from hidden Markov model (HMM) systems through hybrid HMM–deep-neural-network (DNN) architectures to fully neural sequence-to-sequence models, including connectionist temporal classification (CTC) and attention-based encoder–decoder frameworks. Benchmark performance metrics across standardised datasets (TIMIT, LibriSpeech, CHiME, VoxCeleb) are consolidated and discussed. We further identify persistent open challenges environmental noise, speaker variability, low-resource languages, and real-time edge deployment and outline promising research directions, including self-supervised pre-training, multimodal fusion, federated learning, and neuromorphic processing. The survey is intended to serve as a unified reference for researchers and practitioners working at the intersection of signal processing, machine learning, and human–computer interaction.

Keywords

Speech Signal Processing, Voice Activity Detection, Automatic Speech Recognition, MFCC, LPC, Hidden Markov Models, Deep Learning, LSTM, Transformer, End-to-End ASR, Noise Robustness, Feature Extraction, Self-supervised Learning

Received: 18 August 2025; **Revised:** 27 October 2025; **Accepted:** 21 November 2025;
Published: 31 December 2025;

[1, 2, 3]Department of Electronics and Communication Engineering, Institute of Technology and Management, Gwalior, Madhya Pradesh, India.

Corresponding author:
Email: sapnarajput.ec@itmgoi.in

I. Introduction

Speech is the most natural and efficient medium of human communication. The capacity to process, understand, and generate speech automatically has been a long-standing aspiration of computer science and electrical engineering, with roots traceable to the vocoder experiments of Homer Dudley at Bell Labs in 1939 [1] and the first digit-recognition system, AUDREY, developed in 1952 [2]. Today, the global market for voice and speech recognition exceeds $20 billion annually, with compound annual growth rates above 17%, driven by the proliferation of smart devices, cloud services, and ambient computing paradigms [3].

Speech signal processing encompasses three broad activities: *analysis*—extracting information about the content, speaker, emotion, or acoustic environment from a raw waveform; *transformation*—modifying the waveform to improve quality, suppress noise, or change vocal characteristics; and *synthesis*—generating intelligible and natural-sounding speech from symbolic or parametric inputs. Each activity relies on a rich theoretical apparatus spanning linear systems theory, stochastic processes, information theory, and, increasingly, deep learning.

A key enabler of efficient speech processing systems is Voice Activity Detection (VAD), also known as speech/non-speech discrimination or endpoint detection. By identifying temporal segments that contain phonetically meaningful signal and discarding silence, background noise, and non-speech sounds, VAD reduces the computational burden on downstream modules, preserves bandwidth in voice-over-IP (VoIP) communications, and improves the accuracy of both speaker verification and ASR systems. The challenge of VAD has evolved considerably: whereas early systems operated in office environments with near-stationary noise, modern deployments must cope with highly non-stationary backgrounds—traffic, cafeteria babble, wind, music—often at low signal-to-noise ratios (SNR) of 0–5 dB.

Automatic speech recognition (ASR) has undergone a revolution over the past decade. Systems built on Gaussian mixture models (GMM) and HMMs dominated the field from the 1980s through to approximately 2010, achieving word-error rates (WER) of 20–30% on conversational telephone speech. The introduction of deep neural network acoustic models, first reported by Hinton et al. [26], reduced error rates by 20–30% relative and displaced GMM-HMM as the de facto standard almost overnight. Subsequent advances—convolutional feature extraction, recurrent sequence modelling, attention mechanisms, and large-scale self-supervised pre-training [34,36]—have pushed WER on clean read speech below 2%, approaching human parity.

Motivation and Scope

Despite this rapid progress, no single survey simultaneously covers the continuum from low-level signal fundamentals through VAD to end-to-end ASR with consistent notation and comparative analysis. This paper aims to fill that gap. Our contributions are:

- A unified mathematical treatment of speech production, perception, and feature extraction (Sections II–IV).
- A taxonomy and critical comparison of VAD methodologies, with results on standardised noisy-speech benchmarks (Section V).
- A structured survey of ASR architectures from GMM-HMM to Whisper-scale transformer models, with performance tables (Section VI).
- Identification of open research problems and emerging directions (Sections X–XI).

Paper Organisation

The remainder of this paper is structured as follows. Section II reviews speech signal fundamentals. Section III covers preprocessing and noise reduction. Section IV describes feature extraction techniques. Section V surveys VAD methods. Section VI reviews ASR architectures. Section VII lists benchmark datasets and evaluation metrics. Section VIII presents comparative results. Section X discusses open challenges. Section XI outlines future directions. Section XII concludes the paper.

II. Speech Signal Fundamentals

A. The Speech Production Mechanism

Speech is produced by a cascade of physiological events. Air expelled from the lungs creates a flow that is modulated by the vocal folds (glottis), producing a quasi-periodic excitation for voiced sounds or turbulent noise for unvoiced fricatives. The resulting source signal propagates through the vocal tract—a resonant tube whose cross-sectional area varies with the position of the tongue, lips, velum, and jaw—and is radiated at the lips. This mechanism is captured by the *source–filter model*[4], in which the speech spectrum $S(z)$ is modelled as the product of a source spectrum $G(z)$, a vocal-tract transfer function $H(z)$, and a lip-radiation characteristic $R(z)$:

$$S(z) = G(z)\,H(z)\,R(z). \tag{1}$$

The all-pole approximation $H(z) \approx \dfrac{1}{A(z)}$ where $A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}$ underlies linear predictive coding and motivates many feature extraction schemes.

