Buletin Ilmiah Sarjana Teknik Elektro
Vol. 8, No. 2, April 2026, pp. 344-379
ISSN: 2685-9572, DOI: 10.12928/biste.v8i2.15564

Language as the Semantic Bridge in Audio, Music, and Multimodal Artificial Intelligence: A Systematic Review (2021-2025)

Novia Ratnasari, Aji Prasetya Wibawa
Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Indonesia.

ARTICLE INFORMATION

Article History:

Received 08 December 2025
Revised 04 March 2026
Accepted 30 March 2026

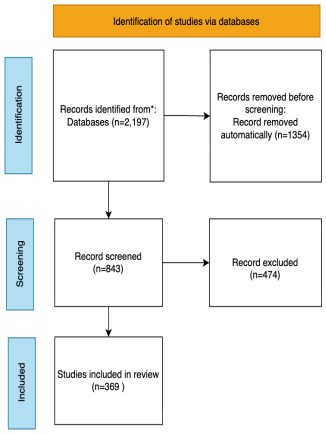
Keywords:

Systematic Literature Review;
PRISMA Framework;
Audio and Music Artificial Intelligence;
Natural Language Processing;
Multimodal Integration

Corresponding Author:

Aji Prasetya Wibawa,
Universitas Negeri Malang,
Jalan Semarang No. 5, Malang,
Jawa Timur 6514, Indonesia.
Email: aji.prasetya.ft@um.ac.id

ABSTRACT



This study presents a systematic review of research in Audio, Music, and Multimodal Artificial Intelligence published between 2021 and 2025, investigating how language operates as a semantic mediation layer between acoustic signals and high-level meaning. The research addresses the fragmentation of existing surveys by introducing a Domain; Modality; Technique; Task (D-M-T-T) taxonomy that systematically differentiates domain focus, modality configuration, modeling techniques, and task objectives. The research contribution is a structured analytical framework that offers a more granular perspective than architecture-centered surveys of Multimodal Large Language Models. Following the PRISMA 2020 protocol, 2,197 Scopus-indexed publications were screened, yielding 369 eligible studies. Language is defined as a representational layer encompassing natural language and structured symbolic encodings that connect acoustic embeddings to semantic interpretation and generative reasoning. Multimodal systems aligning audio and vision without explicit textual grounding are included and analyzed as non-linguistic alignment architectures within the taxonomy. The findings reveal a shift from recognition-based models toward unified multimodal systems in which language conditions alignment, reasoning, and generative synthesis. For instance, text-conditioned music generation demonstrates how linguistic prompts guide compositional structure and emotional expression. These developments reflect an epistemic transition from signal recognition paradigms to language-mediated generative intelligence. Emerging gaps include limited explainability in generative audio systems and insufficient low-resource cross-modal semantic grounding.

Document Citation:

N. Ratnasari and A. P. Wibawa, "Language as the Semantic Bridge in Audio, Music, and Multimodal Artificial Intelligence: A Systematic Review (2021-2025)," *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 8, no. 2, pp. 344-379, 2026, DOI: 10.12928/biste.v8i2.15564.

1. INTRODUCTION

Sound represents one of the most fundamental forms of human expression [1][2] and perception [3][4], encompassing linguistic [5], emotional [6][7], identity [9], and cultural dimensions [10][11]. The evolution of artificial intelligence [12][13], digital signal processing [14], and music knowledge has contributed to the emergence of interdisciplinary developments in digital technology, particularly in Artificial Intelligence for Audio and Music (AI Audio & Music) [15]. Within this field, computational modelling [16] enables systems to recognize [17], interpret, and generate new sounds [18] through signal processing techniques [19] and Natural Language Processing (NLP) [20][21]. These developments reflect the interaction between human cognition and machine intelligence, where sound is treated not only as a physical signal but also as a carrier of meaning. The field of AI in Audio & Music has undergone an epistemic transition from recognition-based paradigms to language-mediated generative intelligence [22][23].

The initial evolution was dominated by architectures such as Convolutional Neural Networks (CNN) [24]-[26] and Recurrent Neural Networks (RNN) [27][28] for speech [29] and music recognition tasks [30]. These methods quickly evolved into learning models that can extract semantic structure [31][32] from audio signals without manual annotation, including Transformers [33][34] and self-supervised models such as Wav2Vec2 [35][36], and MusicBERT [37]. Furthermore, generative architectures utilize Variational Autoencoders (VAE) [38][39]. Generative Adversarial Networks (GAN) [40][41], and diffusion models to enrich synthesis analysis [42][43] These models generate sound [44][45] and music patterns [46], while also enabling the detection of meaning and emotion [47][48]. NLP has shifted beyond its traditional linguistic [49] domain to function as an epistemic bridge that is, a structured representational layer connecting acoustic perception, musical expression [50][51], and higher-level semantic reasoning [52][53]. Through a learning and cross modal process, several studies have been used to construct a space within which text [54], sounds [55], and music interact to generate meaning [56]. This paradigm explains the rapid expansion of technology from signal based computing [57][58] to meaning infused cognition [59][60]. Language is positioned as a semantic bridge that enables systems to understand the context of sound output [61]. Within this broader technological evolution, the present study specifically examines how Natural Language Processing reconfigures these systems into semantically mediated and language-driven frameworks.

However, this epistemic transition toward semantically mediated and language-driven systems also introduces new structural and methodological complexities. During this epistemic transition, new challenges emerge at both epistemic and methodological levels, including dataset bias [62], limited cross-cultural generalization [63], and insufficient evaluation trials [64]This transition foregrounds the importance of meaning and emotional resonance [65] in intelligent systems [66]. Despite rapid advances in models [67] and architectures, research in Audio and Music AI remains conceptually fragmented. Many studies prioritize performance optimization without examining how language reshapes the epistemic foundations of sound-based intelligence [68]. As a result, a longitudinal and integrative synthesis of NLP's evolving role remains limited. This limitation constrains a broader theoretical understanding of how semantic mediation transforms intelligent audio systems and hinders the development of more coherent, context-aware, and human-centered AI applications

Against this backdrop, the present study aims to interpret the conceptual and methodological evolution of NLP [69] within Audio and Music AI between 2021 and 2025. To achieve this, the study employs a Systematic Literature Review (SLR) guided by the PRISMA [70] framework and structured under the Domain, Modality, Technique, Task (D-M-T-T) taxonomy. The analysis maps the architectural, methodological, and epistemic dynamics shaping the development of Audio & Music AI. It also clarifies the position of NLP as a semantic mechanism that unifies perception, expression, and understanding across modalities.

The overarching goal is to formulate a conceptual foundation for developing intelligent systems that not only recognize and generate sound but also interpret meaning, emotion, and aesthetic value mirroring how humans experience and understand the world through language and music. In this study there are 5 research questions, including:

RQ1 : How have NLP based approaches and methodologies evolved within Audio and Music AI research between 2021 and 2025, and in what ways does language function as a semantic bridge across modalities?

RQ2 : Which NLP modalities, techniques, and architectures are most dominant across studies in Audio and Music AI, and how have they shifted from signal processing toward semantic and affective representation learning?

RQ3 : What datasets and evaluation metrics are utilized in NLP based Audio and Music AI research, and how consistent and relevant are they across years and domains?

RQ4 : How has Audio and Music NLP been applied across various domains such as healthcare, creative industries, and human AI interaction and to what extent has it contributed to cross disciplinary innovation and multimodal generation?

RQ5 : What methodological challenges, technical limitations, and future research opportunities exist in advancing Audio and Music NLP toward more contextual, coherent, and semantically affective intelligent systems?

These research questions are designed to clarify how NLP has reshaped the conceptual, methodological, and architectural foundations of Audio and Music AI. By examining not only dominant models and datasets but also epistemic shifts and cross-domain applications, the study situates language as a central mechanism in the evolving intelligence of sound-based systems. By addressing these research questions, this study not only aims to consolidate existing knowledge but also to identify the conceptual and methodological gaps that continue to shape the research trajectory of Audio & Music AI. The Systematic Literature Review (SLR) approach provides a rigorous and structured framework to synthesize findings, highlight dominant trends, and trace the interaction between technological development and theoretical understanding. This review further emphasizes the need to balance acoustic and semantic dimensions, as several areas of musical meaning representation remain relatively underexplored yet hold substantial potential for future scientific inquiry. The research contribution of this study is a longitudinal and conceptually grounded synthesis of NLP's evolving role in Audio and Music AI, positioning language as a central epistemic mechanism in multimodal intelligence.

2. LITERATURE REVIEW

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta Analyses) framework was used as a methodological guideline to improve clarity, comprehensiveness, and transparency in meta analyses. The PRISMA stages used include: a). Identification; b). Screening; c). Eligibility Assessment; d). Inclusion/Exclusion; and e). Final Selection [71]. A Systematic Literature Review (SLR) was conducted on publications indexed in the Scopus database between 2021 and 2025. The PRISMA stages provide a systematic overview of the evolution, taxonomy, and research gaps in the field of NLP in Audio and Music (AI Audio & Music).

