



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

From Signal Processing to LLM-Integrated Speech Intelligence: A Narrative and Conceptual Review of Speech Processing

Duyen Nguyen Thi¹, Duc-Quang Vu¹, Tran Minh Bao^{2*}, Dac-Nhuong Le³

¹Thai Nguyen University of Information and Communication Technology, Thai Nguyen, Vietnam.

²Ho Chi Minh University of Industry and Trade, Ho Chi Minh, Vietnam.

³Haiphong University, Hai Phong, Vietnam.

*Corresponding author: baotm@huit.edu.vn

Abstract

Audio and speech processing has evolved significantly over the past two decades, driven by advances in signal processing, statistical modeling, deep learning, and, more recently, large language models (LLMs). This paper presents a narrative and conceptual review of this evolution, highlighting major paradigm shifts that have shaped the field rather than attempting an exhaustive or systematic synthesis of the literature. We begin by examining classical signal processing approaches, which rely on handcrafted features and spectral analysis, followed by statistical methods that introduce probabilistic modeling for handling temporal variability and uncertainty. We then discuss the deep learning era, where end-to-end neural architectures enable hierarchical representation learning and substantially improve performance across a wide range of tasks, including speech recognition, synthesis, enhancement, and separation. Building on these developments, we analyze the emergence of self-supervised and foundation speech models, which leverage large-scale unlabeled data to learn general-purpose representations with improved robustness and cross-domain generalization. We further explore the integration of speech processing with large language models, emphasizing the transition from transcription-centric systems to multimodal speech intelligence. In this context, we highlight the fundamental challenge of bridging the modality gap between continuous acoustic signals and discrete textual representations. In addition to reviewing methodological advances, we identify several persistent challenges, including noise robustness, multi-speaker interference, data scarcity, evaluation limitations, and computational efficiency. We also discuss emerging research directions, such as unified speech foundation models, speech-native interaction with LLMs, and multimodal integration. This narrative review provides a unified conceptual perspective on the evolution of speech processing and outlines key directions for future research toward more integrated and intelligent speech systems.

Keywords: Speech Processing, Automatic Speech Recognition, Speech Synthesis, Deep Learning, Self-Supervised Learning, Multimodal Speech Processing, Speech Foundation Models

1. Introduction

Speech processing has been one of the most dynamic research areas at the intersection of signal processing, machine learning, and artificial intelligence (Purwins et al. 2019, Mehrish et al. 2023). Over the past two decades, the field has undergone a sequence of major paradigm shifts, evolving from classical signal processing techniques to statistical modeling, deep learning, and, more recently, self-supervised and large language model (LLM)-integrated systems. These transitions have not only improved performance across core tasks but have also fundamentally reshaped how speech is represented, modeled, and utilized in intelligent systems. As illustrated in Fig. 1, speech processing has evolved from signal-level modeling to multimodal and LLM-integrated systems.

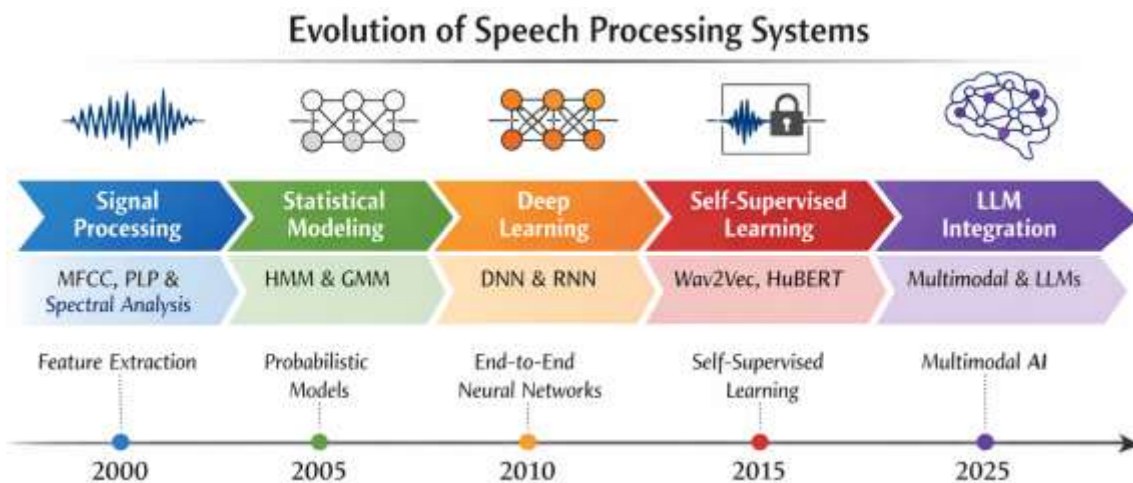


Figure 1. Conceptual evolution of speech processing systems from signal processing to LLM-integrated speech intelligence.

While this review focuses on developments over the past twenty years, earlier foundational work is included where necessary to contextualize subsequent advances. In the classical era, speech was primarily treated as a time-varying signal, and research focused on designing robust and compact representations through signal transformations. Perceptually motivated features such as Mel-frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP) became standard representations for speech analysis [Davis and Mermelstein 1980, Hermansky 1990]. In parallel, noise reduction techniques such as spectral subtraction and minimum mean-square error (MMSE) estimation were developed to improve robustness in adverse acoustic environments [Boll 1979, Ephraim and Malah 1984]. These approaches established the foundation for early speech recognition, enhancement, and speaker modeling systems [Reynolds 1995].

The introduction of statistical modeling marked a significant shift toward probabilistic representations of speech. Hidden Markov models (HMMs) became the dominant framework for modeling temporal dynamics [Rabiner 1989, Gales and Young 2008], typically combined with Gaussian mixture models for acoustic modeling. This paradigm enabled scalable and data-driven systems, supporting both automatic speech recognition (ASR) and text-to-speech synthesis (TTS) [Tokuda et al. 2000, Zen et al. 2009]. Concatenative and parametric synthesis methods further expanded the design space of TTS systems [Hunt and Black 1996, Stylianou 2009]. Despite their success, statistical approaches remained constrained by handcrafted features and modular pipelines, limiting their ability to capture complex acoustic patterns.

The integration of artificial intelligence, connected technologies, and smart sensing systems has expanded the potential of technology-assisted monitoring and decision support in real-world applications (Patra & Singh, 2026). The emergence of deep learning fundamentally transformed speech processing by enabling end-to-end learning of hierarchical representations [Purwins et al. 2019, Nassif et al. 2019, Mehrish et al. 2023]. Deep neural networks (DNNs) significantly improved acoustic modeling [Hinton et al. 2012], while recurrent neural networks (RNNs) and long short-term memory (LSTM) architectures enhanced the modeling of temporal dependencies [Graves et al. 2013, Sak et al. 2014]. Attention-based encoder-decoder models simplified traditional pipelines by directly mapping speech to text [Chan et al. 2016, Li 2022, Prabhavalkar et al. 2024], and transformer architectures further improved scalability and long-range dependency modeling [Vaswani et al. 2017, Kheddar et al. 2024].

In addition, neural approaches revolutionized speech synthesis and separation, achieving substantial improvements in naturalness and robustness [Wang et al. 2017, Shen et al. 2018, Luo and Mesgarani 2019, Yu et al. 2017].

Despite these advances, supervised deep learning approaches remain dependent on large amounts of labeled data [Mohamed et al. 2022]. This limitation has motivated the development of self-supervised learning (SSL), where models learn representations from large-scale unlabeled speech. Early approaches such as wav2vec demonstrated the feasibility of representation learning from raw audio [Schneider et al. 2019], while more advanced models such as wav2vec 2.0, HuBERT, and WavLM have achieved strong performance across a wide range of tasks [Baevski et al. 2020, Hsu et al. 2021, Chen et al. 2022]. These approaches have enabled the emergence of speech foundation models that generalize across tasks, languages, and domains.

More recently, the rapid development of large language models has opened new directions for integrating speech processing with higher-level semantic reasoning. Traditional systems treat speech recognition as the final objective, converting speech into text for downstream processing. However, emerging approaches aim to directly integrate speech representations with language models, enabling multimodal reasoning and

interaction [Zhu et al. 2021, Sheng et al. 2024]. Multimodal learning frameworks such as CLIP demonstrate how different modalities can be aligned within shared representation spaces [Radford et al. 2021], while speech-language models such as AudioLM, VALL-E, and SpeechGPT extend language modeling techniques to the speech domain [Borsos et al. 2023, Wang et al. 2023, Zhang et al. 2023]. These developments highlight a transition from transcription-centric systems toward speech-centric intelligence.

Despite rapid progress, several fundamental challenges remain. These include robustness to noise and multi-speaker environments, data scarcity in low-resource settings, discrepancies between objective metrics and human perception, and the difficulty of aligning speech representations with semantic reasoning. In particular, bridging the modality gap between continuous acoustic signals and discrete textual representations remains a central challenge in LLM-integrated speech systems.

This paper presents a narrative and conceptual review of speech processing, focusing on the evolution of methodological paradigms and the emergence of LLM-integrated speech intelligence. Rather than aiming to provide an exhaustive systematic survey, the paper synthesizes representative developments that illustrate how speech processing has evolved from signal-level modeling toward multimodal and LLM-integrated speech intelligence.

The selected works are used as illustrative anchors for major conceptual transitions in the field. Specifically, we:

- conceptually analyze the transition from signal processing to statistical and deep learning approaches,
- examine the role of self-supervised and foundation models in speech representation learning,
- discuss representative recent advances in integrating speech processing with large language models,
- and identify key challenges and future research directions, with emphasis on modality bridging between speech and text.

The remainder of this paper is organized as follows. Section 2 introduces the taxonomy of speech processing tasks and representations. Sections 3–6 review major methodological paradigms, including classical signal processing, statistical modeling, deep learning, and self-supervised learning. Section 7 discusses LLM-integrated speech systems, followed by cross-cutting challenges in Section 8 and future directions in Section 9. Section 10 concludes the paper.

2. Scope and Conceptual Taxonomy of Speech Processing

To analyze the evolution of speech processing from a narrative and conceptual perspective, it is useful to establish a taxonomy that captures the diversity of tasks, representation levels, and modeling paradigms. Such a framework provides an organizing lens for comparing representative approaches and highlighting how advances in representation and learning have reshaped system design.