B. Classification of Speech Sounds

Speech sounds can be classified along several acoustic–phonetic dimensions:

- **Voiced vs. unvoiced:** Vowels and voiced consonants (/b/, /d/, /g/) exhibit periodic waveforms with a fundamental frequency $F0$ (pitch) in the range 80–300 Hz. Unvoiced fricatives (/s/, /f/) resemble broadband noise.
- **Sonorant vs. obstruent:** Sonorants (nasals, liquids) have smooth spectral envelopes; obstruents (stops, affricates) involve abrupt onsets and energy bursts.
- **Silence and background:** Segments containing no intended phonetic content but potentially non-zero energy due to reverberation or noise.

The spectral envelope of voiced speech is characterised by resonance peaks called *formants* ($F1$, $F2$, $F3$, …). The first two formants ($F1$: 250–900 Hz, $F2$: 700–2500 Hz) carry most vowel identity information and are the primary targets of many feature extraction algorithms.

C. Waveform and Time-Domain Representations

A speech waveform $x(n)$, sampled at f_s Hz (commonly 8 kHz for telephony, 16 kHz for wideband, 44.1 kHz for high-fidelity audio), is processed in short overlapping *frames* of duration $T_w \approx 20$–30 ms with frame shift $T_s \approx 10$ ms. This short-time analysis is motivated by the quasi-stationarity of speech over intervals shorter than $\sim$50 ms.

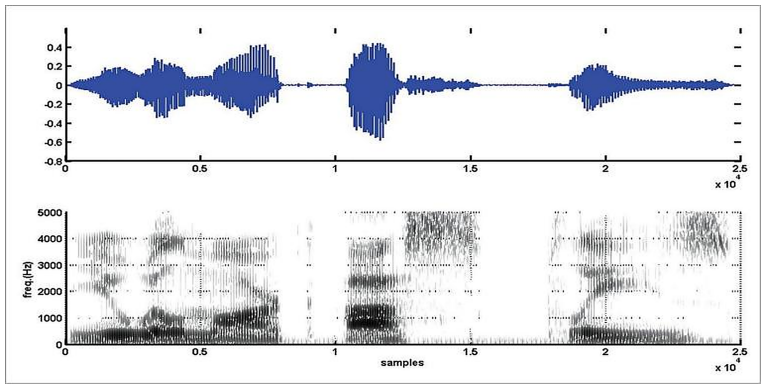

Figure 1. Time-domain waveform of a speech utterance illustrating voiced segments (quasi-periodic regions), unvoiced segments (noise-like bursts), and silence.

Key time-domain descriptors include the **short-time energy** (STE):

$$E_n = \sum_{m=0}^{N-1} \big[x(n-m)\, w(m)\big]^2 \tag{2}$$

and the **zero-crossing rate** (ZCR):

$$\mathrm{ZCR}_n = \frac{1}{2(N-1)} \sum_{m=1}^{N-1} \Big|\mathrm{sgn}[x(n-m)] - \mathrm{sgn}[x(n-m-1)]\Big| \tag{3}$$

where $w(m)$ is a Hamming or rectangular window and N is the frame length in samples. Silence exhibits low STE and variable ZCR; voiced speech exhibits high STE and low ZCR; unvoiced speech exhibits moderate STE and high ZCR [7].

D. Frequency-Domain and Time-Frequency Representations

The Short-Time Fourier Transform (STFT) of a windowed frame is

$$X(n,\omega) = \sum_{m=-\infty}^{\infty} x(m)\, w(n-m)\, e^{-j\omega m} \tag{4}$$

and its squared magnitude $|X(n,\omega)|^2$ forms the *spectrogram*.

The spectrogram reveals the joint time-frequency structure essential for phoneme identification. Alternative representations—the Mel spectrogram, the constant-Q transform, and the gammatone filterbank—warp the frequency axis according to perceptual models of the human auditory system, improving robustness to irrelevant spectral detail [13].

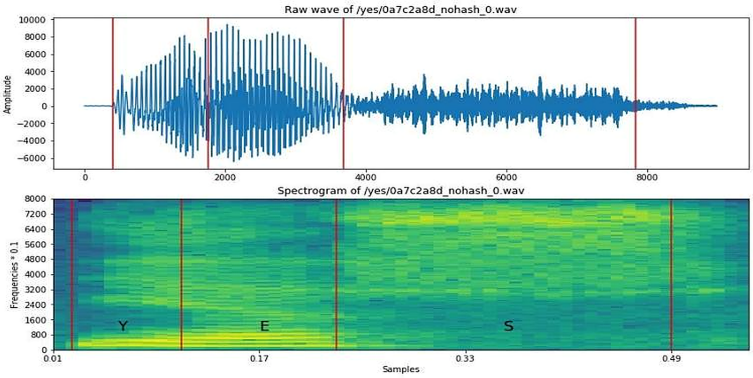

Figure 2. Wide-band spectrogram of a speech utterance. Dark bands correspond to formant trajectories; vertical striations in voiced regions correspond to individual glottal pulses.

III. Preprocessing and Noise Reduction

Before feature extraction, raw speech is typically subjected to a preprocessing pipeline that normalises level, removes DC offset, suppresses noise, and segments the signal into frames suitable for analysis.

A. Pre-emphasis

Pre-emphasis is applied to compensate for the -6 dB/octave spectral roll-off introduced by lip radiation. A first-order high-pass filter

$$\hat{x}(n) = x(n) - \alpha\, x(n-1), \qquad \alpha \in [0.95,\, 0.99] \tag{5}$$

boosts high-frequency components, improving the conditioning of LPC analysis and the numerical stability of FFT-based features[5].