2.1. Eligibility, Planning, and Selection

PRISMA was used to ensure methodological accuracy, process transparency, and analytical coherence. At this stage, the research design, inclusion constraints, and article eligibility criteria were established as the basis for literature selection. The primary objective of this stage was to formulate a conceptual map of Audio & Music AI research, focusing on four analytical dimensions: a). Domain; b). Modality; c). Technique; and d). Task (DMT). These dimensions serve as structural pillars for analyzing how linguistic, acoustic, and computational modalities intersect in this field. During the planning stage, a continuous literature search process is carried out to maintain research consistency by applying several stages, including: a) Identification; b) Filtering; and c) Final Inclusion. This aims to select and identify relevant research that is high-quality, focused, valid, and representative.

2.1.1. Conditions for Inclusion (C)

Inclusion criteria (C1-C3) form a structured framework for assessing each article's eligibility. The sequential evaluation ensures that only studies meeting every specified condition are included in the final dataset.

C1 : Search Condition

Articles must be written in English and contain the terms "Natural Language Processing," "Text Classification," "Music," "Song," "Lyric," or "Voice" in the title, abstract, or keywords, retrieved from the Scopus database.

C2 : Screening Condition

Articles must be research journals or conference proceedings that include empirical experiments, dataset usage, model development, or evaluation within the context of AI applied to audio, music, or sound.

C3 : Eligibility Condition

Eligibility requirements must meet several requirements, including: a). Empirical results; b). Model implementation; c). Dataset utilization; and d). Evaluation metrics in the field of Artificial Intelligence (AI) Audio and Music with full access to files.

2.1.2. Restrictions for Inclusion (R)

The Restrictions for Inclusion (R) were applied after the initial search process to ensure the quality, consistency, and methodological rigor of the dataset in accordance with the PRISMA standards. These restrictions functioned as an additional filtering layer to refine the dataset and to exclude studies that, while initially relevant, did not fully meet the analytical or conceptual requirements of this review.

R1 : Articles that could not be accessed in full text form whether open or restricted access were excluded.

R2 : Studies not written in English were excluded.

R3 : Articles lacking an abstract or summary were not included.

R4 : Studies focusing solely on language teaching or pure linguistics were excluded.

R5 : After quality screening, conference proceedings, workshop papers, and Q3-Q4 journals were excluded. The final dataset included only Q1-Q2 journal articles with clearly defined empirical validation.

The initial stage of this Systematic Literature Review (SLR) began with the application of a search string developed according to Condition C1. This process was conducted using the Scopus database, which is widely recognized as one of the most credible scientific repositories in the fields of computer science, engineering, and artificial intelligence. Scopus was selected because it provides access to a broad range of high-quality, peer-reviewed publications, while also enabling categorization by journal quartile (Q-Q4) and research domain classification.

In the initial search phase, the two core components of this study, namely "Natural Language Processing" and "Audio & Music Artificial Intelligence", served as the conceptual foundation for constructing the search query. In the adequacy of comprehensive findings, the search was not limited to keywords, but was not limited to the same terms with the same meaning, presented as follows:

"(TITLE-ABS-KEY ("Natural LANGUAGE processing") OR TITLE-ABS-KEY ("Text Classification") AND TITLE-ABS-KEY ("music") OR TITLE-ABS-KEY ("song") OR TITLE-ABS-KEY ("lyric") OR TITLE-ABS-KEY ("voice")) AND PUBYEAR > 2020 AND PUBYEAR < 2026 AND (LIMIT-TO (SRCTYPE, "j") OR LIMIT-TO (SRCTYPE, "p")) AND (LIMIT-TO (DOCTYPE, "ar") OR LIMIT-TO (DOCTYPE, "cp")) AND (LIMIT-TO (LANGUAGE, "English"))"

2.2. Study Selection Process

The study selection phase aimed to ensure that each analyzed publication was both relevant and scientifically robust. The process began with an initial screening of publications retrieved from Scopus, where articles were filtered through an examination of titles, keywords, and abstracts to assess their alignment with the scope of Audio and Music AI. Articles with uncertain relevance were not immediately excluded but were reevaluated through full text reading to examine their methodology, data integrity, and conceptual contribution. This approach ensured that potentially valuable studies were not overlooked during the initial filtering phase.

2.3. Data Collection

The data collection process in this Systematic Literature Review (SLR) was carried out in a planned and sequential manner to obtain literature most relevant to the research topic. The primary objective was to address the formulated research questions while establishing a strong conceptual foundation grounded in empirical evidence. The inclusion and exclusion process implemented research design, publication type, language use, and topic relevance in the Scopus database, with an initial search totaling 2,197 publications. The next stage was the filtering process based on title, abstract, and content, which yielded 843 articles. The final results of this filtering stage, which yielded quality, and further eligibility screening, yielded 369 journal articles with quartiles Q1-Q2.

2.4. Data Extraction and Quality Assessment

During data extraction, each article meeting the inclusion criteria was examined in depth to document key information and ensure comparability across studies. The extracted fields covered study descriptors, domain classification, methodological details, task typology, datasets, evaluation metrics, and salient findings or limitations. Subsequently, a structured quality assessment was performed to validate the reliability and interpretive robustness of the dataset. During data extraction, each article meeting the inclusion criteria was analyzed in depth to document the following information:

a) Study characteristics;
b) Research domain focused on audio, music, multimodal, and NLP;
c) Techniques and algorithms used;
d) Task types such as classification;

e) Datasets used, including names and sizes;
f) Evaluation metrics;
g) Key findings and limitations reported by each study;

A quality assessment was conducted to ensure the validity and reliability of the analytical results. This evaluation considered four primary indicators:

a) Domain relevance, the degree to which each study aligns with the scope of Audio and Music AI;
b) Methodological clarity, the transparency of model descriptions, evaluation techniques, and experimental outcomes;
c) Empirical evidence, the extent to which the study employs datasets and quantitative validation; and
d) Document accessibility, the availability of a complete, verifiable manuscript.

This stage ensured that all included publications contributed verifiable scientific evidence and supported the overall objectives of the review. Throughout the process, data management and documentation practices were maintained rigorously using tools such as Microsoft Excel and Mendeley to record metadata, track screening status, and prevent duplication. The implementation of a systematic data governance protocol not only enhanced the transparency and reproducibility of the research process but also reinforced the scientific integrity of this review. Figure 1 presents a flowchart of the PRISMA 2020 [72] article screening and selection process. The initial identification phase yielded 2,197 publications, and the final selection phase yielded 369 articles from Q1-Q2 journals.

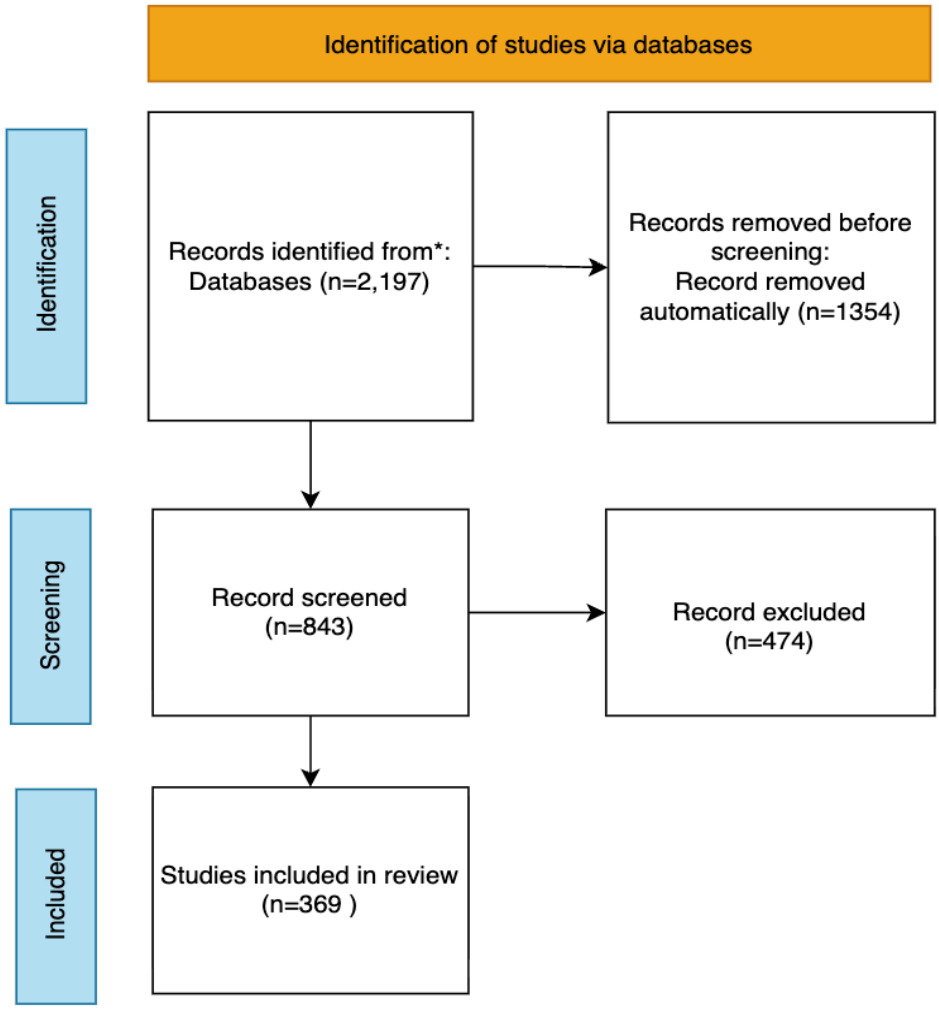


Figure 1. PRISMA Flow Diagram of Study Identification, Screening, and Inclusion (2021-2025)

Figure 1. PRISMA Flow Diagram of Audio and Music AI Study Selection Process (Scopus 2021-2025). This Figure1 illustrates the systematic identification and selection of studies based on the PRISMA 2020

framework. References to the PRISMA flow diagram have been explicitly integrated within the relevant methodological stages to improve narrative continuity. The figure caption has been revised to clearly describe each selection phase (identification, screening, eligibility, and inclusion), thereby enhancing interpretability and alignment with PRISMA reporting standards. A total of 2,197 records were initially identified from the Scopus database using the keywords ("Natural Language Processing" OR "Text Classification") AND (Music OR Song OR Lyric OR Voice). After removing 1,354 duplicates and irrelevant entries, 843 records proceeded to the screening phase, where titles, abstracts, and keywords were examined for domain relevance to Audio and Music AI. Subsequently, 474 records were excluded for not meeting thematic criteria. The remaining 369 high-quality Q1-Q2 journal articles met all inclusion and eligibility standards and were incorporated into the final systematic review dataset. In this study, "Q1-Q2" refers to journals ranked in Quartiles 1 and 2 according to the SCImago Journal Rank (SJR) classification, as accessed and verified in 2025.