2.1. Nature of the Review

This article is designed as a narrative and conceptual review rather than a systematic review or meta-analysis. The aim is not to exhaustively enumerate all studies in speech processing, but to synthesize representative methods, paradigms, and conceptual transitions that have shaped the field. The cited works were selected because they illustrate major methodological shifts, including classical signal processing, statistical modeling, deep learning, self-supervised learning, foundation models, and LLM-integrated speech systems.

Accordingly, the taxonomy proposed in this paper should be understood as an interpretive framework for organizing the evolution of speech processing, rather than as a formal classification derived from a systematic screening protocol.

2.2. Task Taxonomy

Speech processing encompasses a wide range of tasks, which can be broadly categorized into several key groups:

- Automatic Speech Recognition (ASR): Converting speech signals into text, forming the backbone of many speech-driven applications [Rabiner 1989, Hinton et al. 2012, Chan et al. 2016, Nassif et al. 2019, Malik et al. 2021].
- Text-to-Speech Synthesis (TTS): Generating speech from text, including concatenative, statistical, and neural approaches [Hunt and Black 1996, Zen et al. 2009, Shen et al. 2018].
- Speech Enhancement and Denoising: Improving speech quality in noisy or degraded environments [Boll 1979, Ephraim and Malah 1984, Yousif and Mahmmoud 2025, Michelsanti et al. 2021].
- Speech Separation: Extracting individual sources from multi-speaker mixtures, often referred to as the cocktail party problem [Yu et al. 2017, Luo and Mesgarani 2019].
- Speaker Recognition and Diarization: Identifying speakers and segmenting speech streams based on speaker identity [Reynolds 1995, Hansen and Hasan 2015, Anguera et al. 2012, O'Shaughnessy 2025].
- Spoken Language Understanding: Interpreting semantic content beyond transcription, often integrated with language models and conversational systems [Sethiya and Maurya 2025, Latif et al. 2023].

These tasks are increasingly interconnected in modern systems. For example, conversational AI systems integrate ASR, speaker modeling, and language understanding into unified pipelines. Furthermore, emerging applications such as speech-based healthcare and low-resource speech processing extend the scope of the field beyond traditional tasks [Phung et al. 2024].

2.3. Representation Levels

A central dimension in the evolution of speech processing is the progression of representation levels. These can be broadly categorized into four types:

- Signal-level representations: Raw waveforms and time-frequency representations that directly capture acoustic properties of speech.
- Feature-level representations: Handcrafted features such as MFCC and PLP, designed to encode perceptually relevant information [Davis and Mermelstein 1980, Hermansky 1990].
- Latent representations: Learned embeddings derived from deep neural networks and self-supervised models, capturing higher-level acoustic and phonetic structure [Schneider et al. 2019, Baevski et al. 2020].
- Semantic representations: Representations aligned with linguistic meaning and reasoning, often modeled through multimodal and language models [Radford et al. 2021, Ao et al. 2022].

The transition from handcrafted features to learned representations represents one of the most significant shifts in the field. Early systems relied heavily on domain knowledge to design features, whereas modern approaches emphasize data-driven learning that can generalize across tasks, domains, and languages.

2.4. Modeling Paradigms

The evolution of speech processing can be understood through a sequence of major modeling paradigms:

- Signal processing approaches: Focused on deterministic transformations and handcrafted features, forming the foundation of early systems.
- Statistical modeling: Introduced probabilistic frameworks such as HMM-GMM systems [Rabiner 1989, Gales and Young 2008], supported by training methods based on expectation-maximization [Bilmes 1998] and probabilistic modeling theory [Bishop 2006]. Toolkits such as HTK further standardized system development [Young et al. 2006].
- Deep learning: Enabled end-to-end learning and hierarchical representation modeling, significantly improving performance across tasks [Hinton et al. 2012, Graves et al. 2013, Vaswani et al. 2017, Purwins et al. 2019, Mehrish et al. 2023].
- Self-supervised and foundation models: Leveraged large-scale unlabeled data to learn general-purpose representations [Schneider et al. 2019, Baevski et al. 2020, Chen et al. 2022, Mohamed et al. 2022].
- LLM-integrated systems: Combined speech processing with language models to enable multimodal reasoning and interaction [Borsos et al. 2023, Zhang et al. 2023].

These paradigms are not mutually exclusive; instead, they build upon one another. For example, modern neural systems still incorporate concepts such as spectral representations and probabilistic modeling within deep learning frameworks.

2.5. Evaluation Dimensions

Evaluating speech processing systems involves multiple criteria, often with trade-offs between them:

- Accuracy: Task-specific metrics such as word error rate (WER) for ASR.
- Robustness: Performance under noise, reverberation, and domain shifts.
- Naturalness: Perceptual quality of synthesized speech, often evaluated through subjective listening tests [Zen et al. 2009].
- Generalization: Ability to perform well across speakers, languages, and environments [Babu et al. 2022].
- Efficiency: Computational and memory requirements, particularly important for real-time applications.

A persistent issue in the field is the mismatch between objective evaluation metrics and human perception. Improvements in numerical metrics do not always correspond to better perceived quality, especially in tasks such as speech synthesis and enhancement.

2.6 Discussion

The taxonomy presented above highlights that the evolution of speech processing is driven not only by advances in individual tasks but also by changes in representation and modeling paradigms. Classical approaches emphasized signal-level transformations, while modern systems increasingly rely on learned representations and cross-modal integration.

This conceptual framework provides the foundation for the subsequent narrative synthesis, where we analyze how representative paradigms—from classical signal processing to LLM-integrated systems—have shaped the development of speech processing.

3. Classical Signal Processing Approaches

Before the rise of statistical and learning-based methods, speech processing was primarily approached as a signal analysis problem, where speech was modeled as a time-varying acoustic waveform and analyzed using deterministic transformations [Oppenheim and Schaffer 2010]. This signal-centric paradigm emphasized the

design of robust and compact representations that capture perceptually and linguistically relevant information, forming the theoretical foundation for subsequent developments.

3.1 Spectral Analysis and Perceptual Feature Design

A fundamental assumption in classical speech processing is that raw waveforms are highly redundant and difficult to model directly. Time-frequency analysis techniques, such as short-time Fourier transform and filter-bank representations, are therefore used to transform speech into structured spectral representations [Oppenheim and Schaffer 2010].

Perceptually motivated features such as Mel-frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP) further refine these representations by incorporating models of human auditory perception [Davis and Mermelstein 1980, Hermansky 1990]. These features became the standard frontend for speech recognition, speaker identification, and coding systems due to their robustness and compactness.

Beyond cepstral features, alternative representations have been explored to better capture non-stationary properties of speech. Multi-resolution analysis, particularly wavelet-based approaches, provides the ability to capture both transient and long-term characteristics of speech signals. Such methods have been applied to speech coding, feature extraction, and speaker recognition [Nghia et al. 2008]. These works illustrate broader efforts to design signal representations that improve robustness and adaptability across different conditions.

3.2 Noise Reduction and Robust Signal Processing

Robustness to noise and environmental variability has been a central challenge since the early development of speech processing. Classical approaches addressed this problem through explicit signal enhancement techniques. Spectral subtraction [Boll 1979] and minimum mean-square error (MMSE) estimation [Ephraim and Malah 1984] are among the most influential methods, aiming to recover clean speech from noisy observations by modeling noise characteristics in the spectral domain.

These approaches provide interpretability and computational efficiency, enabling explicit control over system behavior. However, they rely on assumptions such as noise stationarity, which are often violated in real-world environments.

3.3. Speech Coding and Compression

Speech coding is another important application of classical signal processing, where the goal is to represent speech efficiently while preserving intelligibility and perceptual quality. Techniques such as transform coding, subband coding, and waveform modeling have been extensively studied [Spanias 1994].

Wavelet-based speech coding represents one direction in this area, leveraging multi-resolution analysis to capture both fine-grained and coarse acoustic structures. More broadly, classical coding approaches highlight fundamental trade-offs between compression efficiency, computational complexity, and perceptual quality, which continue to influence modern neural compression methods [Carnero and Drygajlo 1999].

3.4. Speaker Modeling and Early Recognition Pipelines

Classical signal processing also played a crucial role in speaker-related tasks and early speech recognition systems. In speaker recognition, spectral and cepstral features are combined with statistical models such as Gaussian mixture models (GMMs) to distinguish between speakers [Reynolds 1995]. These approaches rely heavily on the stability and discriminability of the extracted features.

Early speech recognition systems were typically organized as modular pipelines, consisting of preprocessing, feature extraction, optional enhancement, and classification. While this design provides interpretability and flexibility, it also introduces limitations, as each component is optimized independently. As a result, such systems often struggle to adapt to variability in speakers, environments, and speaking styles [Rabiner 1989, Gales and Young 2008].

3.5. Language-Specific Signal Modeling

A distinctive characteristic of classical approaches is their reliance on domain knowledge, particularly for handling language-specific properties. In tonal languages, pitch contours carry critical lexical information, requiring explicit modeling of fundamental frequency (F0).

Research in this direction explored methods for transforming language trajectories and developing speech synthesis systems tailored to specific language characteristics [Phung et al. 2012]. These approaches demonstrate the flexibility of signal processing techniques in incorporating linguistic knowledge, while also highlighting their limitations in scalability across languages and domains.

3.6. Strengths and Limitations

The classical signal processing paradigm offers several important strengths. It provides interpretable and mathematically grounded methods, enables efficient implementations suitable for real-time applications, and incorporates perceptual knowledge into system design.

However, the limitations of this paradigm are also significant. Handcrafted features impose strong assumptions about what information is important, limiting representational flexibility. Modular pipelines prevent joint optimization and adaptation. Furthermore, classical approaches struggle to capture complex nonlinear relationships and higher-level abstractions inherent in speech signals.

These limitations motivated the transition toward statistical modeling approaches, where variability and uncertainty could be handled more systematically [Rabiner 1989, Gales and Young 2008]. Importantly, this transition builds upon, rather than replaces, the foundations established by classical signal processing.