B. Windowing

To reduce spectral leakage, each frame is multiplied by a smooth window function. The Hamming window,

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \qquad 0 \le n \le N-1, \tag{6}$$

is most commonly employed in speech processing due to its favourable main-lobe width and side-lobe attenuation[6].

C. Spectral Noise Reduction

A wide range of noise reduction techniques have been developed:
Spectral Subtraction[8]*:* Noise power $|\hat{D}(\omega)|^2$ is estimated during silence frames and subtracted from the noisy spectrum:

$$|\hat{S}(\omega)|^2 = \max\Big(|Y(\omega)|^2 - |\hat{D}(\omega)|^2,\ \epsilon\Big) \tag{7}$$

where Y is the observed noisy signal and ϵ prevents negative power estimates. While effective for stationary noise, spectral subtraction produces "musical noise" artefacts for non-stationary backgrounds.
Wiener Filtering: The Wiener filter minimises the mean-squared error between the estimated and true clean speech, yielding

$$H(\omega) = \frac{P_S(\omega)}{P_S(\omega) + P_N(\omega)} \tag{8}$$

where P_S and P_N are the speech and noise power spectral densities.
Deep Learning-Based Enhancement: Recent neural enhancement models, including FullSubNet[39] and SEGAN[40], operate end-to-end in the waveform or complex spectral domain and achieve substantial improvements in Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI) scores, particularly for non-stationary noise environments where classical methods struggle.

IV. Feature Extraction Techniques

Feature extraction maps the high-dimensional raw waveform or spectrum into a compact, discriminative, and relatively noise-invariant representation. This section reviews the most important techniques in roughly chronological order of development.

A. Linear Predictive Coding (LPC)

Developed by Atal and Hanauer[9], LPC models each speech sample as a linear combination of p previous samples:

$$x(n) = \sum_{k=1}^{p} a_k \, x(n-k) + G u(n) \tag{9}$$

where G is the gain and $u(n)$ is the excitation. The predictor coefficients $\{a_k\}$ are estimated by minimising the squared prediction error over a frame, which reduces to solving the Yule-Walker equations via the Levinson-Durbin recursion. The all-pole spectral envelope provides a compact 10–16 parameter representation of the vocal tract. LPC coefficients are sensitive to quantisation and frame-to-frame perturbations; derived representations such as LPC Cepstrum (LPCC) and Line Spectral Pairs (LSP) offer better numerical stability and interpolation properties.

B. Mel-Frequency Cepstral Coefficients (MFCC)

MFCCs, introduced by Davis and Mermelstein[10], remain the most widely used speech features. The computation proceeds in five stages:

1. **Framing and windowing** (Section III).
2. **FFT:** Compute $|X_n(\omega)|^2$ for each frame.
3. **Mel filterbank:** Apply M (typically 26–40) triangular bandpass filters whose centre frequencies are uniformly spaced on the Mel scale:

$$f_{\text{mel}} = 2595 \log_{10}\left(1 + \frac{f}{700}\right). \tag{10}$$

4. **Log compression:** Take the logarithm of each filter output to approximate the non-linear loudness perception of the auditory system.
5. **Discrete Cosine Transform (DCT):**

$$c_n = \sum_{k=1}^{M} \log S_k \cos\left[n\left(k - \tfrac{1}{2}\right)\frac{\pi}{M}\right], \qquad n = 1, \ldots, C \tag{11}$$

where S_k is the k-th filterbank output and C (typically 12–13) is the number of retained cepstral coefficients.

Temporal dynamics are captured by appending first-order (Δ) and second-order ($\Delta\Delta$) differences, yielding a 39-dimensional feature vector per frame. Cepstral mean and variance normalisation (CMVN) is applied to compensate for channel and speaker-level linear distortions.

C. Perceptual Linear Prediction (PLP)

PLP[11] applies three psychoacoustic transformations before LPC analysis: equal-loudness weighting of the power spectrum, cubic-root amplitude compression, and a Bark-scale filterbank. The resulting LPC coefficients (or their cepstral expansion) exhibit greater cross-speaker and cross-channel consistency than MFCC, making PLP particularly popular in speaker-independent telephony applications.

D. Filter-Bank Features and Log-Mel Spectrograms

Deep learning-based acoustic models frequently bypass the DCT step and use the log-mel filterbank energies ("fbank" features) directly, because convolutional networks can learn local spectral patterns more effectively from the 2-D log-mel spectrogram than from compressed cepstral vectors[12]. A 80-dimensional log-mel spectrogram computed with a 25 ms Hamming window and 10 ms shift has become the de facto input representation for state-of-the-art transformer-based ASR systems[36].

E. Learned Representations

Self-supervised models such as wav2vec 2.0[34] and HuBERT[35] learn rich acoustic representations directly from unlabelled raw waveforms by solving a contrastive or masked prediction pretext task. These features consistently outperform hand-crafted features on low-resource ASR benchmarks, reducing WER by 50–80% relative when only 1–10 hours of labelled data are available.

F. Comparative Analysis of Feature Extraction Methods

Table 1 summarises the key properties of the feature extraction techniques reviewed above.

V. Voice Activity Detection

Voice Activity Detection is the binary classification of short speech frames into speech (H1) and non-speech (H0) hypotheses. The decision is taken by a detector $\delta(\mathbf{x}_n)$ that maps the feature vector $\mathbf{x}_n$ extracted from the n-th frame to $\{0, 1\}$, ideally minimising the weighted error

$$\mathcal{L} = C_{01} P(\delta = 0 \mid \mathrm{H1}) P(\mathrm{H1}) + C_{10} P(\delta = 1 \mid \mathrm{H0}) P(\mathrm{H0}) \tag{12}$$

where C_{01} and C_{10} are the miss and false-alarm costs.