3. METHODS

To enhance methodological clarity and provide a visual overview of the study selection procedure, a structured flowchart is presented below. This diagram summarizes the sequential stages of the Systematic Literature Review (SLR), beginning with identification and proceeding through screening, eligibility assessment, and final inclusion. The flowchart complements the PRISMA-based documentation by illustrating the logical progression of filtering and validation steps applied to construct the final dataset. Figure 2 Methodological Flow of the Systematic Literature Review (SLR) Based on PRISMA 2020. The diagram illustrates the sequential stages of identification, screening, eligibility assessment, and inclusion applied in constructing the final dataset.

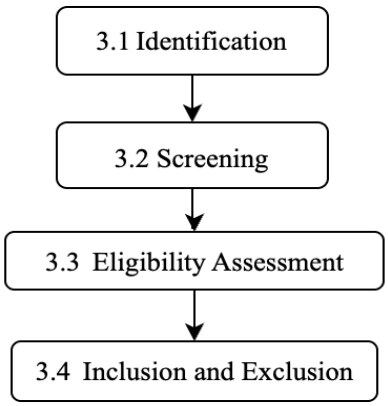


Figure 2. Methodological Flow of the Systematic Literature Review (SLR) Based on PRISMA 2020

3.1. Identification

The identification stage was conducted using the Scopus database due to its structured indexing system and comprehensive coverage of peer-reviewed publications in Computer Science, Engineering, and Artificial Intelligence. The search strategy was designed to retrieve studies related to Audio and Music Artificial Intelligence (AI), including intersections with Natural Language Processing and multimodal systems. The search covered publications between 2021 and 2025 and yielded 2,197 records. All retrieved records were exported into a structured database to facilitate systematic screening and transparent documentation of selection decisions. The complete search string and retrieval parameters are reported to ensure transparency and reproducibility.

3.2. Screening

During the screening phase, titles and abstracts of the 2,197 records were examined to assess alignment with the conceptual and technical scope of Audio and Music AI. Predetermined inclusion criteria (C1-C3) and exclusion criteria (R1-R5) were applied consistently. Studies were retained if they addressed audio, music, or multimodal systems involving sound, employed AI-based computational methodologies, and reported empirical or quantitative findings. Terminologies such as music generation [73], audio classification [74], lyric-based modeling, transformer architectures, and multimodal learning were considered indicators of thematic relevance. Articles were excluded if they were not written in English, lacked full-text availability, did not employ AI-driven approaches, or were purely conceptual without empirical validation. Following this stage,

843 records demonstrated sufficient thematic and methodological alignment to proceed to full-text eligibility assessment.

3.3. Eligibility Assessment

The eligibility stage involved a comprehensive full-text assessment of the 843 studies retained after title and abstract screening. Each article was examined systematically to ensure methodological rigor, empirical validity, and substantive alignment with the scope of Audio and Music Artificial Intelligence (AI). The evaluation focused on research design transparency, dataset specification, clarity of computational methodology, and the presence of reproducible experimental procedures. Particular attention was given to whether the study explicitly described its model architecture, data sources, evaluation metrics, and reported quantitative results. Studies were excluded if they lacked empirical validation, failed to provide sufficient methodological detail, did not clearly report evaluation protocols, or were not directly situated within audio, music, or multimodal sound-based systems involving AI-based approaches. Conceptual discussions without experimental implementation were also excluded at this stage. Through this structured full-text evaluation, the dataset was refined to 369 journal articles that met all predefined inclusion criteria and satisfied the methodological standards established for this review. These retained studies form the validated dataset for subsequent data extraction, comparative analysis, and longitudinal synthesis. Data management and screening documentation were conducted using Microsoft Excel to ensure traceability of inclusion and exclusion decisions.

3.4. Inclusion and Exclusion

The final dataset consists of 369 journal articles indexed in Scopus and published between 2021 and 2025. The designation Q1-Q2 refers to journals ranked in Quartiles 1 and 2 according to the SCImago Journal Rank (SJR), as accessed in 2025. This quartile classification was applied as a predefined quality criterion during study selection. All inclusion and exclusion decisions were systematically documented in the PRISMA flow diagram, which records the progression of records from identification through screening and eligibility to final inclusion. This documentation ensures transparency of the selection process and traceability of decision-making at each stage. The resulting dataset constitutes the validated empirical foundation for the subsequent data extraction, structured analysis, and longitudinal synthesis conducted under the Domain, Modality, Technique, and Task (D-M-T-T) framework.

4. RESULT AND DISCUSSION

4.1. Quantitative Results: Descriptive Overview of the Dataset

This section presents the quantitative structure of the final corpus using the Domain-Modality-Technique-Task (D-M-T-T) framework as an analytical lens. The distribution of the 369 included studies is examined across five dimensions: publication trends over time (2021–2025), domain composition, modality representation, technique adoption, and task classification. Together, these descriptive statistics establish the structural configuration of the dataset prior to the qualitative synthesis of evolutionary and conceptual developments.

4.1.1. Publication Distribution (2021-2025)

This section presents the quantitative distribution of studies included in the final dataset. A total of 369 peer-reviewed publications published between 2021 and 2025 were analyzed to establish the structural composition of the corpus. The annual and quartile-based distribution of these studies is summarized in Figure 3. Figure 3 depicts the annual distribution of Q1- and Q2-indexed publications included in the final dataset for the period 2021-2025. The corpus comprises 369 peer-reviewed studies, with annual totals of 42 (2021), 52 (2022), 76 (2023), 91 (2024), and 108 (2025), indicating a progressive increase across the review window. Of the total corpus, 230 studies (62.3%) were published in Q1 journals and 139 (37.7%) in Q2 journals. The combined output of 2024 and 2025 accounts for 199 publications (54.0% of the dataset), representing the highest concentration within the five-year span. These quantitative distributions establish the structural composition of the corpus and frame the subsequent domain-, task-, and architecture-level analyses.

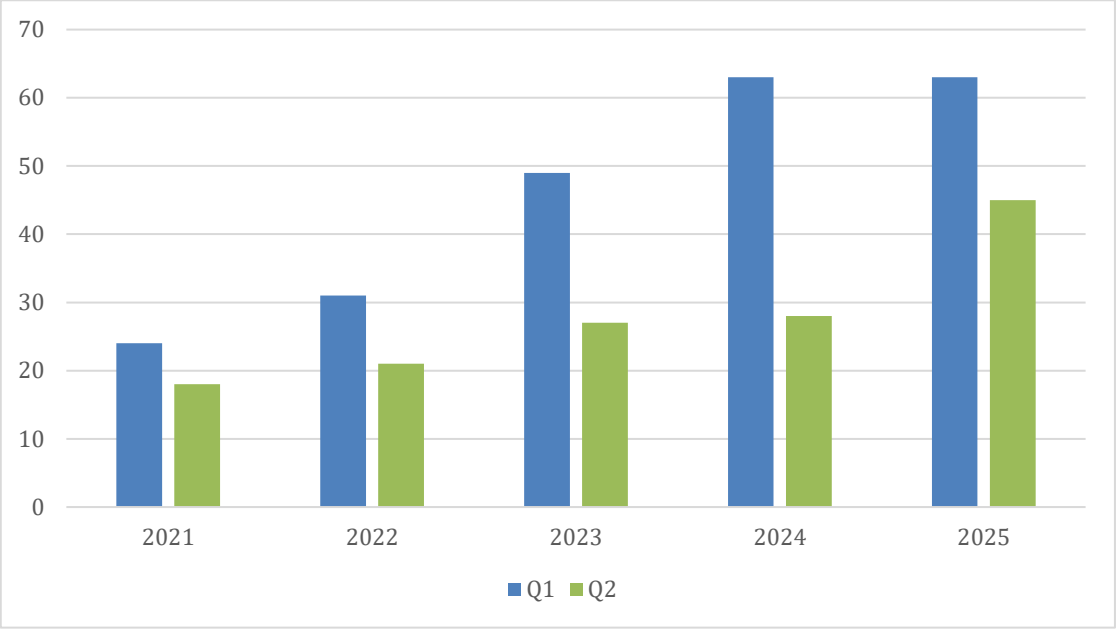


Figure
Figure 3. Annual Distribution of Q1 and Q2 Publications (2021-2025)

4.1.2. Domain Distribution

This subsection reports the distribution of the 369 included studies across the primary research domains. Each publication was classified according to its dominant disciplinary focus under the Domain-Modality-Technique-Task (D-M-T-T) framework. The resulting domain-level composition provides a structural overview of the research landscape prior to modality-, technique-, and task-level analyses. Figure 4 presents the annual distribution of the 369 included studies across the three primary research domains: Multimodal AI, Audio NLP, and Music NLP. Multimodal AI constitutes the largest share of the corpus with 204 studies (55.3%), followed by Audio NLP with 120 studies (32.5%) and Music NLP with 45 studies (12.2%). At the annual level, Multimodal AI increased from 27 studies in 2021 to 62 studies in 2025, showing a continuous upward trend. Audio NLP fluctuated between 17 and 30 studies per year, reaching its highest count in 2024 (30 studies). Music NLP represents the smallest proportion across all years, with counts ranging from 4 to 14 studies. These Figure 4 describe the structural distribution of domain-level research within the reviewed period.