The architectural evolution of speech systems from modular pipelines to integrated models is illustrated in Fig. 2.

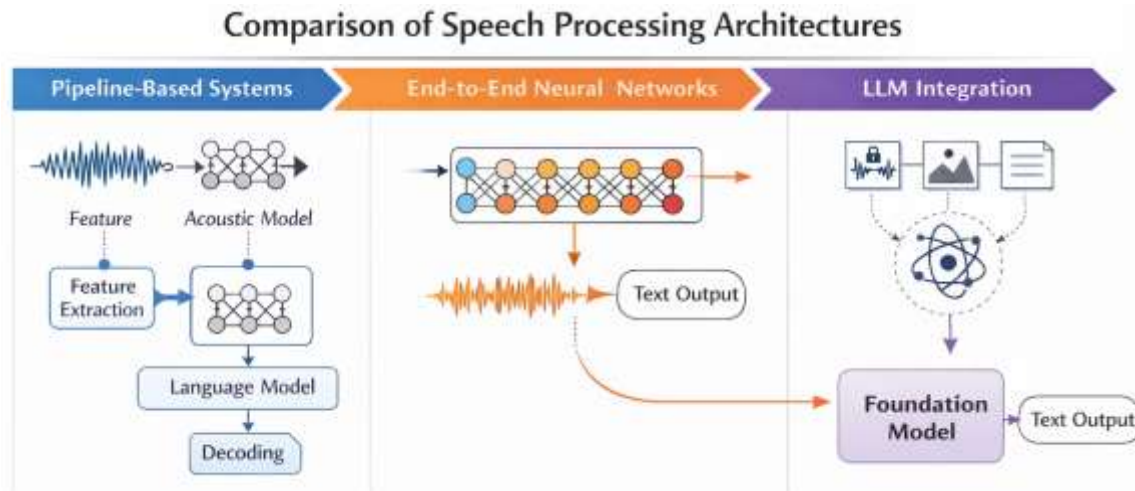


Figure 2. Conceptual comparison of speech processing architectures using automatic speech recognition (ASR) as a representative example, including pipeline-based systems, end-to-end neural models, and LLM-integrated multimodal systems.

4. Statistical and Model-Based Speech Processing

The limitations of purely signal-driven approaches led to the emergence of statistical modeling as the dominant paradigm in speech processing from the late 1990s to the early 2010s. In this paradigm, speech is modeled as a stochastic process, enabling systematic handling of variability across speakers, environments, and linguistic contexts [Rabiner 1989, Gales and Young 2008]. This shift marked a transition from deterministic transformations toward probabilistic reasoning and data-driven modeling.

4.1. Hidden Markov Models and Acoustic Modeling

Hidden Markov models (HMMs) became the core framework for modeling temporal dynamics in speech signals [Rabiner 1989, Gales and Young 2008]. By representing speech as a sequence of latent states with probabilistic transitions, HMMs provide a natural mechanism for capturing time-dependent variability.

In practice, HMMs are typically combined with Gaussian mixture models (GMMs) to model acoustic observations, forming the widely adopted GMM-HMM architecture. Training of these models is commonly performed using the expectation-maximization (EM) algorithm [Bilmes 1998], grounded in probabilistic modeling principles [Bishop 2006]. Toolkits such as HTK further standardized system development and enabled large-scale experimentation [Young et al. 2006].

This probabilistic framework offers several advantages, including explicit modeling of uncertainty, scalability to large datasets, and integration of acoustic and linguistic information. However, the performance of these systems remains constrained by the quality of handcrafted features such as MFCC and PLP.

4.2. Statistical Parametric Speech Synthesis

Statistical modeling also led to significant advances in speech synthesis. HMM-based text-to-speech (TTS) systems generate acoustic parameters such as spectral envelopes and fundamental frequency (F0) from linguistic input [Tokuda et al. 2000, Zen et al. 2009]. Compared with concatenative synthesis [Hunt and Black 1996], which selects and concatenates recorded speech segments, statistical parametric synthesis offers greater flexibility, reduced storage requirements, and improved adaptability.

In addition, voice conversion techniques further expanded the capabilities of statistical speech processing by enabling transformation of speaker characteristics across different voices [Stylianou 2009]. Comprehensive treatments of speech synthesis systems highlight the trade-offs between naturalness, flexibility, and computational efficiency [Taylor 2009, Black and Tokuda 2005].

Despite these advantages, statistical parametric methods often suffer from over-smoothing, where generated speech lacks natural variability and expressiveness. This issue arises from the averaging effects inherent in statistical estimation.

4.3. Improving Perceptual Quality and Variability

To address limitations such as over-smoothing, various techniques have been proposed to improve perceptual quality in statistical speech synthesis. These include variance enhancement, post-filtering, and hybrid approaches that combine statistical modeling with unit selection.

Representative studies have explored hybrid TTS systems and perceptual improvements in HMM-based synthesis [Phung et al. 2013, Phung et al. 2013]. In addition, research on smoothness measures and perceptual evaluation has provided insights into the relationship between statistical modeling and perceived speech quality [Nghia et al. 2016]. Further investigations into over-smoothing phenomena have highlighted the limitations of conventional statistical estimation methods in capturing fine-grained acoustic variations [Nguyen and Phung 2017].

These works reflect broader efforts to balance model flexibility, data efficiency, and perceptual realism in statistical speech systems.

4.4. Robust Modeling and Noise Compensation

Statistical modeling also enabled new approaches to robustness by incorporating noise characteristics directly into probabilistic frameworks. Instead of modifying the signal itself, these methods model the joint distribution of clean speech and noise.

4.5. Strengths and Limitations

The statistical modeling paradigm introduced several important advances. It provided a principled framework for handling uncertainty and temporal dynamics, enabled scalable training on large datasets, and improved flexibility compared with purely signal-based methods.

Nevertheless, the limitations of this paradigm became increasingly evident. The reliance on handcrafted features constrained representational capacity, while modular system design hindered end-to-end optimization. In addition, statistical models struggled to capture highly nonlinear relationships and complex acoustic structures.

These limitations ultimately motivated the transition toward deep learning approaches, which aim to learn representations directly from data and enable end-to-end optimization. Importantly, the statistical paradigm serves as a critical bridge between classical signal processing and modern neural approaches, providing many of the theoretical foundations that continue to influence contemporary systems.

5. Deep Learning for Speech Processing

The emergence of deep learning marked a fundamental turning point in speech processing, shifting the field from handcrafted feature engineering and modular pipelines toward data-driven representation learning and end-to-end optimization [Purwins et al. 2019, Mehrish et al. 2023]. Unlike statistical approaches, which rely on predefined features and relatively shallow models, deep neural networks enable hierarchical modeling of complex and nonlinear relationships directly from data [Hinton et al. 2012]. This paradigm builds upon earlier statistical frameworks while overcoming limitations related to feature design and model capacity [Rabiner 1989, Gales and Young 2008].

As shown in Fig. 3, deep learning enables hierarchical representation learning from low-level acoustic signals to high-level semantics.

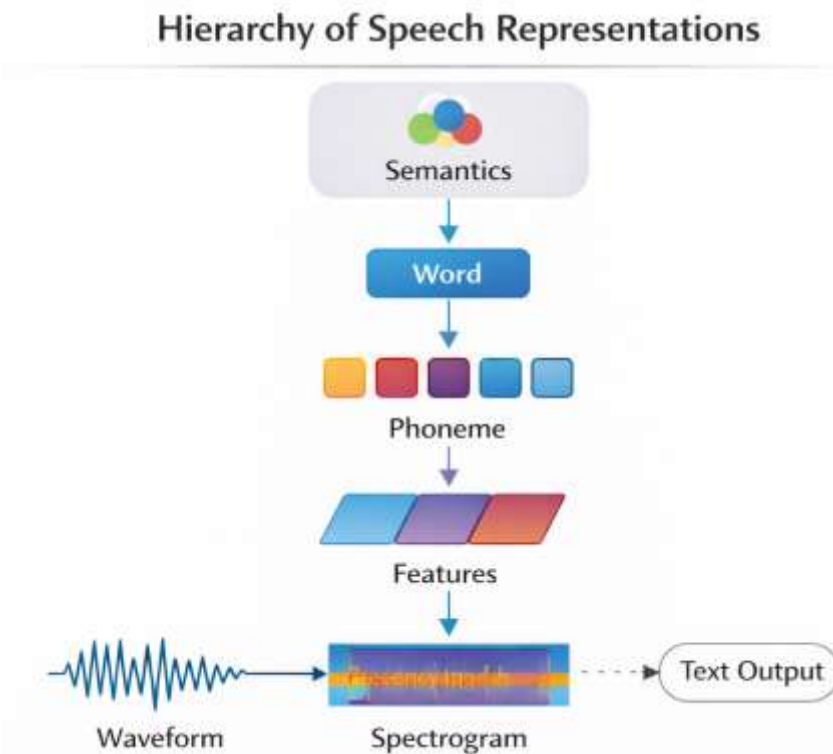


Figure 3. Hierarchical representation of speech signals from acoustic features to semantic understanding.

5.1. Neural Acoustic Modeling

One of the earliest successes of deep learning in speech processing was in acoustic modeling for automatic speech recognition (ASR). Deep neural networks (DNNs) significantly improved acoustic likelihood estimation compared to Gaussian mixture models [Hinton et al. 2012]. These models provide enhanced representational capacity and can capture complex acoustic variability.

Recurrent neural networks (RNNs), particularly long short-term memory (LSTM) architectures, further improved performance by modeling temporal dependencies in speech sequences [Graves et al. 2013, Sak et al. 2014]. These architectures allow modeling of long-range contextual information, which is critical for accurate recognition.

The introduction of attention-based encoder-decoder models marked a major shift toward end-to-end ASR systems [Li 2022, Prabhavalkar et al. 2024]. Models such as Listen, Attend and Spell (LAS) directly map acoustic inputs to output sequences without requiring explicit alignment [Chan et al. 2016]. Transformer architectures further extended this paradigm by leveraging self-attention mechanisms for efficient modeling of long-range dependencies [Vaswani et al. 2017]. Large-scale neural ASR systems have demonstrated near-human performance in conversational settings, highlighting the effectiveness of deep learning approaches [Xiong et al. 2018].