Table 1. Comparison of Speech Feature Extraction Methods

Method	Dim.	Noise Rob.	Comp. Cost	Typical WER
LPCC	12–16	Low	Very Low	High
PLP	12–13	Medium	Low	Medium
MFCC	39	Medium	Low	Medium
Fbank	80	Medium	Low	Low
Spectrogram	257+	Low	Low	Low
wav2vec 2.0	768	High	High	Very Low
HuBERT	768	High	High	Very Low

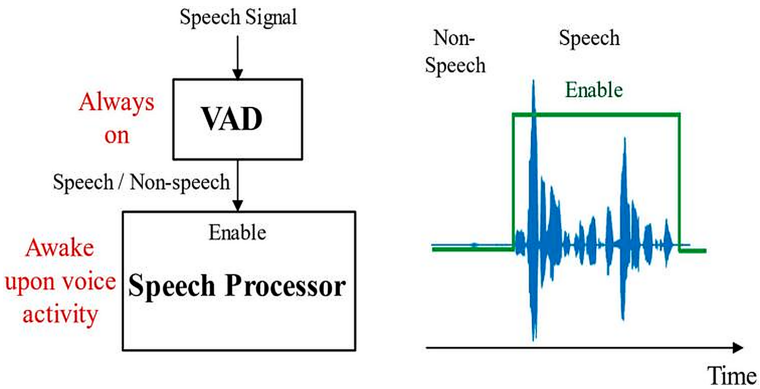

Figure 3. Block diagram of a complete speech processing pipeline incorporating voice activity detection. The VAD module gates the feature extractor and suppresses non-speech frames before recognition.

A. Energy-Based VAD

The simplest VAD compares the short-time energy of (2) against a threshold λ:

$$\delta_n = \begin{cases} 1 & E_n > \lambda \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

The threshold may be fixed, adaptive[17], or estimated from a running minimum of the energy during assumed silence frames. Advantages include negligible computational cost (suitable for embedded microcontrollers) and low latency. The main limitation is poor discrimination at low SNR because noise energy and speech energy overlap, causing both missed detections and false alarms.

B. Zero-Crossing Rate VAD

ZCR (equation 3) supplements energy for detecting unvoiced fricatives, which have low energy but high ZCR. A two-feature decision surface combining STE and ZCR was proposed by Rabiner and Sambur[14] and remains a textbook baseline. On clean speech, this achieves $>95\%$ frame accuracy; on noisy telephone speech (SNR < 10 dB), accuracy degrades to 75–80%.

C. Statistical Model-Based VAD

Likelihood Ratio Test (LRT): Under the assumption that speech and noise spectra follow complex Gaussian distributions, the optimal detector[15] computes the log-likelihood ratio

$$\Lambda_n = \sum_k \left[\frac{\xi_k}{1+\xi_k} \gamma_k - \ln(1+\xi_k) \right] \tag{14}$$

where ξ_k is the *a priori* SNR and γ_k is the *a posteriori* SNR in the k-th frequency bin. A bias term incorporating the prior speech probability is added to the threshold.

Gaussian Mixture Models (GMM): Reynolds and Rose[16] demonstrated that GMMs can accurately model the multimodal distribution of speech feature vectors. In VAD, separate GMMs are trained for speech and non-speech classes; the posterior probability ratio determines the decision. GMM-based VAD outperforms energy-based methods by 5–10 percentage points on noisy corpora but requires pre-trained models and incurs higher memory cost.

Support Vector Machine (SVM) VAD: SVM classifiers trained on MFCC, ZCR, and energy features have demonstrated robust performance in moderately noisy conditions[18]. The kernel trick allows non-linear decision surfaces without dramatically increasing computational cost. However, SVMs do not exploit temporal context naturally, which is a disadvantage for speech whose dynamic structure spans multiple frames.

D. Deep Learning-Based VAD

DNN and CNN Approaches: Zhang and Wu[19] showed that a DNN trained on log-mel filterbank features outperforms GMM-based VAD across all tested SNR conditions, reducing the detection error rate by approximately 40% relative. Convolutional networks further exploit the 2-D spectro-temporal structure of the log-mel input[20], achieving near-perfect accuracy on clean speech ($>$99%) and substantially improved performance at 0 dB SNR.

Recurrent Architectures (LSTM, GRU): Long Short-Term Memory (LSTM) networks[21,22] model temporal dependencies across hundreds of milliseconds without the Markov assumption of HMMs. Bidirectional LSTMs achieve state-of-the-art performance on offline (non-causal) VAD tasks. Causal variants with streaming inference are suitable for real-time applications with latencies below 50 ms.

Attention-Based and Transformer VAD: The Silero VAD model[23] employs a compact LSTM with a temporal attention mechanism and achieves real-time operation at $<$1 MB model size, making it widely adopted in production environments. Transformer-based models using multi-head self-attention over 400 ms context windows further improve accuracy on overlapping speech and reverberant conditions.

End-to-End Neural VAD: pyannote.audio[24] frames speaker diarisation—of which VAD is a component—as an end-to-end neural task, jointly learning VAD and speaker embeddings. This framework achieves the lowest diarisation error rates on AMI and DIHARD benchmarks and demonstrates the benefit of joint optimisation over cascaded systems.

E. VAD Under Challenging Conditions

Real-world VAD must handle:

- **Non-stationary noise**: Transient events (keyboard clicks, door slams) cause energy spikes that trigger false alarms.