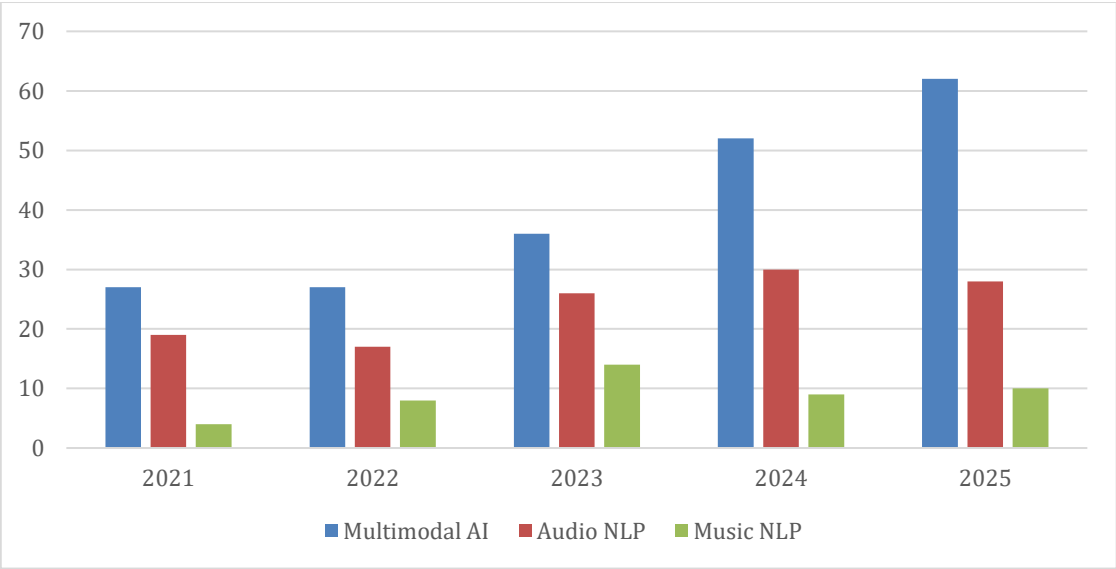


Figure 4. Annual Distribution of Studies by Research Domain (2021-2025)

4.1.3. Modality Distribution

This subsection presents the distribution of modality configurations across the 369 reviewed studies. Each publication was categorized based on the primary data modality or combination of modalities employed in the proposed system, including single-modality and cross-modal configurations. This classification provides a structural overview of how different forms of data representation are utilized within the reviewed research landscape prior to further analytical interpretation. Figure 5 presents the annual distribution of modality configurations across the 369 reviewed studies published between 2021 and 2025. Overall, single-modality research remains substantial throughout the period [75][76], particularly in Audio-only (n = 96) and Text-only (n = 82) studies. However, a clear structural shift emerges over time. In the early phase (2021-2022), the literature is primarily characterized by unimodal designs [77], with Audio-only (15-18 studies per year) and Text-only (12-14 studies per year) dominating the landscape, while cross-modal integration remains comparatively limited (Text+Audio = 8-10 studies). Beginning in 2023, a noticeable transition occurs, marked by a sharp increase in bimodal configurations, especially Text+Audio (18 studies in 2023; 25 in 2024; 33 in 2025). This trajectory suggests an intensifying convergence between linguistic and acoustic representations. By 2025, the expansion of ≥3 modality studies (n = 19 in 2025; total = 30) further indicates the growing prominence of fully multimodal architectures. Collectively, these patterns reflect a longitudinal evolution from modality-isolated modeling toward integrative and multimodal systems, reinforcing the broader epistemic movement from signal-centered processing to semantically coordinated, cross-modal intelligence.

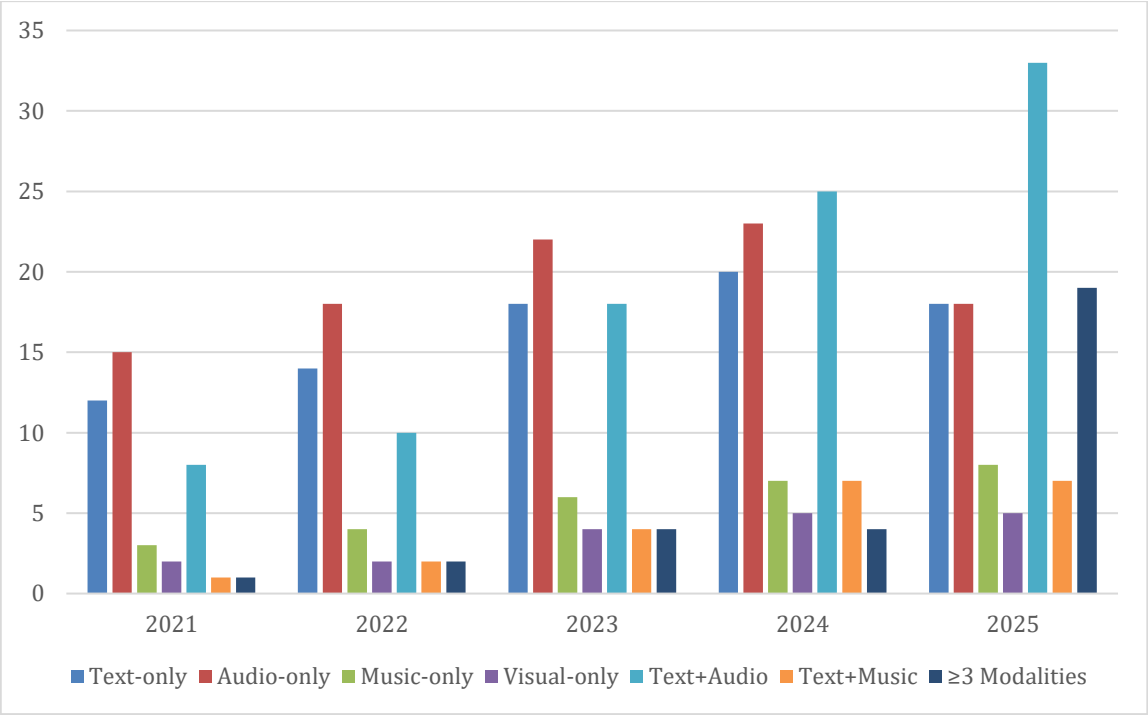


Figure 5. Annual Distribution of Modality Configurations (2021-2025)

4.1.4. Technique Distribution

This subsection examines the distribution of computational techniques employed across the 369 selected studies to identify methodological patterns in Audio, Music, and Multimodal AI between 2021 and 2025. Figure 6 presents the longitudinal distribution of computational techniques employed in Audio, Music, and Multimodal AI studies between 2021 and 2025. Across the 369 selected studies, Deep Learning architectures (CNN, LSTM, Transformer) [78][79] constitute the largest category with 100 studies, increasing from 15 in 2021 to a peak of 28 in 2023 before declining to 11 in 2025. This confirms their position as the dominant computational backbone of the field. Traditional Machine Learning methods [80] (SVM, Random Forest, Logistic Regression) [81][82] follow closely with 94 studies, showing relatively stable usage across all five years. This indicates methodological continuity, where classical models coexist alongside deep neural architectures rather than being fully replaced [83][84]. Non-deep NLP techniques (TF-IDF, N-gram, LDA) account for 67 studies [85][86], peaking in 2023 (17), while Speech/Audio Processing methods [87][88] (ASR,

MFCC-based pipelines) [89][90] total 56 studies, remaining relatively stable between 11-13 studies during the mid-period. These approaches continue to function as foundational or complementary components within broader modeling frameworks. Multimodal approaches comprise 37 studies, gradually increasing toward 2024 (11 studies) [91][92], reflecting growing cross-modal integration [93]. Large Language Models [94][95] (e.g., GPT, LLaMA) represent the smallest but emerging category with 15 studies, expanding from 0 in 2021 to 6 in 2024, indicating a recent yet notable methodological shift. Overall, the distribution demonstrates a structural transition from feature-centric and task-specific pipelines toward deep, multimodal, and increasingly language-centered computational architectures.