5.2. Neural Speech Synthesis

Deep learning has also transformed speech synthesis by enabling highly natural and expressive speech generation. Sequence-to-sequence models such as Tacotron directly map text to intermediate acoustic representations [Wang et al. 2017], while subsequent systems such as Tacotron 2 further improve stability and perceptual quality [Shen et al. 2018].

Compared with statistical parametric synthesis [Zen et al. 2009], neural TTS systems reduce artifacts such as over-smoothing and better capture prosodic variation. These models also provide greater flexibility in modeling speaker identity and speaking style. However, they typically require large-scale datasets and significant computational resources.

The transition from statistical to neural TTS highlights a broader shift from explicit modeling of acoustic parameters to implicit learning of representations from data.

5.3. Speech Enhancement and Separation

Deep learning has driven major advances in speech enhancement and separation, particularly in multi-speaker and noisy environments. Neural networks are commonly used to estimate time-frequency masks or directly reconstruct clean speech signals.

A key breakthrough in this area is Conv-TasNet, which demonstrated that time-domain end-to-end models can outperform traditional frequency-domain approaches [Luo and Mesgarani 2019]. These models learn representations directly from raw waveforms, enabling more effective separation of overlapping speech.

Permutation invariant training (PIT) addresses the label ambiguity problem in multi-speaker scenarios by allowing models to learn speaker-independent mappings [Yu et al. 2017]. Comprehensive analyses of deep learning-based speech separation provide further insights into model design and performance trade-offs [Wang and Chen 2018].

Recent work has explored attention mechanisms and embedding-based approaches to improve separation performance in challenging conditions [Thi et al. 2026, Tan et al. 2026]. In addition, survey studies provide systematic evaluations of deep learning approaches for speech separation and enhancement [Thi et al. 2025]. These works serve as representative examples of broader research trends.

5.4. Expanding Applications of Deep Speech Models

The flexibility of deep learning has enabled speech processing to expand beyond traditional tasks. One emerging direction is speech-based health monitoring, where acoustic signals are used to detect medical conditions. For example, deep learning models have been applied to snoring sound analysis for obstructive sleep apnea detection.

Another important direction is low-resource and multilingual speech processing. Deep learning models, particularly when combined with transfer learning, can leverage shared representations across languages and domains. This enables improved performance in under-resourced settings, although challenges related to data availability and domain mismatch remain [Phung et al. 2024].

5.5. Strengths and Limitations of Deep Learning

The deep learning paradigm offers several key advantages. First, it enables end-to-end optimization, allowing models to jointly learn representations and task-specific mappings. Second, it significantly improves performance across benchmarks when sufficient data is available. Third, it reduces reliance on handcrafted features, shifting the focus toward representation learning.

However, these benefits come with important limitations. Deep learning models are data-intensive and computationally expensive, making them difficult to deploy in resource-constrained environments. In addition, their internal representations are often difficult to interpret, raising challenges for transparency and trust.

Furthermore, while deep models excel at pattern recognition, they do not inherently capture higher-level semantic reasoning. This limitation motivates the transition toward more general representation learning frameworks.

Table 1, 2 summarizes the major paradigms in speech processing, highlighting their core modeling approaches and key limitations.

5.6. Transition to Self-Supervised Learning

The dependence on labeled data in supervised deep learning has motivated the development of self-supervised learning approaches. By leveraging large-scale unlabeled speech data, these methods aim to learn general-purpose representations that can be adapted to multiple tasks.

Early work such as wav2vec demonstrated the feasibility of learning representations directly from raw audio [Schneider et al. 2019]. Subsequent models such as wav2vec 2.0, HuBERT, and WavLM further improved performance and robustness across tasks [Baevski et al. 2020, Hsu et al. 2021, Chen et al. 2022].

This transition represents a natural progression from task-specific deep learning models to scalable representation learning frameworks, paving the way for foundation models and LLM-integrated speech systems.

Table 1. Conceptual summary of major paradigms in speech processing

Paradigm	Time	Features	Data
Signal processing	Pre-2005	MFCC, PLP	Low
Statistical	2005–2015	Engineered	Medium
Deep learning	2015–2020	Learned	High
Self-supervised	2020–2023	Raw speech	Very high
LLM-integrated	2023–	Multimodal	Massive

Table 2. Representative strengths and limitations of major paradigms

Paradigm	Strength	Limitation
Signal processing	Interpretable, efficient	Limited abstraction
Statistical	Probabilistic modeling	Independence assumptions
Deep learning	End-to-end learning	Data-intensive
Self-supervised	Strong generalization	High compute cost
LLM-integrated	Semantic reasoning	Modality gap

6. Self-Supervised and Foundation Speech Models

While deep learning significantly improved performance in speech processing, its reliance on large amounts of labeled data remained a fundamental limitation. Collecting and annotating speech data is expensive and time-consuming, particularly for low-resource languages and specialized domains. This limitation has driven the development of self-supervised learning (SSL), which enables models to learn meaningful representations from large-scale unlabeled speech data [Baevski et al. 2020, Hsu et al. 2021, Chen et al. 2022].

6.1. Self-Supervised Representation Learning

Self-supervised learning in speech processing is based on exploiting intrinsic structure in the signal to define learning objectives without manual labels [Mohamed et al. 2022]. Early approaches such as wav2vec demonstrated that models can learn useful representations directly from raw waveforms by predicting contextual information [Schneider et al. 2019]. This line of work builds upon earlier representation learning paradigms in deep learning while addressing the limitations of supervised training.

Wav2vec 2.0 introduced a masked prediction framework combined with contrastive learning, significantly improving representation quality and downstream task performance [Baevski et al. 2020]. In this approach, segments of the input are masked, and the model learns to predict latent representations, encouraging contextual understanding of speech.

Subsequent models further refined this paradigm. HuBERT leverages iterative clustering to generate pseudo-labels, enabling the model to capture both acoustic and linguistic structure [Hsu et al. 2021]. WavLM extends this approach by incorporating denoising objectives and robustness-oriented training strategies, improving performance across tasks such as speaker verification, speech separation, and ASR [Chen et al. 2022].

6.2. Scaling with Unlabeled Data

A key advantage of self-supervised learning is its ability to leverage large-scale unlabeled datasets. Studies have shown that increasing the scale of pretraining data leads to substantial improvements in representation quality and downstream performance.

Large-scale datasets such as Libri-Light have enabled training on tens of thousands of hours of unlabeled speech, significantly improving model generalization [Kahn et al. 2020]. Similarly, multilingual pretraining approaches such as XLS-R demonstrate the ability to learn cross-lingual representations that generalize across languages, improving performance in low-resource scenarios [Babu et al. 2022].

These developments highlight a shift from task-specific optimization toward large-scale pretraining, where models learn general-purpose representations that can be adapted to a wide range of tasks.

6.3. From Representation Learning to Foundation Models

Self-supervised learning has enabled the emergence of speech foundation models, which serve as general-purpose backbones for multiple speech tasks. Unlike traditional models trained for specific tasks, foundation models provide reusable representations that can be fine-tuned or adapted for different applications.

Recent systems such as Whisper demonstrate the effectiveness of large-scale pre-training for robust speech recognition across diverse languages and acoustic conditions [Radford et al. 2022]. Similarly, unified architectures such as SpeechT5 integrate speech and text modeling within a single framework, enabling both recognition and generation tasks [Ao et al. 2022].

These models reflect a broader trend toward unifying speech and language representations, bridging the gap between acoustic signals and linguistic structure.

6.4. Generalization and Transfer Learning

One of the key strengths of SSL-based models is their ability to generalize across domains and tasks. By learning representations from diverse and large-scale datasets, these models can be adapted to new conditions with relatively small amounts of labeled data.

This capability is particularly important for low-resource languages and domain-specific applications, where labeled data is scarce. Self-supervised models provide a strong foundation for transfer learning, enabling improved performance in under-resourced settings [Phung et al. 2024].

However, domain mismatch remains a challenge. Differences in recording conditions, speaker characteristics, and language distributions can still lead to performance degradation, highlighting the need for more robust and adaptive representations.

6.5. Limitations of Self-Supervised Learning

Despite their success, self-supervised speech models have several limitations. First, their performance depends heavily on the scale and diversity of pretraining data. Biases in the data can lead to uneven performance across languages and speaker groups.

Second, SSL models are typically optimized using surrogate objectives, such as reconstruction or contrastive learning, which may not align directly with downstream tasks. As a result, fine-tuning remains necessary, and performance can vary depending on adaptation strategies.

Third, while SSL models capture acoustic and phonetic structure effectively, they do not inherently model higher-level semantic reasoning. This limitation becomes critical when moving toward applications that require understanding, reasoning, or interaction.

6.6. Bridging Toward Speech-Language Models

The emergence of self-supervised and foundation models has created a natural pathway toward integrating speech processing with language models. By providing high-quality acoustic representations, SSL models facilitate alignment between speech and text.

Recent work on generative and multimodal systems, such as AudioLM and related approaches, demonstrates how speech can be modeled using discrete representations and language modeling techniques [Borsos et al. 2023]. Similarly, advances in speech-language modeling highlight the potential for unified systems that jointly process speech and text [Zhang et al. 2023, Wang et al. 2023].

These developments suggest a convergence between speech processing and language modeling, paving the way for LLM-integrated speech intelligence systems, which will be discussed in the next section.

7. LLM-Integrated Speech Systems and Modality Bridging

The rapid development of large language models (LLMs) has introduced a new paradigm in speech processing, shifting the focus from signal transformation and representation learning toward semantic understanding and multimodal reasoning. While traditional speech systems primarily aim to convert speech into text, emerging research increasingly treats speech as a rich modality that should be directly integrated with higher-level reasoning systems [Radford et al. 2021, Borsos et al. 2023, Zhang et al. 2023]. The fundamental gap between continuous speech signals and discrete text representations is illustrated in Fig. 4.

7.1. From Pipeline Architectures to Integrated Systems

Conventional speech processing systems are typically organized as pipelines, where speech is first transcribed into text using automatic speech recognition (ASR), and the resulting text is then processed by downstream models. This paradigm has been highly successful, particularly with the integration of deep learning-based ASR systems [Hinton et al. 2012, Chan et al. 2016, Vaswani et al. 2017].