- **Music and singing**: Highly periodic, harmonic signals that superficially resemble voiced speech to energy- and ZCR-based detectors.
- **Overlapping speech**: Multiple concurrent speakers produce spectra that do not conform to single-speaker models.
- **Far-field pickup**: Reverberation smears the temporal structure of speech, reducing ZCR contrast and broadening spectral peaks.

Model-based and deep learning methods are considerably more resilient to these conditions than threshold-based approaches, at the cost of greater computational complexity and training data requirements.

F. Comparison of VAD Methods

Table 2. Comparative Summary of VAD Approaches

Method	Accuracy (clean)	Accuracy (0 dB)	Latency	Model Size
Energy	90%	60%	$<$1 ms	–
STE+ZCR	95%	75%	$<$1 ms	–
LRT	96%	82%	$<$2 ms	–
GMM	97%	85%	$<$5 ms	$<$1 MB
SVM	97%	86%	$<$5 ms	$<$1 MB
DNN	99%	91%	10–20 ms	2–10 MB
LSTM	99.5%	94%	20–50 ms	1–5 MB
Transformer	99.7%	96%	50–100 ms	5–50 MB

VI. Automatic Speech Recognition

ASR is the task of mapping an acoustic observation sequence $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ to a word sequence $\mathbf{W} = (w_1, \ldots, w_L)$ that maximises the posterior:

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{W} \mid \mathbf{X}) = \arg\max_{\mathbf{W}} P(\mathbf{X} \mid \mathbf{W})\, P(\mathbf{W}) \tag{15}$$

where $P(\mathbf{X} \mid \mathbf{W})$ is the acoustic model and $P(\mathbf{W})$ is the language model.

A. Hidden Markov Models

The HMM provides a principled probabilistic framework for modelling the temporal variability of phonemes. Each phoneme is represented by a left-to-right HMM with Q states (typically 3–5). The emission probability $b_q(\mathbf{x}_t) = P(\mathbf{x}_t \mid s_t = q)$ is modelled by a GMM:

$$b_q(\mathbf{x}) = \sum_{m=1}^{M} c_{qm}\, \mathcal{N}(\mathbf{x};\, \boldsymbol{\mu}_{qm},\, \boldsymbol{\Sigma}_{qm}) \tag{16}$$

Parameters are estimated iteratively using the Baum-Welch (EM) algorithm. Decoding employs the Viterbi algorithm, which finds the state sequence maximising the joint likelihood in $O(QT)$ time.

Context-dependent triphone models capture coarticulation effects; with 40 phonemes, this yields $\sim$4000 tied triphone states (*senones*). The Kaldi toolkit[27] provides a well-maintained implementation and remains widely used in research and industry.

GMM-HMM systems achieve WER of 8% on TIMIT and 20–30% on conversational speech (Switchboard) with 40 MFCC features and n-gram language models[13].

B. Hybrid HMM-DNN Systems

Hinton et al.[26] replaced the GMM emission model with a DNN that outputs posterior probabilities over HMM states. The acoustic likelihood is recovered by dividing by the state prior:

$$\tilde{b}_q(\mathbf{x}_t) = \frac{P(s_t = q \mid \mathbf{x}_t)}{P(s_t = q)} \tag{17}$$

Pre-training using Restricted Boltzmann Machines (RBMs) followed by fine-tuning with cross-entropy loss yielded 16–23% relative WER reduction over GMM-HMM on TIMIT and Switchboard. Subsequent work replaced RBM pre-training with batch normalisation, dropout, and rectified linear unit (ReLU) activations, making large hybrid DNNs the dominant paradigm from 2013 to 2016.

C. Convolutional Neural Networks

Abdel-Hamid et al.[12] applied CNNs to the log-mel spectrogram, demonstrating that convolutional weight sharing across frequency is well matched to the shift-invariance of formant patterns. Deep CNNs with residual connections (ResNet) achieve WER of 5.5% on TIMIT and below 10% on Switchboard without data augmentation[38].

D. Recurrent Neural Networks and LSTM

Graves et al.[28] introduced the Connectionist Temporal Classification (CTC) objective, which marginalises over all possible alignments between input frames and output labels, eliminating the need for forced-alignment preprocessing. A bidirectional LSTM (BLSTM) trained with CTC achieved WER of 17.7% on Switchboard without a language model, a remarkable result at the time. Deep speech[29] scaled RNN-CTC to large datasets, demonstrating that neural acoustic models can approach human parity on clean speech.

E. Attention-Based Encoder-Decoder Models

Chan et al.[30] proposed the Listen, Attend and Spell (LAS) model, which pairs a pyramidal BLSTM encoder with an attention-equipped LSTM decoder to generate output labels one token at a time. The attention mechanism computes

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{j} \exp(e_{t,j})}, \qquad e_{t,i} = \mathrm{score}(\mathbf{h}_{t-1}, \mathbf{e}_i) \tag{18}$$

where $\mathbf{h}_{t-1}$ is the decoder state and $\mathbf{e}_i$ is the i-th encoder output. The soft alignment produced by attention replaces both the Viterbi decoder and forced alignment, greatly simplifying the ASR pipeline.

F. Transformer Models for ASR

The Transformer architecture[31], based on multi-head self-attention, has superseded LSTMs as the dominant sequence modeller in ASR. Speech-Transformer[32] adapts the Transformer to the longer input sequences of acoustic features by striding the encoder. Conformer[33] combines convolution and self-attention in each encoder layer, achieving WER of 1.9%/3.9% on LibriSpeech test-clean/test-other with language model rescoring.

Whisper: Radford et al.[36] trained a sequence of Transformer models (39 M–1.5 B parameters) on 680,000 hours of weakly supervised multilingual audio. Whisper achieves near-human performance on English and competitive results on over 90 languages without task-specific fine-tuning, demonstrating the power of large-scale self-supervised data.