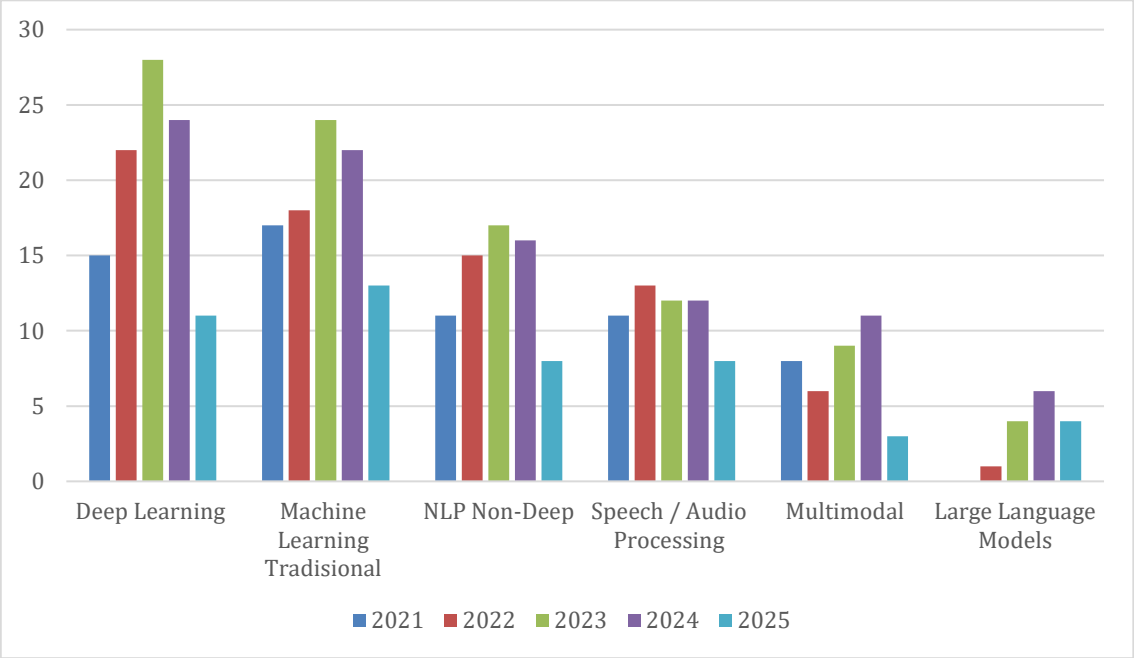


Figure 6. Longitudinal Distribution of Computational Techniques in Audio, Music, and Multimodal AI (2021-2025)

4.1.5. Task Distribution

This subsection examines the distribution of research tasks across the 369 selected studies, aiming to identify dominant problem orientations and their evolution within Audio, Music, and Multimodal AI between 2021 and 2025. Figure 7 illustrates the annual distribution of dominant research tasks across the 369 reviewed studies. An analysis of 369 research tasks conducted between 2021 and 2025 indicates that classification remains the dominant problem orientation, accounting for 120 studies (32.5%). This category includes emotion classification [96], sentiment analysis [97][98], genre classification [99][100], speaker identification [101][102], and disease classification [103][104], and consistently represents the primary research focus across all years. The second-largest category is detection, comprising 57 studies (15.4%). This includes deepfake detection [105][106], fraud/phishing detection [107][108], bias detection [109][110], and clinical disorder screening using speech and text [111]. A noticeable increase occurred in 2023-2024, reflecting growing attention to system security and reliability [112]. Transcription tasks [113] (e.g., ASR, speech-to-text, alignment) [114][115], account for 49 studies (13.3%) and remain stable across years, serving as foundational infrastructure for downstream systems [116]. Similarly, generation tasks [117] total 48 studies (13.0%), covering text generation [118][119], music generation [120][121], text-to-speech synthesis [122], and dialogue generation [123][124] with growth particularly evident in 2024-2025. Other categories appear in smaller but meaningful proportions: retrieval / information extraction [125][126] (29 studies; 7.9%), separation / enhancement [127] (23 studies; 6.2%), and recommendation systems [128] (20 studies; 5.4%). Additionally, instruction-following / control systems account for 18 studies (4.9%) [129], and multimodal reasoning represents 16 studies (4.3%) [130], both showing gradual growth in 2023-2024. Studies focused on evaluation / assessment remain limited, with 11 studies (3.0%). Overall, the 2021-2025 research landscape is characterized by the dominance of classification tasks (32.5%), followed by detection (15.4%), transcription (13.3%), and generation (13.0%), with a gradual shift toward generative and multimodal systems in recent years.

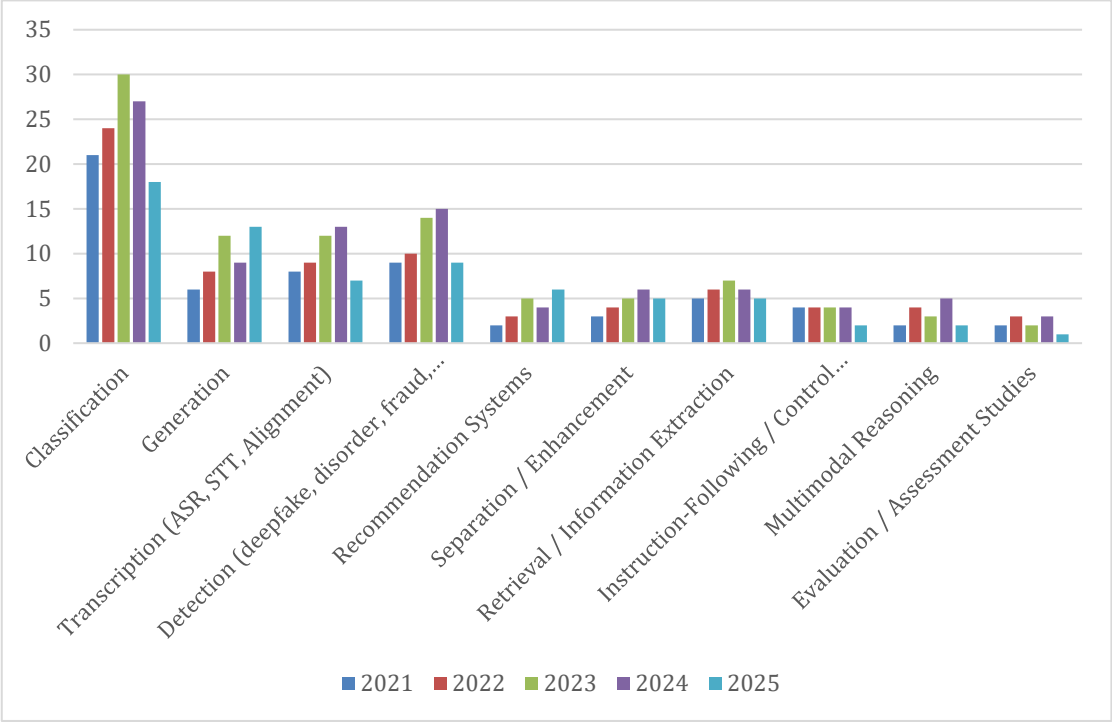


Figure 7. Longitudinal Task Distribution Across Audio, Music, and Multimodal AI Domains (2021-2025)

4.1.6. Comparative Positioning with Prior Reviews

This subsection compares the present review with previous systematic literature reviews published between 2021 and 2025. The aim is to clarify how this study differs in scope, structure, and analytical focus. While prior reviews often concentrate on specific tasks or domains within audio or music research, the present study provides a broader longitudinal synthesis across Audio NLP, Music NLP, and Multimodal AI using the D-M-T-T framework. Table 1 presents a comparative positioning of prior systematic literature reviews published between 2021 and 2025 in relation to the present study, with particular attention to differences in analytical scope and reported findings. Earlier reviews tend to concentrate on specific tasks or subdomains. For example, Lozano [131] identified context-aware recommender systems as predominantly driven by feature engineering and metadata-based personalization strategies. In contrast, our longitudinal analysis (2021-2025) reveals a progressive shift toward multimodal embedding alignment and transformer-based architectures that integrate audio [135], lyrics [136], and contextual signals within unified semantic spaces. Similarly, Barnett [132] examined generative audio models primarily from an ethical and governance perspective, highlighting concerns related to bias, misuse, and accountability. While these normative insights are essential, the study does not provide structural analysis of architectural evolution. Our findings extend this perspective by demonstrating a measurable increase in language-conditioned generative models and multimodal transformer frameworks between 2023 and 2025, indicating a broader epistemic transition beyond ethical discourse alone. Mohammad [133] reported continued dominance of recognition-oriented sound event detection pipelines. However, our cross-domain synthesis indicates that recognition tasks increasingly coexist with generative and multimodal reasoning tasks, suggesting diversification rather than persistence of a single paradigm. More recently, Budi Putra [134] reviewed music information retrieval techniques for genre classification and confirmed the sustained importance of classification-based pipelines in MIR [137]. In contrast, our analysis shows that classification, while still prevalent, is gradually complemented by generation, retrieval, and instruction-following tasks, reflecting functional expansion across the Audio-Music-Multimodal ecosystem. Overall, unlike prior reviews that remain task-specific, domain-specific, or thematically bounded, the present study adopts a broader and longitudinal perspective spanning 2021-2025. By integrating Audio NLP, Music NLP, and Multimodal AI under the unified Domain-Modality-Technique-Task (D-M-T-T) framework, this review provides a structured mapping of architectural evolution, task diversification, and semantic convergence trends. Methodologically, the analysis is grounded in Scopus-indexed, English-language Q1-Q2 publications, ensuring rigor while maintaining transparency in scope definition.