However, this pipeline introduces several limitations. Transcription acts as a bottleneck, discarding paralinguistic information such as prosody, emotion, and speaker characteristics. In addition, recognition errors propagate to downstream tasks, affecting overall system performance. These limitations have motivated the development of integrated architectures that aim to connect speech processing directly with language modeling.

Recent work has explored large-scale speech recognition models that improve robustness and generalization across languages and acoustic conditions [Radford et al. 2022]. At the same time, unified frameworks aim to integrate speech and text processing within a single architecture, reducing reliance on intermediate representations [Ao et al. 2022].

7.2. Speech-Text Alignment and Representation Gap

A central challenge in LLM-integrated speech systems is the alignment between speech and text representations. Speech signals are continuous and high-dimensional, while text is discrete and symbolic. Bridging this gap requires mapping between fundamentally different representational spaces.

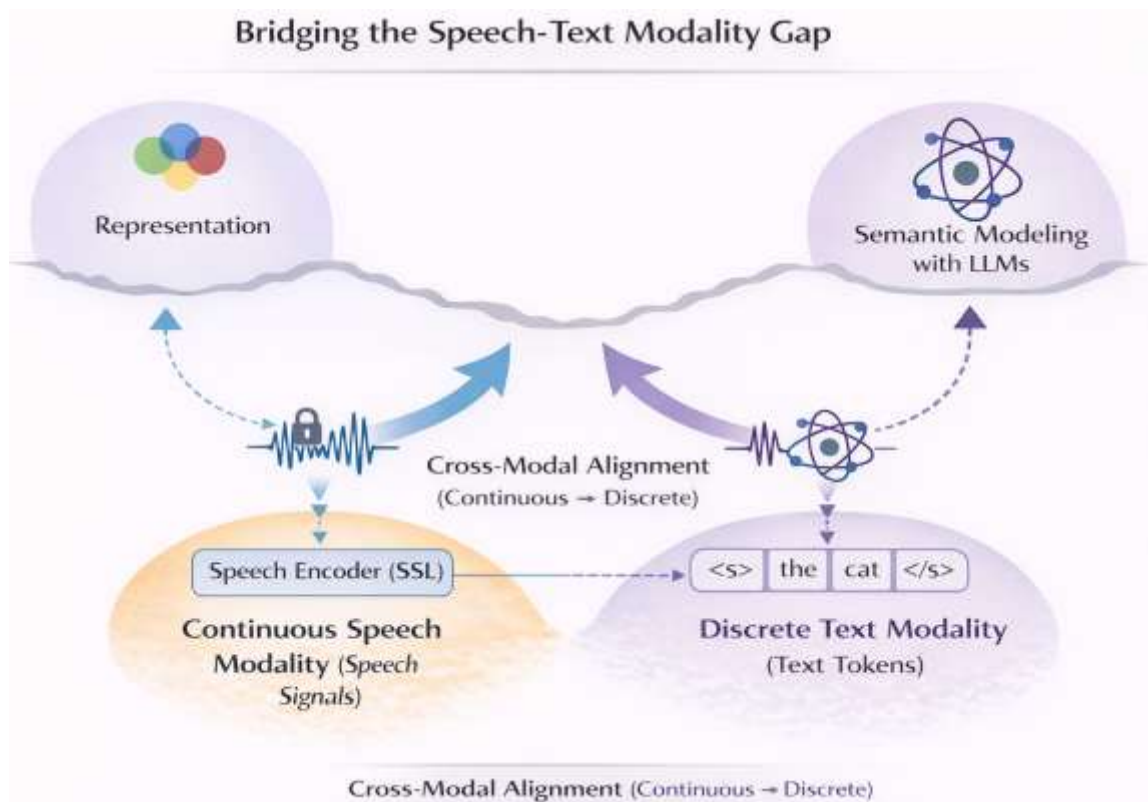


Figure 4. Bridging the modality gap between continuous speech signals and discrete text representations through cross-modal alignment.

Multimodal learning frameworks provide one approach to this problem. For example, contrastive learning methods such as CLIP demonstrate how different modalities can be aligned within a shared embedding space [Radford et al. 2021]. Similar ideas have been extended to speech-text alignment, where models learn joint representations that connect acoustic patterns with linguistic units.

Generative models such as AudioLM further bridge this gap by representing speech as discrete tokens that can be modeled using language modeling techniques [Borsos et al. 2023]. Extensions of this approach, including VALL-E, demonstrate the ability to synthesize high-quality speech by leveraging discrete acoustic representations [Wang et al. 2023]. These models suggest that speech can be treated as a sequence modeling problem, analogous to text.

7.3. Speech-Language Models and Multimodal Reasoning

Speech-language models aim to jointly process speech and text within unified architectures. These models combine pretrained speech encoders with large language models, enabling cross-modal reasoning and interaction.

Recent systems such as SpeechGPT illustrate how language models can be extended to handle speech inputs, enabling tasks such as spoken dialogue, question answering, and speech summarization [Zhang et al. 2023]. These models treat speech as an input modality to language models, allowing semantic reasoning over spoken content.

In parallel, advances in speech foundation models provide high-quality representations that facilitate integration with language models [Baevski et al. 2020, Chen et al. 2022]. These representations enable more effective alignment between acoustic and linguistic information, supporting multimodal learning.

7.4. Beyond Transcription: Speech as a Rich Modality

Traditional speech processing systems focus primarily on transcription, reducing speech to textual representations. However, speech conveys significantly more information than text alone, including prosody, timing, speaker identity, and emotional cues.

LLM-integrated systems open the possibility of leveraging these additional dimensions. By incorporating acoustic features into multimodal representations, systems can capture aspects of communication that are lost in text-only processing. This is particularly important for applications such as conversational AI, affective computing, and human-computer interaction.

The integration of speech with multimodal learning frameworks suggests a shift from speech processing toward broader speech intelligence systems that model communication holistically.

7.5. Revisiting Classical and Statistical Insights

Interestingly, many challenges in LLM-integrated systems echo earlier issues encountered in classical and statistical paradigms. For example, robustness to noise and variability remains a critical concern, as highlighted in early signal processing approaches [Boll 1979, Ephraim and Malah 1984]. Similarly, modeling temporal dynamics continues to rely on principles introduced in statistical frameworks such as HMMs [Rabiner 1989, Gales and Young 2008].

In speech synthesis, issues such as over-smoothing and perceptual quality, previously studied in statistical parametric models [Nghia et al. 2016, Nguyen and Phung 2017], reappear in neural and generative models, albeit in different forms. These connections highlight the continuity of research challenges across paradigms.

7.6. Challenges in LLM-Integrated Speech Systems

Despite recent progress, several challenges remain in integrating speech with large language models:

- Information loss: Transcription-based pipelines discard important acoustic information.
- Representation alignment: Bridging continuous speech signals and discrete text remains inherently difficult.
- Scalability: Multimodal models require large-scale datasets and computational resources.
- Latency: Real-time applications impose strict constraints on model size and inference speed.

In addition, domain-specific challenges such as low-resource languages and specialized applications continue to require targeted solutions [Phung et al. 2024].

Table 3. Illustrative evolution of key speech processing tasks across paradigms

Task	Evolution
ASR	HMM-GMM → End-to-end → SSL (wav2vec, HuBERT) → Multimodal LLM
TTS	Unit selection / HMM-TTS → Tacotron → Neural vocoders → Generative speech models
Speech Separation	Spectral masking → Deep models (TasNet) → SSL-enhanced → Multimodal scene understanding
Speaker Recognition	GMM-UBM → x-vector → SSL embeddings → Multimodal identity modeling
Speech Understanding	Rule-based → Neural NLU → Unified embeddings → LLM reasoning

7.7. Discussion

The integration of LLMs represents a shift from task-specific speech processing toward multimodal intelligence. Rather than focusing solely on recognition or synthesis, modern systems aim to understand and interact with spoken language in context.

This transition reflects a broader convergence between speech processing and language modeling. By combining advances in representation learning, multimodal alignment, and generative modeling, LLM-integrated systems have the potential to redefine the role of speech in artificial intelligence.

However, fully bridging the modality gap between speech and text remains an open problem. Addressing this challenge will require new modeling paradigms that jointly capture acoustic, linguistic, and semantic information, building upon insights from both classical and modern approaches.

8. Cross-Cutting Challenges in Speech Processing

Despite substantial progress across multiple paradigms, several fundamental challenges have persisted throughout the evolution of speech processing. These challenges are not specific to any single methodology, but instead reflect intrinsic properties of speech as a complex, variable, and multimodal signal. Understanding these cross-cutting issues provides a unifying perspective on the field and highlights directions for future research.

Table 3 provides an illustrative summary of how selected speech processing tasks have evolved across different paradigms, from classical approaches to LLM-integrated systems.

8.1. Robustness to Noise and Environmental Variability

Robustness to noise and environmental variability has been a central challenge from classical signal processing to modern deep learning systems [Yousif and Mahmmod 2025]. Early approaches addressed this issue through explicit signal enhancement techniques such as spectral subtraction and MMSE estimation [Boll 1979, Ephraim and Malah 1984]. Statistical models later incorporated noise modeling into probabilistic frameworks [Rabiner 1989, Gales and Young 2008], while deep learning methods rely on large-scale training and data augmentation [Hinton et al. 2012, Luo and Mesgarani 2019].

8.2. Multi-Speaker Scenarios and Source Separation

The cocktail party problem remains one of the most challenging tasks in speech processing [Michelsanti et al. 2021]. Classical approaches provided limited solutions, while statistical models struggled to represent overlapping speech signals.

Deep learning has significantly advanced this area, with models such as Conv-TasNet achieving strong performance in speech separation tasks [Luo and Mesgarani 2019]. Permutation invariant training further addressed the label ambiguity problem in multi-speaker scenarios [Yu et al. 2017]. Comprehensive analyses highlight both the progress and limitations of current approaches [Wang and Chen 2018].

Recent work has explored attention mechanisms and embedding-based approaches to improve separation in complex environments [Thi et al. 2026, Tan et al. 2026], while survey studies provide systematic evaluations of deep learning techniques [Thi et al. 2025]. Despite these advances, robust separation in highly dynamic real-world scenarios remains unresolved.