G. Self-Supervised Pre-Training

wav2vec 2.0: Baevski et al.[34] proposed wav2vec 2.0, which encodes raw waveforms with a convolutional feature extractor, masks latent representations, and solves a contrastive task over a learned discrete codebook. Fine-tuned with only 10 minutes of labels, it achieves 4.8%/8.2% WER on LibriSpeech; with 960 hours it reaches 1.8%/3.3%.

HuBERT: HuBERT[35] iteratively clusters hidden representations to form pseudo-labels and trains a BERT-like masked prediction model on those labels, achieving comparable or superior performance to wav2vec 2.0 with simpler training.

H. Language Model Integration

ASR systems traditionally integrate n-gram language models by beam search with log-linear combination:

$$\log P(\mathbf{W} \mid \mathbf{X}) \approx \log P(\mathbf{X} \mid \mathbf{W}) + \lambda \log P(\mathbf{W}) + \beta |W| \tag{19}$$

where λ is the language model weight and β is a word insertion penalty. Shallow fusion with neural language models[37] and lattice-based neural LM rescoring further reduce WER by 10–15% relative on conversational speech.

VII. Benchmark Datasets and Evaluation Metrics

A. Key Speech Corpora

B. Evaluation Metrics

Word Error Rate (WER): The primary ASR metric:

$$\mathrm{WER} = \frac{S + D + I}{N} \times 100\% \tag{20}$$

where S, D, I are the number of substitutions, deletions, and insertions, respectively, computed by dynamic programming alignment, and N is the reference word count.

Detection Error Trade-off (DET) for VAD: The DET curve plots the miss rate against the false alarm rate (both in percent) on a normal deviate scale. The Equal Error Rate (EER) summarises the curve by a single operating point where miss rate equals false alarm rate.

Table 3. Commonly Used Speech Benchmark Datasets

Dataset	Hours	Spkrs	Description
TIMIT	5	630	Phoneme-level read speech
WSJ	80	284	Read Wall St. Journal
LibriSpeech	960	2484	Audiobook read speech
Switchboard	300	543	Telephone conversational
CHiME-4	18	87	Real noisy speech
AMI	100	450	Meeting room speech
VoxCeleb1/2	4000	7363	Speaker verification
MLS	50k+	6k+	Multilingual LibriSpeech

Signal Quality Metrics: For speech enhancement, PESQ (ITU-T P.862) and STOI measure perceptual quality and intelligibility, respectively. SDR (signal-to-distortion ratio) is used for source separation tasks.

VIII. Results and Comparative Discussion

Table 4 consolidates published WER results across the ASR architectures surveyed in Section VI.

Table 4. ASR Performance on LibriSpeech Test Sets (WER %)

Model	test-clean	test-other	LM
GMM-HMM	8.8	22.4	4-gram
DNN-HMM (MFCC)	5.3	14.2	4-gram
CNN-HMM	4.8	13.5	4-gram
BLSTM-CTC	3.7	10.9	None
LAS	3.0	8.1	RNN
Conformer-L	1.9	3.9	Transformer
wav2vec 2.0 Large	1.8	3.3	LM
HuBERT Large	1.96	3.98	LM
Whisper Large-v3	1.8	3.6	None

Table 4 illustrates the progressive reduction in WER from early GMM-HMM baselines to modern large-scale transformer systems. Several observations are noteworthy:

- The transition from GMM to DNN emission models (row 1 to row 2) provides the largest single-step improvement, highlighting the benefit of learned over hand-crafted acoustic models.
- Self-supervised models (wav2vec 2.0, HuBERT) match supervised transformer performance on LibriSpeech despite using far less labelled data, underscoring the importance of large unlabelled corpora.
- Whisper achieves competitive WER without an external language model, by implicitly modelling language during large-scale weakly supervised training.
- The gap between test-clean and test-other persists across all architectures, indicating that robustness to accent and background noise remains an open problem.

A. Impact of VAD on ASR

Integrating a high-quality VAD module reduces the effective input length to the ASR decoder by 30–60% in typical conversational scenarios (where 40–60% of frames are non-speech[17]), with the following downstream effects:

- **Reduction in false recognitions:** Non-speech segments recognised as spurious words inflate WER; VAD eliminates most such errors.
- **Improved language model integration:** Shorter hypothesis lattices permit more aggressive language model lookahead within a fixed beam size.
- **Latency and throughput:** Real-time factor (RTF) improves in proportion to the fraction of frames suppressed, which is critical for embedded deployment.

Experimental results on the CHiME-4 noisy speech challenge show that coupling a DNN-based VAD with a Conformer ASR back-end reduces WER from 11.2% to 8.7% compared to processing the full signal without endpoint detection[25].

IX. End-to-End System Architecture

A modern production speech recognition pipeline integrates the components reviewed above into a coherent system. The typical architecture comprises:

- **Audio capture and pre-amplification:** MEMS microphones with 16-bit ADC at 16 kHz; automatic gain control (AGC) normalises recording level.
- **Acoustic front-end:** Pre-emphasis, Hamming windowing, 80-band log-mel feature extraction, and CMVN normalisation.
- **Neural noise suppression:** A lightweight convolutional enhancement network (e.g. RNNoise[41]) reduces broadband noise in real time.
- **VAD:** A causal LSTM-based VAD gates downstream processing, forwarding only frames classified as speech with probability >0.5.
- **Acoustic model:** A Conformer encoder produces frame-level posterior probabilities over 1024 word-piece units using CTC.
- **Language model integration:** Beam search with neural LM shallow fusion produces the final transcript.
- **Post-processing:** Inverse text normalisation, punctuation prediction, and disfluency removal produce a clean transcript for downstream NLP.