Table 1. Comparative Positioning with Prior Systematic Literature Reviews (2021-2025)

Author	Scope	Time	Coverage	Methodological Boundary
Lozano [131]	Music recommender	Pre-2021	Task-specific	Limited to recommendation systems
Barnett [132]	Generative audio ethics	Pre-2023	Ethics-focused	No structural AI taxonomy
Mohammad [133]	Sound event detection	Pre-2024	Detection-focused	Recognition pipelines only
Budi Putra [134]	MIR genre classification	2025	Domain-specific	No multimodal integration
Present Study	Audio-Music-Multimodal AI	2021-2025	Cross-domain & longitudinal	Scopus-indexed; English-only; Q1-Q2 filtering

4.2. Qualitative Discussion

4.2.1. Main Findings

4.2.1.1. Paradigm Shift in Research (2021-2025)

Between 2021 and 2025, research in Audio and Music Artificial Intelligence underwent a significant paradigm shift from signal-centered recognition systems toward language-mediated semantic modeling [81] and generative intelligence [138]. In the early phase, most studies focused on acoustic feature extraction [139] and classification tasks [102], emphasizing phonetic accuracy [140], emotion detection [141], and genre recognition [142] through conventional convolutional [143] and recurrent architectures [144]. Sound was primarily treated as a physical signal requiring precise representation and decoding [116]. Over time, however, advances in self-supervised learning [145] and Transformer-based architectures enabled models to move beyond surface-level recognition toward contextual understanding [146] and meaning construction [147]. Language progressively emerged as a central representational layer, mediating the relationship between acoustic perception and semantic interpretation [148]. Rather than merely transcribing or classifying audio signals [149], systems increasingly aimed to interpret intent, affect, and contextual meaning embedded within sound. By the later stages of the reviewed period, generative and multimodal frameworks became prominent, transforming audio and music AI into ecosystems capable not only of recognition but also of semantic synthesis and co-creative generation. This evolution reflects an epistemic transition in which sound is no longer treated solely as a signal to be decoded, but as a linguistic and affective construct situated within a broader multimodal semantic space.

4.2.1.2. Evolution of Domains and Modalities

The development of Artificial Intelligence (AI) research in the fields of audio [150] and music shows that technical progress is determined not only by advances in learning models but also by how domains and modalities are defined and integrated. The domain represents a disciplinary focus, such as Audio AI [151][152], which emphasizes sound signal processing [153][154]; Music AI [155][156], which highlights musical structure and affective expression [157][158]; and Multimodal AI [159][160], which combines multiple forms of data to produce unified meaning. Meanwhile, modality refers to the form of data representation text, audio, image, or other signals that serves as the foundation for perception, representation, and generation in AI systems. Within the context of research from 2021 to 2025, the relationship between domain and modality has become increasingly interwoven, reflecting a paradigm shift from single modality processing [161][162] to language mediated multimodal systems [163][164]. Accordingly, this section discusses the evolution of domain concepts and the role of modality as a semantic connector [165], enabling AI systems to understand, interpret, and generate meaning across different forms of representation.

Table 2 presents the evolution of domains and modalities in Audio, Music, and Multimodal AI research between 2021 and 2025. Overall, the developmental trend indicates a transition from acoustic signal processing to language-mediated semantic understanding [166]. In the early phase, Audio AI and Music AI focused on signal recognition and emotion analysis [48][167], while Multimodal AI primarily linked text and audio descriptively [168]. Over time, all three evolved toward contextual and generative systems [169], with language functioning as the semantic bridge across modalities. By 2025, the integration of six modalities text, audio, music, visual, emotion, and memory positions language as the central mechanism for cross-domain meaning construction.

Table 2 presents a qualitative overview of the evolution of domains and modalities in Audio, Music, and Multimodal AI research from 2021 to 2025. Between 2021 and 2025, research in Audio and Music AI moved from acoustic signal recognition to systems capable of interpreting meaning, context, and emotion in sound [187]. Language became the semantic connective tissue [188], linking text, sound, music, and vision within shared representational spaces [189][190]. This development unfolded across Audio AI, Music AI, and Multimodal AI, which gradually converged toward cross-modal semantic representation.

In AudioNLP, early work centered on acoustic features such as MFCC [191], spectrograms [192], and mel filterbanks, using CNN [193][194] and RNN architectures [195][196] for ASR [197][198] and SER [199][200], focusing mainly on signal recognition [201]. The introduction of self-supervised models such as Wav2Vec2 [202] and HuBERT enabled systems to learn linguistic representations from raw speech, connecting acoustic signals to speaker intent [203][204]. This shift supported tasks such as intent detection [205] and emotion-aware modeling. Later, deeper NLP integration appeared through BLSTM-CRF [206][207], Transformer-based semantic parsing [208], and SpeechSQLNet [161], enabling speech-to-SQL translation. Applications expanded to education [209]-[211] and healthcare [212][213] via systems such as SAIL [214] and DAX™ [215][216]. Recent models, including multilingual ASR [217][218] and SSL-based EmoSDS [219], support adaptive and emotionally responsive dialogue. Overall, audio is no longer treated only as a signal, but as a linguistic expression carrying meaning.

In Music AI, research progressed from signal-based analysis [220] to linguistic and semantic modeling [221]. Early studies linked lyrics [222], emotion [223][224], and musical features such as tempo and tonality [225], using Word2Vec and FastText embeddings [226]-[228]. Later, models such as BiLSTM [229][230] and BERT [231][232] enabled lyric music alignment and text-conditioned emotion generation. The field then entered a generative phase with architectures such as MuseBERT, MusicLM, and MusicCaps-CLAP [233], which translated text into coherent musical output. Datasets such as MusicCaps and GTZAN-Fusion [46][234] strengthened semantic alignment, while cross-lingual studies expanded cultural understanding [235]. Recent contrastive frameworks unified text [236], mood [237], and melody [238], and models such as Jukebox-XL and Udio T5 generated music aligned with linguistic prompts [239], supporting applications in therapy and education [240][241].

Figure 8 illustrates the ontological and epistemic relationship between the two principal research domains Audio AI and Music AI. Based on the synthesis of studies from 2021 to 2025, Audio AI primarily concerns the processing and perception of acoustic signals, whereas Music AI is oriented toward the construction of semantic meaning and affective expression derived from sound. Consequently, from an epistemic standpoint, Music AI is positioned as a subset of Audio AI, representing a higher level of understanding in which sound is not merely recognized but also interpreted linguistically and emotionally.

Figure 9 underscores the position of Natural Language Processing (NLP) as the semantic axis connecting perception in Audio AI [242][243] and expression in Music AI [244][245], ultimately converging toward integrated understanding within Multimodal AI [246][247]. NLP functions not merely as a linguistic tool but as an epistemic mechanism that translates sound into meaning and meaning into emotional expression [248][249]. The evolution of Audio, Music, and Multimodal AI research between 2021 and 2025 reveals that the primary role of NLP has shifted from a purely linguistic component to a semantic core that interlinks perception, expression, and cross modal understanding [250][251]. Audio AI represents the perceptual dimension, focusing on the recognition [252][253] and interpretation of acoustic signals, whereas Music AI operates within the expressive dimension [254][255], constructing emotional [256][257] and semantic structures from sound [258][259]. Situated between these layers, NLP serves as an epistemic bridge, connecting language, meaning, and affect, thereby allowing AI systems to convert acoustic perception into meaningful musical expression. Collectively, these three domains illustrate an epistemic shift from signal level processing to cross modal meaning representation. Despite significant progress, cross domain research continues to face challenges in maintaining semantic representation stability, achieving affective alignment across modalities, and developing cross cultural benchmarks capable of evaluating semantic coherence comprehensively.