8.3. Data Scarcity and Generalization

Although modern systems benefit from large-scale datasets, data scarcity remains a critical issue for many languages and domains [Mohamed et al. 2022, Kheddar et al. 2024]. Low-resource languages and domain-specific applications often lack sufficient labeled data for effective training.

Self-supervised learning has alleviated this problem by enabling representation learning from unlabeled data [Schneider et al. 2019, Baevski et al. 2020, Hsu et al. 2021, Chen et al. 2022]. Large-scale datasets such as Libri-Light further improve generalization [Kahn et al. 2020], while multilingual approaches such as XLS-R enable cross-lingual transfer [Babu et al. 2022].

However, domain mismatch remains a persistent challenge. Differences in acoustic conditions, speaker characteristics, and linguistic distributions can significantly affect performance [Phung et al. 2024].

8.4. Evaluation Metrics and Perceptual Quality

Evaluating speech processing systems remains a complex problem. Objective metrics such as word error rate (WER) and signal-to-distortion ratio (SDR) provide standardized benchmarks, but they do not always correlate with human perception.

In speech synthesis, statistical approaches have long struggled with perceptual issues such as over-smoothing [Zen et al. 2009, Nghia et al. 2016, Nguyen and Phung 2017]. While neural approaches improve naturalness [Shen et al. 2018], discrepancies between objective and subjective evaluation persist.

These challenges highlight the need for evaluation frameworks that better capture perceptual quality and user experience.

8.5. Interpretability and Model Transparency

As speech processing systems become increasingly complex, interpretability has emerged as an important concern. Classical signal processing methods are inherently interpretable, with clearly defined transformations and parameters [Oppenheim and Schaffer 2010]. In contrast, deep learning models operate as high-dimensional nonlinear systems, making it difficult to understand their internal behavior [Hinton et al. 2012].

This lack of transparency poses challenges for debugging, trust, and deployment in critical applications such as healthcare and security. Developing interpretable models and analysis tools remains an important research direction.

8.6. Computational Efficiency and Scalability

Modern speech processing systems, particularly those based on deep learning and foundation models, require substantial computational resources. Large-scale pretraining and inference can be computationally expensive, limiting deployment in real-time and resource-constrained environments.

While advances in model design and hardware acceleration have improved efficiency, balancing performance and computational cost remains a key challenge. This issue is particularly important for applications such as edge devices and mobile systems.

8.7. Bridging Speech, Text, and Semantics

Perhaps the most fundamental challenge in modern speech processing is bridging the gap between acoustic signals and semantic understanding. While significant progress has been made in mapping speech to text, true understanding of spoken language requires capturing context, intent, and implicit meaning.

Recent work on multimodal and speech-language models highlights the potential for integrating speech with language models [Radford et al. 2021, Borsos et al. 2023, Zhang et al. 2023, Wang et al. 2023, Zhu et al. 2021, Sheng et al. 2024, Sethiya and Maurya 2025]. However, aligning continuous speech representations with discrete semantic structures remains difficult.

Generative models such as AudioLM demonstrate how speech can be modeled using discrete tokens [Borsos et al. 2023], while unified architectures such as SpeechT5 aim to bridge speech and text within a single framework [Ao et al. 2022]. Despite these advances, fully resolving the modality gap remains an open problem.

8.8. Discussion

The challenges outlined above highlight that progress in speech processing is not solely driven by advances in individual paradigms, but also by the ability to address fundamental limitations that persist across methodologies.

Importantly, these challenges are interconnected. Robustness depends on representation learning, which in turn depends on data availability and model design. Similarly, bridging speech and semantics requires advances in both acoustic modeling and language understanding.

Addressing these challenges will require integrated approaches that combine insights from signal processing, statistical modeling, deep learning, and multimodal learning, paving the way toward more robust and intelligent speech systems.

9. Future Directions in Speech Intelligence

The evolution of speech processing over the past two decades suggests that the field is entering a new phase characterized by the convergence of speech, language, and multi-modal intelligence. While substantial progress has been achieved across signal processing, statistical modeling, and deep learning paradigms, several emerging directions are likely to shape the next generation of speech systems.

9.1. Toward Unified Speech Foundation Models

One of the most important trends is the development of unified speech foundation models that support multiple tasks within a single architecture [Mohamed et al. 2022, Prabhavalkar et al. 2024]. Early systems relied on task-specific pipelines, while statistical approaches introduced modular probabilistic frameworks [Rabiner 1989, Gales and Young 2008]. Deep learning further enabled end-to-end modeling [Hinton et al. 2012, Vaswani et al. 2017], but models remained largely task-dependent.

Self-supervised learning has fundamentally changed this paradigm by enabling large-scale pretraining on unlabeled data [Baeovski et al. 2020, Hsu et al. 2021, Chen et al. 2022]. These models provide general-purpose representations that can be adapted across tasks, languages, and domains. Future research is expected to focus on scaling these approaches and integrating them into unified architectures that jointly model speech, text, and other modalities.

9.2. Speech-Native Interaction with Language Models

Most existing systems still rely on a pipeline in which speech is transcribed into text before being processed by language models. While effective, this approach introduces information loss and limits the ability to capture the full richness of speech signals.

A promising direction is the development of speech-native interfaces for language models, where speech is processed directly without intermediate transcription. Recent advances in large language models [OpenAI 2023] and speech-language systems [Zhang et al. 2023, Borsos et al. 2023, Wang et al. 2023] suggest that such integration is feasible. These approaches build upon earlier work in speech synthesis and sequence modeling [Taylor 2009, Black and Tokuda 2005], while extending them toward unified generative frameworks.

9.3. Multimodal Integration and Cross-Modal Learning

Speech is inherently multimodal, often accompanied by visual and contextual information [Zhu et al. 2021, Sheng et al. 2024]. Future systems are likely to integrate speech with other modalities, enabling richer and more robust understanding.

Multimodal learning frameworks such as CLIP demonstrate the effectiveness of joint representation learning across modalities [Radford et al. 2021]. Extending these approaches to speech could enable applications such as audiovisual speech recognition, multimodal dialogue systems, and context-aware assistants.

9.4. Robustness, Generalization, and Low-Resource Learning

Despite advances in large-scale models, robustness and generalization remain key challenges [Yousif and Mahmmod 2025, Kheddar et al. 2024, Ahlawat et al. 2025]. Classical approaches addressed noise through signal enhancement [Boll 1979, Ephraim and Malah 1984], while statistical models incorporated variability through probabilistic frameworks [Rabiner 1989]. Deep learning and self-supervised models improve generalization but still struggle under domain mismatch.

Future work is expected to focus on improving robustness across speakers, environments, and languages. Multilingual and cross-lingual approaches such as XLS-R [Babu et al. 2022] demonstrate the potential of shared representations, while low-resource studies highlight the importance of transfer learning [Phung et al. 2024].

9.5. Efficient and Scalable Speech Systems

As models continue to scale, computational efficiency becomes increasingly important. Classical systems were designed for efficiency and real-time processing, while modern neural models require substantial computational resources.

Future research will need to balance performance and efficiency, particularly for deployment on edge devices. This includes model compression, efficient architectures, and hardware-aware optimization strategies.

9.6. Toward Speech Intelligence

The long-term trajectory of the field points toward a broader concept of speech intelligence, where systems move beyond signal processing to understanding and interacting with spoken language in a human-like manner.

This vision involves integrating perception, representation, reasoning, and generation within unified frameworks. It builds upon decades of research, from signal processing and statistical modeling to deep learning and LLM-integrated systems.

Importantly, many challenges identified in earlier paradigms—such as robustness, variability, and perceptual quality—remain relevant. Issues such as over-smoothing in statistical synthesis [Nghia et al. 2016, Nguyen and Phung 2017] and limitations in representation learning continue to influence modern systems.

Discussion

The future of speech processing lies in the convergence of multiple paradigms. Rather than replacing earlier approaches, modern systems build upon insights from signal processing, statistical modeling, and machine learning.

By integrating these perspectives, future research can move toward more robust, efficient, and intelligent systems that better capture the complexity of human communication. Bridging the gap between speech, text, and semantics remains a central challenge, but also a key opportunity for advancing the field. As a narrative and conceptual review, this paper has several limitations. It does not follow a systematic search and screening protocol, and the cited studies are not intended to represent an exhaustive corpus of speech processing research. Instead, the review prioritizes representative works that help clarify major conceptual transitions in the field. Future systematic or scoping reviews could complement this perspective by mapping the literature quantitatively across databases, tasks, datasets, and evaluation protocols.

10. Conclusion

This paper has presented a narrative and conceptual review of the evolution of speech processing, tracing the progression from classical signal processing to statistical modeling, deep learning, self-supervised learning, and, more recently, LLM-integrated speech intelligence. Across these paradigms, the field has undergone a fundamental transformation in how speech is represented, modeled, and utilized in intelligent systems.

A key insight from this review is that advances in speech processing are closely tied to shifts in representation. Early approaches relied on handcrafted features and signal transformations, while later paradigms progressively moved toward learned representations with increasing levels of abstraction. Self-supervised and foundation models further generalized this trend by enabling scalable representation learning from large-scale unlabeled data. The integration of speech with large language models represents the next step in this evolution, extending speech processing from perception to semantic reasoning and multimodal interaction.

Despite substantial progress, several fundamental challenges remain unresolved. Issues such as robustness to real-world environments, multi-speaker interference, data scarcity, evaluation limitations, and computational efficiency continue to constrain system performance. In addition, the modality gap between speech and text presents a central challenge for LLM-integrated systems, highlighting the need for new modeling paradigms that can jointly capture acoustic, linguistic, and semantic information.

Looking forward, the field is moving toward unified and multimodal frameworks that integrate speech with language, vision, and other modalities. This transition reflects a broader shift from task-specific systems toward general-purpose speech intelligence. Achieving this vision will require advances in representation learning, multimodal modeling, efficient computation, and system integration.

In summary, this narrative synthesis suggests that the evolution from signal processing to LLM-integrated speech intelligence illustrates both the progress and the conceptual complexity of the field. By building on established foundations while addressing persistent challenges, future research can enable more robust, flexible, and intelligent speech systems that better support real-world communication.