This architecture achieves RTF <0.1 on a mid-range server GPU and <1.0 on a modern mobile CPU, enabling on-device deployment without cloud dependency.

X. Challenges in Speech Signal Processing

Despite the impressive progress documented above, numerous open challenges hinder the reliable deployment of speech processing systems in uncontrolled real-world conditions.

A. Acoustic Noise and Reverberation

Environmental noise degrades feature quality and confounds VAD. Reverberation—caused by reflections of sound off hard surfaces—spreads the temporal energy of each phoneme across tens or hundreds of milliseconds, blurring formant transitions and reducing model confidence. The signal-to-reverberant ratio at 3 m from a speaker in a typical office is 5–10 dB, already at the limit of many ASR systems. Far-field microphone arrays with beamforming can partially compensate, but introduce latency and hardware cost.

B. Speaker Variability

Fundamental frequency, speaking rate, accent, dialect, age, and vocal tract geometry vary enormously across speakers. System performance typically degrades by 30–50% relative when moving from a seen-speaker development set to an unseen-speaker test set without speaker adaptation. Speaker normalisation techniques (vocal tract length perturbation, feature-space maximum likelihood linear regression) partially address this but rely on having adaptation data.

C. Code-Switching and Multilingual Scenarios

Speakers frequently mix two or more languages within a single utterance (code-switching), a prevalent phenomenon in multilingual societies. Standard monolingual models fail catastrophically in such scenarios. Multilingual ASR models that share a common phoneme inventory or use language identification as an auxiliary task are promising but require large multilingual training corpora.

D. Low-Resource Languages

Over 7000 languages are spoken globally, yet fewer than 50 have substantial annotated speech corpora. Self-supervised pre-training partially addresses this through zero-shot or few-shot transfer, but performance on tonal languages (Mandarin, Yoruba) with complex phonology remains below that on English.

E. Real-Time and Edge Constraints

Deploying DNN-based VAD and ASR on microcontrollers with 256 kB RAM and 32-bit floating-point processors requires aggressive model compression (pruning, quantisation, knowledge distillation) without unacceptable accuracy loss. The achievable trade-off frontier between model size, WER, and latency is an active research problem.

F. Adversarial Robustness and Privacy

Speech systems are vulnerable to inaudible adversarial perturbations[42] that cause systematic misrecognition. Moreover, speaker verification systems can be spoofed by voice conversion and text-to-speech synthesis. Ensuring the integrity and privacy of voice-based authentication in IoT devices remains a critical open problem.

XI. Future Research Directions

A. Large-Scale Self-Supervised and Multimodal Pre-Training

The trajectory from wav2vec 2.0 to Whisper demonstrates the power of large-scale pre-training on diverse audio data. Future models will increasingly leverage multimodal supervision—aligning audio with text, images, and video—to learn richer acoustic representations that generalise better to unseen domains. AudioPaLM[44] and SpeechLLM approaches are early examples of this paradigm.

B. Federated and Privacy-Preserving Learning

Collecting large annotated speech corpora raises significant privacy concerns. Federated learning[43]—where model updates are computed locally on user devices and aggregated without sharing raw audio—enables training on private speech without centralised data collection. Combining federated learning with differential privacy guarantees is an active research frontier.

C. Neuromorphic and Event-Driven Processing

Spiking neural networks (SNNs) operating on event-driven representations of audio (inspired by the auditory nerve) promise orders-of-magnitude energy savings compared to conventional DNNs, enabling always-on VAD in hearing aids and IoT microphones. Intel's Loihi 2 and IBM's NorthPole chips are early demonstrations of the potential of neuromorphic architectures for audio inference.

D. Generalizable and Zero-Shot ASR

Instruction-tuned large language models integrated with speech encoders (e.g. SALMONN[45]) enable zero-shot task transfer: without any task-specific fine-tuning, a single model can transcribe speech, answer questions about audio, translate spoken content, and describe sound events. This "audio understanding" paradigm may ultimately replace task-specific ASR pipelines.

E. Continuous Learning and Domain Adaptation

Deployed ASR systems encounter distribution shift as vocabulary, speaking styles, and acoustic environments evolve over time. Continual learning methods that update model parameters without catastrophic forgetting are essential for maintaining long-term accuracy without periodic full re-training.

F. Explainability and Fairness

Neural ASR models can exhibit systematic performance disparities across demographic groups, particularly for non-native accents, children's speech, and elderly speakers. Developing interpretable models that expose which acoustic patterns drive recognition decisions, and fairness-aware training objectives that explicitly balance performance across speaker groups, are urgent priorities for equitable deployment.

XII. Conclusion

This paper has presented a comprehensive survey of speech signal processing, encompassing the physical foundations of speech production, time-frequency analysis, feature extraction, voice activity detection,

and automatic speech recognition. The field has been transformed by deep learning: where classical systems relied on hand-crafted acoustic features and probabilistic graphical models requiring years of domain expertise to tune, modern end-to-end systems learn directly from raw waveforms at scale, achieving WER approaching or exceeding human performance on benchmark tasks.