Table 2. Qualitative Overview of Domains and Modalities (2021-2025)

Year	Audio AI	Music AI	Multimodal AI
2021	Signal & prosody recognition [170][171]; CNN/RNN architectures [172][173]; ASR, SER	Lyric emotion analysis; Word2Vec/FastText embeddings s [174]	Dual-modality (Text Audio); AudioCLIP, AudioCaps
2022	Linguistic representation learning [175]; SSL (Wav2Vec2, HuBERT) [176][92]; semantic abstraction [177]	Lyric emotion mapping [178]; mood tagging; text acoustic correspondence	Tri-modality (Text Audio Music) [179]; genre [180] & mood recognition
2023	Transformer-based semantic alignment *t* [171]; prosody meaning integration	Text-conditioned generation; BERT, BiLSTM [181][182].	Quad-modality (+Emotion); CLAP, MusicLM
2024	NLP acoustic convergence [94]; hybrid systems (SpeechSQLNet, MAM-BERT) *s* [183]	Semantic affective generative phase; MusicLM, MuseBERT, [184] CLAP	Penta-modality (Visual, Social); VOICE, CAMFF
2025	Cognitive & adaptive dialogue; EmoSDS, Vox Calculi	Dynamic linguistic control; Udio-T5, Jukebox-XL	Hexa-modality (+Memory); VISION [185], V2M [186]

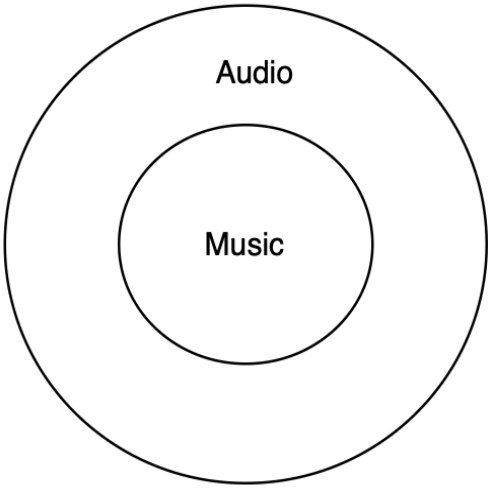


Figure 8. Ontological and Epistemic Relation between Audio AI and Music AI

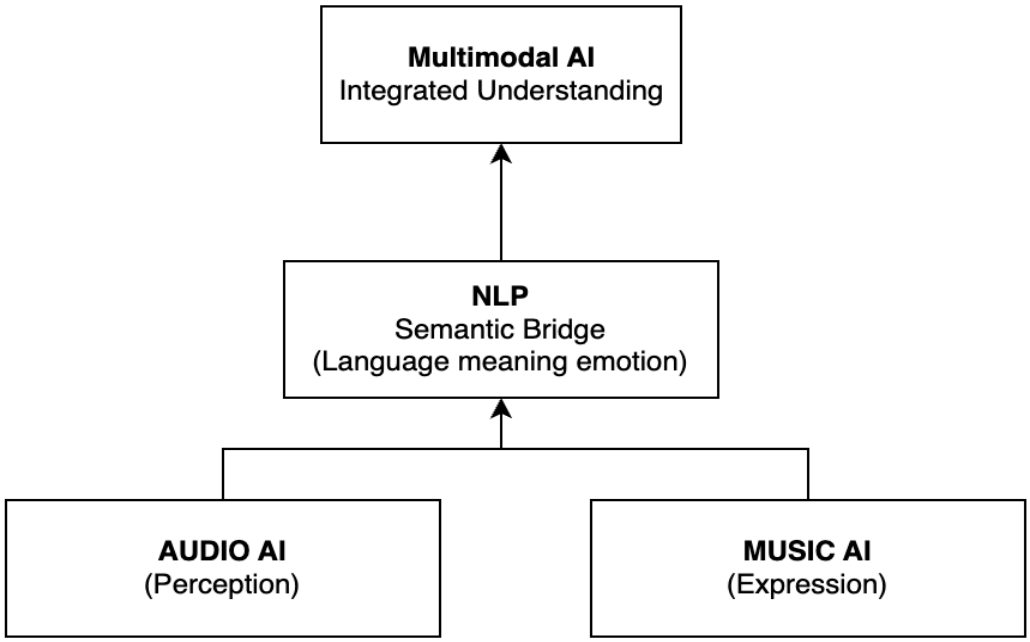


Figure 9. The Role of NLP as a Semantic Bridge among Audio AI, Music AI, and Multimodal AI

4.2.1.3. Evolution of Task Classification and Applications in Audio & Music AI

The classification of tasks and applications in Audio and Music AI highlights the rapid evolution that occurred between 2021 and 2025. Through diverse tasks such as speech recognition [260][261], lyric analysis, and cross modal integration, the interaction between audio, music, and language processing has become the fundamental basis for developing intelligent systems capable of comprehending, interpreting, and creating contextual meaning. Table 3 illustrates the evolutionary trajectory across the three primary domains Audio AI, Music AI, and Cross-Modal AI from 2021 to 2025. The transition moves from signal driven perceptual systems to cross modal meaning generation. In the early years, conventional models such as CNN and LSTM dominated recognition and classification tasks in both speech processing [262][263] and lyric analysis. With advancing technology, self-supervised models like Wav2Vec2 and HuBERT enabled the emergence of linguistic representations without manual annotation. Throughout this evolution, language has served as the semantic bridge uniting acoustic perception, emotional expression, and symbolic representation signifying a paradigm shift from signal recognition toward semantic affective cognition [264], which now forms the epistemic foundation of modern Audio and Music AI.

Table 3. Qualitative Analysis of the Evolution of Domains, Tasks, Models, Applications, and Research Directions (2021-2025)

Year	Domain & Main Focus	Representative Tasks	Dominant Models / Architectures	Applications & Implementation Context	Research Direction & Emerging Trends
2021	Audio AI Acoustic Signal Level Recognition	ASR [265], KWS [266], SER [267]	CNN [268], RNN [269], CRNN [270]	Speech recognition [267], voice command [271], emotion tagging [272]	Transition from signal recognition to early linguistic mapping
	Music AI Early Linguistic Analysis of Lyrics	Lyric classification [273], lyric emotion recognition	LSTM [274], BERT Lyric [275], CNN Attention [276]	Affective lyric analysis and music emotion classification	Strengthening correlations between text and musical expression
2022	Cross Modal AI Early Text Audio Integration	Text to-audio retrieval, audio captioning [277]	AudioCLIP, WavCaps (dual encoder)	Text based sound description, sound retrieval	Early semantic mapping between text and sound
	Audio AI Self Supervised Representation Learning	Multilingual ASR, Contextual KWS [278]	Wav2Vec2 [279], HuBERT [280]	Multilingual ASR [281], intent recognition [282]	Label free learning and cross lingual fine tuning
	Music AI Strengthening Text Music Semantic Relations	Lyric mood correlation [283], emotion tagging [284]	BiLSTM, FastText, Word2Vec	Mood analysis and semantic correlation in music [162]	Expansion of affective datasets and text music alignment
2023	Cross Modal AI Tri Modal Integration (Text Audio Music)	Text music alignment, semantic tagging	MusicCLIP, AudioSet Music Subset	Text music annotation [164], genre and mood tagging	Formation of joint representational space across modalities
	Audio AI Semantic Alignment and Contextual Understanding	SER, Speech Intent Recognition	Transformer, WavLM, SSL EmoSDS	Conversational AI, emotional voice assistant [285]	Semantic integration across tasks and domains
	Music AI Semantic Alignment Between Lyrics and Music	Lyric to music generation, affective mapping [286]	Transformer, Contrastive Learning [287]	Text based melody generation, affective synthesis	Semantic synchronization between words, rhythm, and harmony [239]
2024	Cross Modal AI Contrastive Pretraining and Generation	Text to audio/music generation	CLAP, AudioLDM, MusicLM	Text to sound and text to-music generation	Convergence of text conditioned generation
	Audio AI Domain Expansion and Prosody Modeling	Emotion aware TTS [288] , SpeechSQL	Stack Transformer [289], BLSTM-CRF, MAM BERT	Healthcare [290]-[292], education [293][294], industrial command systems [295][296]	Full integration between signal and semantic representation
	Music AI Generative Linguistic Musical Systems	Text conditioned music generation, lyric aware composition	MusicLM, MuseNet Lyric, MelodyGPT [297]	Generative music based on linguistic narrative	Consolidation of semantic affective modeling
2025	Cross Modal AI Context Aware Multimodal Fusion	Emotion recognition, instruction following	RCFANN BLSO, CAMFF, VOICE	Interactive multimodal dialogue, visual audio reasoning	Integration of emotional and visual context in multimodal fusion
	Audio AI Unified Semantic Representation	SpeechSQLNet 2.0, Adaptive Dialogue	CNN GRU hybrid, instruction-tuned LLMs	Voice reasoning, cognitive query translation	Semantic generalization and architectural efficiency
	Music AI Co Creative Generative Systems	Lyric conditioned composition, semantic affective generation	Udio T5, LyricFusionNet 2, Jukebox XL	Adaptive generative music and AI based music therapy	Aesthetic cognition and language emotion integration
	Cross Modal AI Integrated Multimodal Intelligence	Cross modal composition, multimodal narrative synthesis [286]	V2M, VISION, Memory Augmented Transformer	Video to music generation, contextual sentiment analysis [298][299]	Integrated processing based on cross domain semantic context [300]

4.2.1.4. Technical Evolution

The technical evolution of Audio and Music AI between 2021 and 2025 reflects a fundamental shift from feature based signal processing toward language mediated semantic and generative representation learning. In the early phase, research primarily focused on acoustic feature extraction [301][302] and emotion pattern

recognition using conventional architectures such as CNN [303],[122],[173] and LSTM [304][305]. This progression demonstrates a growing convergence among Audio AI, Music AI, and cross domain systems, where language functions as the central mechanism linking perception, representation, and expression within artificial intelligence cognition.

Table 4 presents the evolution of techniques and methodological approaches across Audio, Music, and Multimodal AI from 2021 to 2025. Overall, the trajectory moves from feature driven signal processing toward language-mediated semantic representation learning. Audio AI has evolved from acoustic recognition to Transformer [306] and SSL based linguistic comprehension systems. Music AI has transitioned from statistical modeling to generative frameworks that employ NLP to control emotional expression and musical structure. Meanwhile, Multimodal AI expands the semantic space by integrating text, sound, music, vision, and emotion within a unified representational framework. Collectively, the integration of language emerges as the key transformative factor, enabling systems not only to recognize sounds and melodies but also to interpret, generate, and construct meaning across domains.