Declarations

Funding

This research received no external funding.

Competing Interests

The authors declare that they have no competing interests.

Data Availability

No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Research Involving Human and/or Animals

This article is a review study and does not involve human participants or animals.

Informed Consent

Not applicable.

Use of AI Tools

The authors used ChatGPT (OpenAI) to assist with language editing and drafting. All content was reviewed, verified, and approved by the authors, who take full responsibility for the manuscript.

References

1. [Purwins et al. 2019] Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.-Y., Sainath, T. N.: "Deep Learning for Audio Signal Processing"; IEEE Journal of Selected Topics in Signal Processing 13, 2 (2019), 206–219. <https://doi.org/10.1109/JSTSP.2019.2908700>.
2. [Patra & Singh (2026)]. Technostress Among Farmers Using AI-Based Agricultural Technologies: A Conceptual Framework. Journal of Artificial Intelligence in Agriculture and Applied Research, 2, 1(2026), 1–12.
3. [Mehrish et al. 2023] Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., Poria, S.: "A Review of Deep Learning Techniques for Speech Processing"; Information Fusion 99 (2023), 101869. <https://doi.org/10.1016/j.inffus.2023.101869>.
4. [Davis and Mermelstein 1980] Davis, S. B., Mermelstein, P.: "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences"; IEEE Transactions on Acoustics, Speech, and Signal Processing 28, 4 (1980), 357–366. <https://doi.org/10.1109/TASSP.1980.1163420>.
5. [Hermansky 1990] Hermansky, H.: "Perceptual Linear Predictive (PLP) Analysis of Speech"; The Journal of the Acoustical Society of America 87, 4 (1990), 1738–1752. <https://doi.org/10.1121/1.399423>.
6. [Boll 1979] Boll, S. F.: "Suppression of Acoustic Noise in Speech Using Spectral Subtraction"; IEEE Transactions on Acoustics, Speech, and Signal Processing 27, 2 (1979), 113–120. <https://doi.org/10.1109/TASSP.1979.1163209>.
7. [Ephraim and Malah 1984] Ephraim, Y., Malah, D.: "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator"; IEEE Transactions on Acoustics, Speech, and Signal Processing 32, 6 (1984), 1109–1121. <https://doi.org/10.1109/TASSP.1984.1164453>.
8. [Reynolds 1995] Reynolds, D. A.: "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models"; Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (1995). <https://doi.org/10.1109/ICASSP.1995.479394>.
9. [Rabiner 1989] Rabiner, L. R.: "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition"; Proceedings of the IEEE 77, 2 (1989), 257–286. <https://doi.org/10.1109/5.18626>.
10. [Gales and Young 2008] Gales, M., Young, S.: "The Application of Hidden Markov Models in Speech Recognition"; Foundations and Trends in Signal Processing 1, 3 (2008), 195–304. <https://doi.org/10.1561/2000000004>.
11. [Tokuda et al. 2000] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T.: "Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis"; Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 3, pp. 1315–1318 (2000). <https://doi.org/10.1109/ICASSP.2000.861820>.
12. [Zen et al. 2009] Zen, H., Tokuda, K., Black, A. W.: "Statistical Parametric Speech Synthesis"; Speech Communication 51, 11 (2009), 1039–1064. <https://doi.org/10.1016/j.specom.2009.04.004>.
13. [Hunt and Black 1996] Hunt, A. J., Black, A. W.: "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database"; Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp. 373–376 (1996). <https://doi.org/10.1109/ICASSP.1996.541110>.
14. [Stylianou 2009] Stylianou, Y.: "Voice Transformation: A Survey"; Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3585–3588 (2009). <https://doi.org/10.1109/ICASSP.2009.4960401>.
15. [Nassif et al. 2019] Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., Shaalan, K.: "Speech Recognition Using Deep Neural Networks: A Systematic Review"; IEEE Access 7 (2019), 19143–19165. <https://doi.org/10.1109/ACCESS.2019.2896880>.
16. [Hinton et al. 2012] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., Kingsbury, B.: "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups"; IEEE Signal Processing Magazine 29, 6 (2012), 82–97. <https://doi.org/10.1109/MSP.2012.2205597>.
17. [Graves et al. 2013] Graves, A., Mohamed, A., Hinton, G.: "Speech Recognition with Deep Recurrent Neural Networks"; Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6645–6649 (2013). <https://doi.org/10.1109/ICASSP.2013.6638947>.
18. [Sak et al. 2014] Sak, H., Senior, A., Beaufays, F.: "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling"; Proceedings of Interspeech 2014, pp. 338–342 (2014).

- <https://doi.org/10.21437/Interspeech.2014-80>.
19. [Chan et al. 2016] Chan, W., Jaitly, N., Le, Q. V., Vinyals, O.: "Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition"; Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4960– 4964 (2016).
<https://doi.org/10.1109/ICASSP.2016.7472621>.
20. [Li 2022] Li, J.: "Recent Advances in End-to-End Automatic Speech Recognition"; APSIPA Transactions on Signal and Information Processing (2022). <https://doi.org/10.1561/116.00000050>.
21. [Prabhavalkar et al. 2024] Prabhavalkar, R., Hori, T., Sainath, T. N., Schlüter, R., Watanabe, S.: "End-to-End Speech Recognition: A Survey"; IEEE/ACM Transactions on Audio, Speech, and Language Processing 32 (2024), 325– 351. <https://doi.org/10.1109/TASLP.2023.3328283>.
22. [Vaswani et al. 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I.: "Attention Is All You Need"; Advances in Neural Information Processing Systems, vol. 30, pp. 5998– 6008 (2017). <https://papers.nips.cc/paper/7181-attention-is-all-you-need>.
23. [Kheddar et al. 2024] Kheddar, H., Hemis, M., Himeur, Y.: "Automatic Speech Recognition Using Advanced Deep Learning Approaches: A Survey"; Information Fusion 109 (2024), 102422.
<https://doi.org/10.1016/j.inffus.2024.102422>.
24. [Wang et al. 2017] Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q. V., Agiomyrgiannakis, Y., Clark, R., Saurous, R. A.: "Tacotron: Towards End-to-End Speech Synthesis"; Proceedings of Interspeech 2017, pp. 4006–4010 (2017).
<https://doi.org/10.21437/Interspeech.2017-1452>.
25. [Shen et al. 2018] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y., Wu, Y.: "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions"; Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4779– 4783 (2018). <https://doi.org/10.1109/ICASSP.2018.8461368>.
26. [Luo and Mesgarani 2019] Luo, Y., Mesgarani, N.: "Conv-TasNet: Surpassing Ideal Time– Frequency Magnitude Masking for Speech Separation"; IEEE/ACM Transactions on Audio, Speech, and Language Processing 27, 8 (2019), 1256–1266. <https://doi.org/10.1109/TASLP.2019.2915167>.
27. [Yu et al. 2017] Yu, D., Kolbæk, M., Tan, Z.-H., Jensen, J.: "Permutation Invariant Training of Deep Models for Speaker-Independent Multi-Talker Speech Separation"; Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2017). <https://doi.org/10.1109/ICASSP.2017.7952154>.
28. [Mohamed et al. 2022] Mohamed, A., Lee, H.-Y., Borgholt, L., Havtorn, J. D., Edin, J., Igel, C., Kirchhoff, K., Li, S.-W., Livescu, K., Maaløe, L., Sainath, T. N., Watanabe, S.: "Self-Supervised Speech Representation Learning: A Review"; IEEE Journal of Selected Topics in Signal Processing 16, 6 (2022), 1179–1210.
<https://doi.org/10.1109/JSTSP.2022.3207050>.
29. [Schneider et al. 2019] Schneider, S., Baevski, A., Collobert, R., Auli, M.: "wav2vec: Unsupervised Pre-Training for Speech Recognition"; Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3465–3469 (2019). <https://doi.org/10.1109/ICASSP.2019.8682335>.
30. [Baevski et al. 2020] Baevski, A., Zhou, H., Mohamed, A., Auli, M.: "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations"; Advances in Neural Information Processing Systems, vol. 33, pp. 12449– 12460 (2020).
<https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html>.
31. [Hsu et al. 2021] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhota, K., Salakhutdinov, R., Mohamed, A.: "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units"; IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021), 3451–3460.
<https://doi.org/10.1109/TASLP.2021.3122291>.
32. [Chen et al. 2022] Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Zeng, M., Yu, X., Wei, F.: "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing"; IEEE Journal of Selected Topics in Signal Processing 16, 6 (2022), 1505–1518.
<https://doi.org/10.1109/JSTSP.2022.3188113>.
33. [Zhu et al. 2021] Zhu, H., Luo, M.-D., Wang, R., Zheng, A.-H., He, R.: "Deep Audio-Visual Learning: A Survey"; Machine Intelligence Research 18, 3 (2021), 351–376. <https://doi.org/10.1007/s11633-021-1293-0>.
34. [Sheng et al. 2024] Sheng, C., Kuang, G., Bai, L., Hou, C., Guo, Y., Xu, X., Pietikäinen, M., Liu, L.: "Deep Learning for Visual Speech Analysis: A Survey"; IEEE Transactions on Pattern Analysis and Machine Intelligence 46, 9 (2024), 6001–6022. <https://doi.org/10.1109/TPAMI.2024.3376710>.
35. [Radford et al. 2021] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: "Learning Transferable Visual Models From Natural Language Supervision"; Proceedings of the 38th International Conference on Machine Learning, vol. 139, pp. 8748–8763 (2021). <https://proceedings.mlr.press/v139/radford21a.html>.
36. [Borsos et al. 2023] Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Roblek, D., Teboul, O., Grangier, D., Tagliasacchi, M., Zeghidour, N.: "AudioLM: A Language Modeling Approach to Audio Generation"; IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2023), 2523–2533.
<https://doi.org/10.1109/TASLP.2023.3288409>.
37. [Wang et al. 2023] Wang, S., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S., Wei, F.: "Neural Codec Language Models Are Zero-Shot Text to Speech Synthesizers"; arXiv preprint