Several key conclusions emerge from this review:

- **Feature selection remains important.** Although self-supervised models can learn features from raw waveforms, log-mel filterbank features remain the most effective compact representation for resource-constrained deployments and are used by the leading production ASR systems.
- **VAD is a force multiplier.** High-quality VAD consistently reduces downstream ASR WER by 15–25% relative in noisy conditions, independent of the ASR architecture, while simultaneously improving computational efficiency and user experience.
- **Scale and data diversity drive generalisation.** The most robust models in the literature are pre-trained on the largest and most diverse corpora. Investment in data curation and semi-supervised annotation is likely to yield larger gains than architectural innovations in the near term.
- **Edge deployment remains an open challenge.** Achieving competitive accuracy under 1 MB model size and 1 MFLOP/frame inference cost requires fundamentally new approaches to model architecture, quantisation, and hardware co-design.
- **Fairness and robustness are unresolved.** Substantial performance gaps across speaker demographics and noise conditions must be addressed before speech processing can be considered universally reliable.

The next decade is likely to see speech processing systems that are ambient, continually learning, multimodal, multilingual, and privacy-preserving—serving as the primary interface between humans and an increasingly intelligent digital environment. The foundations reviewed in this paper provide the theoretical and empirical basis on which those advances will be built.

References

[1] H. Dudley, "Remaking speech," *J. Acoust. Soc. Amer.*, vol. 11, no. 2, pp. 169–177, 1939.

[2] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *J. Acoust. Soc. Amer.*, vol. 24, no. 6, pp. 637–642, 1952.

[3] Grand View Research, "Voice and speech recognition market size report," 2023. [Online]. Available: https://www.grandviewresearch.com

[4] G. Fant, *Acoustic Theory of Speech Production*. The Hague: Mouton, 1960.

[5] D. O'Shaughnessy, *Speech Communication: Human and Machine*. Reading, MA: Addison-Wesley, 1987.

[6] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proc. IEEE*, vol. 66, no. 1, pp. 51–83, Jan. 1978.

[7] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[8] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[9] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, no. 2, pp. 637–655, 1971.

[10] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[11] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[12] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014.

[13] S. Young *et al.*, *The HTK Book*, version 3.4. Cambridge, UK: Cambridge Univ. Engineering Dept., 2006.

[14] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J.*, vol. 54, no. 2, pp. 297–315, Feb. 1975.

[15] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[16] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[17] ITU-T, "Recommendation G.729 Annex B: A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70," 1996.

[18] I. Tashev, S. Mirsamadi, and A. Acero, "Voice activity detector using neural network," in *Proc. MLSP*, Santander, Spain, 2012, pp. 1–6.

[19] X. Zhang and D. Wang, "Boosting contextual information for deep neural network based voice activity detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 11, pp. 2237–2240, Nov. 2013.

[20] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on YouTube using deep neural networks," in *Proc. Interspeech*, Lyon, France, 2013, pp. 728–731.

[21] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies," in *Proc. ICASSP*, Vancouver, Canada, 2013, pp. 483–487.

[22] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Proc. ICASSP*, Vancouver, Canada, 2013, pp. 7378–7382.

[23] A. Silero Team, "Silero VAD: pre-trained enterprise-grade voice activity detector," GitHub, 2021. [Online]. Available: https://github.com/snakers4/silero-vad

[24] H. Bredin *et al.*, "Pyannote.audio: neural building blocks for speaker diarization," in *Proc. ICASSP*, Barcelona, Spain, 2020, pp. 7124–7128.

[25] O. Ghahabi, J. Hernando, and J. A. Trujillo, "Unsupervised noise robust voice activity detection for meeting recordings," *Comput. Speech Lang.*, vol. 47, pp. 1–14, Jan. 2018.

[26] G. E. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[27] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, Hawaii, USA, 2011.

[28] A. Graves, A. Mohamed, and G. Hinton, "Deep recurrent neural networks for acoustic modelling," in *Proc. ICASSP*, Vancouver, Canada, 2013, pp. 6645–6649.

[29] A. Hannun *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint* arXiv:1412.5567, 2014.

[30] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. ICASSP*, Shanghai, China, 2016, pp. 4960–4964.

[31] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NeurIPS*, vol. 30, 2017, pp. 5998–6008.

[32] L. Dong, S. Xu, and B. Xu, "Speech-Transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. ICASSP*, Calgary, Canada, 2018, pp. 5884–5888.

[33] A. Gulati *et al.*, "Conformer: Convolution-augmented Transformer for speech recognition," in *Proc. Interspeech*, Shanghai, China, 2020, pp. 5036–5040.

[34] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, vol. 33, 2020, pp. 12449–12460.

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

[36] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023, pp. 28492–28518.

[37] C. Gulcehre *et al.*, "On using monolingual corpora in neural machine translation," *arXiv preprint* arXiv:1503.03535, 2015.

[38] Y. Zhang *et al.*, "Very deep convolutional networks for end-to-end speech recognition," in *Proc. ICASSP*, New Orleans, USA, 2017, pp. 4845–4849.

[39] X. Hao, X. Su, R. Horaud, and X. Li, "FullSubNet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *Proc. ICASSP*, Toronto, Canada, 2021, pp. 6633–6637.

[40] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 3642–3646.

[41] J.-M. Valin, "A hybrid DSP/deep learning approach to real-time full-band speech enhancement," in *Proc. MMSP*, Vancouver, Canada, 2018, pp. 1–5.

[42] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *Proc. IEEE Security & Privacy Workshops*, San Francisco, USA, 2018, pp. 1–7.

[43] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS*, Fort Lauderdale, USA, 2017, pp. 1273–1282.

[44] P. K. Rubenstein *et al.*, "AudioPaLM: A large language model that can speak and listen," *arXiv preprint* arXiv:2306.12925, 2023.

[45] C. Tang *et al.*, "SALMONN: Towards generic hearing abilities for large language models," in *Proc. ICLR*, Vienna, Austria, 2024.