- arXiv:2301.02111, (2023). <https://www.microsoft.com/en-us/research/publication/neural-codec-language-models-are-zero-shot-text-to-speech-synthesizers/>.
38. [Zhang et al. 2023] Zhang, D., Li, S., Zhang, X., Zhan, J., Wang, P., Zhou, Y., Qiu, X.: "SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities"; Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 15757– 15773 (2023). <https://aclanthology.org/2023.findings-emnlp.1055/>.
39. [Malik et al. 2021] Malik, M., Malik, M. K., Mehmood, K., Makhdoom, I.: "Automatic Speech Recognition: A Survey"; Multimedia Tools and Applications 80, 6 (2021), 9411–9457. <https://doi.org/10.1007/s11042-020-10073-7>.
40. [Yousif and Mahmmod 2025] Yousif, S. T., Mahmmod, B. M.: "Speech Enhancement Algorithms: A Systematic Literature Review"; Algorithms 18, 5 (2025), 272. <https://doi.org/10.3390/a18050272>.
41. [Michelsanti et al. 2021] Michelsanti, D., Tan, Z.-H., Zhang, S.-X., Xu, Y., Yu, M., Yu, D., Jensen, J.: "An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation"; IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021), 1368– 1396. <https://doi.org/10.1109/TASLP.2021.3066303>.
42. [Hansen and Hasan 2015] Hansen, J. H. L., Hasan, T.: "Speaker Recognition by Machines and Humans: A Tutorial Review"; IEEE Signal Processing Magazine 32, 6 (2015), 74–99. <https://doi.org/10.1109/MSP.2015.2462851>.
43. [Anguera et al. 2012] Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., Vinyals, O.: "Speaker Diarization: A Review of Recent Research"; IEEE Transactions on Audio, Speech, and Language Processing 20, 2 (2012), 356–370. <https://doi.org/10.1109/TASL.2011.2125954>.
44. [O'Shaughnessy 2025] O'Shaughnessy, D.: "Speaker Diarization: A Review of Objectives and Methods"; Applied Sciences (2025). <https://doi.org/10.3390/app15042002>.
45. [Sethiya and Maurya 2025] Sethiya, N., Maurya, C. K.: "End-to-End Speech-to-Text Translation: A Survey"; Computer Speech and Language (2025). <https://doi.org/10.1016/j.csl.2024.101751>.
46. [Latif et al. 2023] Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., Schuller, B.: "A Survey of Deep Representation Learning for Speech Emotion Recognition"; IEEE Transactions on Affective Computing 14, 2 (2023), 1634–1654. <https://doi.org/10.1109/TAFFC.2021.3114365>.
47. [Phung et al. 2024] Phung, T.-N., Nguyen, D., Pham, N.: "A Review on Speech Recognition for Under-Resourced Languages: A Case Study of Vietnamese"; International Journal of Knowledge and Systems Science 15, 1 (2024), 1–16. <https://doi.org/10.4018/IJKSS.332869>.
48. [Ao et al. 2022] Ao, J., Wang, R., Zhou, L., Wang, C., Ren, S., Wu, Y., Liu, S., Ko, T., Li, Q., Zhang, Y., Wei, Z., Qian, Y., Li, J., Wei, F.: "SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing"; Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 5723–5738 (2022). <https://doi.org/10.18653/v1/2022.acl-long.393>.
49. [Bilmes 1998] Bilmes, J. A.: "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models"; Technical Report ICSITR-97-021, International Computer Science Institute, (1998). <https://people.ece.uw.edu/bilmes/p/pgs/b2hd-bilmes1997-em.html>.
50. [Bishop 2006] Bishop, C. M.: "Pattern Recognition and Machine Learning"; Springer, New York, (2006). <https://link.springer.com/book/9780387310732>.
51. [Young et al. 2006] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: "The HTK Book (for HTK Version 3.4)"; Technical Report, Cambridge University Engineering Department, (2006). <https://htk.eng.cam.ac.uk/>.
52. [Babu et al. 2022] Babu, A., Wang, C., Tjandra, A., Lakhota, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., Auli, M.: "XLS-R: Self-Supervised Cross-Lingual Speech Representation Learning at Scale"; Proceedings of Interspeech 2022, pp. 2278–2282 (2022). <https://doi.org/10.21437/Interspeech.2022-143>.
53. [Oppenheim and Schaffer 2010] Oppenheim, A. V., Schaffer, R. W.: "Discrete-Time Signal Processing"; Pearson, Upper Saddle River, NJ, (2010). <https://www.pearson.com/en-us/subject-catalog/p/Oppenheim-Discrete-Time-Signal-Processing-3rd-Edition/P200000003226>.
54. [Nghia et al. 2008] Nghia, P. T., Binh, P. V., Thai, N. H., Ha, N. T., Kumsawat, P.: "A Robust Wavelet-Based Text-Independent Speaker Identification"; Proceedings of the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007), vol. 2, pp. 219–223 (2008). <https://doi.org/10.1109/ICCIMA.2007.149>.
55. [Spanias 1994] Spanias, A. S.: "Speech coding: A tutorial review"; Proceedings of the IEEE 82, 10 (1994), 1541–1582. <https://doi.org/10.1109/5.326413>.
56. [Carnero and Drygajlo 1999] Carnero, B., Drygajlo, A.: "Perceptual speech coding and enhancement using frame-synchronized fast wavelet packet transform algorithms"; IEEE Transactions on Signal Processing 47, 6 (1999), 1622–1635. <https://doi.org/10.1109/78.765133>.
57. [Phung et al. 2012] Phung, T.-N., Mai, C. L., Akagi, M.: "A Concatenative Speech Synthesis for Monosyllabic Languages with Limited Data"; Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2012) (2012). https://www.apsipa.org/proceedings_2012/.
58. [Taylor 2009] Taylor, P.: "Text-to-Speech Synthesis"; Cambridge University Press, Cambridge, (2009). https://books.google.com/books/about/Text_to_Speech_Synthesis.html?id=T0O-NHZx7kIC.
59. [Black and Tokuda 2005] Black, A. W., Tokuda, K.: "The Challenge of Synthesizing Speech from Limited Data"; Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 200–206 (2005). https://www.cs.cmu.edu/~awb/papers/ASRU2005/black_tokuda_asru2005.pdf.

60. [Phung et al. 2013] Phung, T.-N., Luong, C. M., Akagi, M.: "A Hybrid TTS between Unit Selection and HMM-Based TTS under Limited Data Conditions"; Proceedings of the 8th ISCA Workshop on Speech Synthesis (SSW 2013), pp. 279–284 (2013). https://www.isca-archive.org/ssw_2013/phung13_ssw.html.
61. [Phung et al. 2013] Phung, T.-N., Phan, T.-S., Vu, T. T., Luong, M. C., Akagi, M.: "Improving Naturalness of HMM-Based TTS Trained with Limited Data by Temporal Decomposition"; IEICE Transactions on Information and Systems E96-D, 11 (2013), 2417–2426. <https://doi.org/10.1587/transinf.E96.D.2417>.
62. [Nghia et al. 2016] Nghia, P. T., Tao, N. V., Huong, P. T. M., Diep, N. T. B., Hien, P. T. T.: "A Measure of Smoothness in Synthesized Speech"; REV Journal on Electronics and Communications 6, 1–2 (2016), 35–39. <https://www.rev-jec.org/index.php/rev-jec/article/download/106/109>.
63. [Nguyen and Phung 2017] Nguyen, G. N., Phung, T.-N.: "Reducing Over-Smoothness in HMM-Based Speech Synthesis Using Exemplar-Based Voice Conversion"; Journal of Audio, Speech, and Music Processing 2017 (2017), 14. <https://doi.org/10.1186/s13636-017-0113-5>.
64. [Xiong et al. 2018] Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., Stolcke, A.: "The Microsoft 2017 Conversational Speech Recognition System"; Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5934–5938 (2018). <https://doi.org/10.1109/ICASSP.2018.8461870>.
65. [Wang and Chen 2018] Wang, D., Chen, J.: "Supervised Speech Separation Based on Deep Learning: An Overview"; IEEE/ACM Transactions on Audio, Speech, and Language Processing 26, 10 (2018), 1702–1726. <https://doi.org/10.1109/TASLP.2018.2842159>.
66. [Thi et al. 2026] Thi, D. N., Tan, H. M., Vu, D.-Q., Phung, T.-N.: "Multi Learning on Discriminative Embedding Vector and Masking for Cocktail Party Effect"; IEEE Access (2026). <https://doi.org/10.1109/ACCESS.2026.3681807>.
67. [Tan et al. 2026] Tan, H. M., Tan, V. N., Tran, T. D., Thi, D. N., Vu, D.-Q., Phung, T.-N.: "Cocktail Party Effect with Fine-Tuning Multi-head Self-attention"; Advances in Information and Communication Technology, vol. 1832, pp. 526–534 (2026). https://doi.org/10.1007/978-3-032-18159-6_55.
68. [Thi et al. 2025] Thi, D. N., Tan, H. M., Phung, T.-N., Vu, D.-Q.: "Deep Learning for Speech Separation: A Comprehensive Review"; Advances in Information and Communication Technology, vol. 1205, pp. 222–229 (2025). https://doi.org/10.1007/978-3-031-80943-9_25.
69. [Kahn et al. 2020] Kahn, J., Rivière, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P. E., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., Likhomanenko, T., Synnaeve, G., Joulin, A., Mohamed, A., Dupoux, E.: "Libri-Light: A Benchmark for ASR with Limited or No Supervision"; Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020). <https://doi.org/10.1109/ICASSP40776.2020.9052942>.
70. [Radford et al. 2022] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: "Robust Speech Recognition via Large-Scale Weak Supervision"; OpenAI, (2022). <https://cdn.openai.com/papers/whisper.pdf>.
71. [OpenAI 2023] OpenAI: "GPT-4 Technical Report"; OpenAI, (2023). <https://cdn.openai.com/papers/gpt-4.pdf>.
72. [Ahlawat et al. 2025] Ahlawat, H., Aggarwal, N., Gupta, D.: "Automatic Speech Recognition: A Survey of Deep Learning Techniques and Approaches"; International Journal of Cognitive Computing in Engineering 6 (2025), 201–237. <https://doi.org/10.1016/j.ijcce.2024.12.007>.