Table 4. Evolution of Linguistic Roles in Audio, Music, and Multimodal AI (2021-2025)

Year	Domain	Dominant Approach	Key Models	Role of Language	Core Implication
2021	Audio AI	Signal level processing	CNN, RNN, CRNN	Absent; phonetic focus	Pre-semantic acoustic recognition [307]
	Music AI	Feature based emotion mapping	Seq2Seq, CNN LSTM [190][308]	Lyrics as labels	Music treated as numerical patterns
	Multimodal	Contrastive dual encoders	AudioCLIP, WavCaps	Captioning	Initial text–audio alignment
2022	Audio AI	Self-supervised learning	Wav2Vec2, HuBERT	Emerging mediator	Label-free phonetic semantics
	Music AI	Lyric-informed affect modeling	BiLSTM, CNN [309]-[311]	Emotion tagging	Early musical semantics
	Multimodal	Tri-modal fusion	MusicCLIP	Semantic binder	Linguistic musical unification
2023	Audio AI	Language-conditioned Transformers	Whisper, WavLM [312]	Parsing & decoding	Toward semantic voice understanding
	Music AI	Deep semantic representations	MusicBERT, Lyric2Vec	Meaning interpreter	Language shapes musical meaning
	Multimodal	Large-scale contrastive pretraining	CLAP, MusicLM	Generative controller	Text conditioned generation
2024	Audio AI	Context aware fine tuning	SpeechSQLNet, WavLM	Logical structure	Reasoning over spoken meaning
	Music AI	Prompt driven generation	MuseBERT, MusicLM	Semantic controller	Language governs music structure
	Multimodal	Attentive multimodal fusion	VOICE [313]	Interaction interface	Language mediates human AI context [314]
2025	Audio AI	Instruction tuned SSL	SpeechSQLNet 2.0	Reasoning module	Autonomous meaning understanding [315][316]
	Music AI	Generative diffusion	Udio-T5, Jukebox-XL	Central controller	Linguistic co-creation
	Multimodal	Memory augmented Transformers	V2M, VOICE	Semantic bridge	Language as epistemic core

4.2.1.5. Representative Datasets and Models

This section examines the evolution of representative datasets and models underlying Audio and Music AI from 2021 to 2025, highlighting a shift from signal-based corpora and convolutional architectures toward multimodal datasets and Transformer-based and Large Multimodal Models capable of unified semantic representation. As summarized in Table 5 early datasets relied on single-signal phonetic annotations such as LibriSpeech [317] and MoodyLyrics4Q [318]. From 2022 onward, paired text audio resources including AudioCaps and MusicCaps enabled explicit semantic alignment between sound and language. The 2023 2024 period saw rapid expansion into large-scale multimodal datasets, notably WavCaps and the CLAP Corpus, which integrated visual and affective dimensions for contrastive and generative learning. By 2025, hyper multimodal and co creative datasets such as LyricFusion PromptBank [319] and V2M emerged to support context-aware generation with Large Multimodal Language Models (LMLMs), underscoring the transformation of datasets from passive signal repositories into active semantic spaces where language, sound, and emotion jointly shape meaning-centered artificial intelligence.

In Audio AI, dataset evolution reflects a shift from phonetic annotation corpora toward multimodal resources enriched with linguistic and affective meaning. Early datasets such as LibriSpeech [307], CommonVoice [58], TED-LIUM3[330], and the AMI Meeting Corpus emphasized precise transcription and underpinned models like Wav2Vec2 [279] and HuBERT. From 2022, semantic oriented datasets including AudioCaps, Clotho v2, and AudioSet expanded analysis from speech recognition to meaning representation through paired audio text descriptions. Large multilingual corpora such as VoxCeleb2, MLS, and VoxPopuli further enabled cross-lingual learning in 2023 [331], followed by task- and context-specific datasets in 2024, including SpeechSQLNet, SAIL, and MAM-BERT. By 2025, datasets such as the Smart Voice Assistant Corpus and Clinical Tele Follow Up Corpus integrated signal, text, and affect, explicitly linking prosody, emotion, and cognition [332][333]. Marking a transition from signal-level annotation to semantic alignment.

Table 5. Qualitative Analysis of Dataset Evolution in Audio, Music, and Multimodal Domains (2021-2025)

Year	Domain & Dataset Characteristics	Example Datasets	Primary Functions & Applications	Research Direction & Emerging Trends
2021	Audio AI	LibriSpeech, CommonVoice, TED LIUM3, AMI	Supervised ASR and SER training through alignment	Focus on phonetic accuracy and basic linguistic labeling
	Music AI	MoodyLyrics4Q, LyricFind, MetroLyrics	Lyric emotion classification, genre tagging [320]	Early exploration of correlations between text and affective expression
	Multimodal AI	ESC 50, UrbanSound8K, GTZAN	Sound/music classification	Reliance on spectral features for signal based detection
2022	Audio AI	AudioCaps, Clotho v2, AudioSet	Audio captioning, text grounded tagging	Shift toward text grounded semantic learning
	Music AI	MELD Lyric, MIREX LyricEmotion	Lyric melody alignment, emotion-aware generation	Strengthening cross modal representation learning
	Multimodal AI	MusicCLIP subset	Text music annotation and tagging	Formation of joint representational spaces
2023	Audio AI	VoxCeleb2, MLS, VoxPopuli	Cross-lingual ASR and SER [321]	Advancements in cross lingual and transfer learning [111]
	Music AI	MusicCaps, MELD Lyric, MIREX Emotion	Lyric to music generation, semantic alignment	Enhancement of semantic embedding alignment
	Multimodal AI	WavCaps, LAION Audio	Text audio retrieval and generation	Scaling up toward universal semantic representation
2024	Audio AI Task	SpeechQL, SAIL, MAM BERT	Speech to SQL, morphology aware TTS, voice-based health AI [322]-[324]	Expansion into semantic prosodic contextualization
	Music AI	LyricSet-XL, MuseData Fusion, PolyglotLyrics	Text to music generation, lyric aware composition	Growth of prompt-based, multicultural datasets
	Multimodal AI	CLAP Corpus, AudioCLIP, IEMOCAP	Emotion recognition, multimodal captioning [325]	Fusion of visual, acoustic, and emotional features
2025	Audio AI	Smart Voice Assistant Corpus, Clinical Follow Up [326][327]	Elderly dialogue modeling, affective response detection [328]	Semantic emotional interpretation and memory modeling
	Music AI	LyricFusion PromptBank, Udio Creative Pairs	LLM-driven music generation, lyric therapy [329]	Human AI co creation guided by aesthetic preference
	Multimodal AI	V2M, VOICE, VISION	Video to music generation, multimodal reasoning	Training of Large Multimodal Language Models (LMLMs)

4.2.1.6. Architecture and Computational Mechanisms

This section examines the evolution of architectures and computational mechanisms in Audio and Music AI between 2021 and 2025, highlighting a shift from signal based models toward language- and semantics driven systems within multimodal frameworks. Through the convergence of deep neural networks [293][294], Transformer based architectures, and diffusion generative models, AI systems have progressed beyond sound and structure recognition toward semantic understanding, reasoning, and generation mediated by linguistic and

affective representations. As summarized in Table 6, early architectures were dominated by CNN RNN models [334] emphasizing acoustic feature extraction [335][336] and basic classification. Since 2022, self supervised and contrastive learning approaches have enabled explicit alignment between sound and text representations [337], narrowing the gap between perception and semantics. The 2023-2024 period marked the integration of Transformers, Large Language Models (LLMs) [338][339] and diffusion architectures to establish shared cross-modal representational spaces. By 2025, architectures evolved into language conditioned systems combining neural computation with symbolic reasoning [340][341], positioning language as the central semantic controller for adaptive and affectively coherent audio and music generation.

Table 6. Representative Models and Architectures in Audio, Music, and Multimodal AI (2021-2025)

Year	Domain	Architectural Paradigm	Representative Models	Core Advances	Research Outcomes
2021	Audio AI	Signal-based CNN RNN	CNN, CRNN, DeepSpeech 2, PANNs	MFCC/spectrogram feature learning	Phonetic accuracy; no semantic modeling
	Music AI	Sequential statistical models	CNN, LSTM, MFCC and Word2Vec	Basic lyric melody embedding	Early linguistic acoustic alignment
	Multimodal AI	Monomodal encoders	PANNs, Musicnn, Wav2Vec2	Signal to label learning	Isolated domains without cross-modal semantics
2022	Audio AI	Self supervised Transformers	Wav2Vec2, HuBERT	Masked prediction; lower WER	Shift toward semantic representations
	Music AI	Transformer enhanced modeling	BERT, CRNN, Attention	Prosody syntax emotion alignment	Context-aware musical semantics
	Multimodal AI	Dual encoder contrastive learning	CLAP, AudioCLIP	Shared latent spaces	Language as semantic mediator
2023	Audio AI	Unified encoder decoder	SpeechT5, CLAP	Multi task semantic embedding	Integrated recognition and generation
	Music AI	Generative multimodal models	MusicBERT, MuLan, Lyric2Vec	Text conditioned melody generation	Music as linguistic expression
	Multimodal AI	Text conditioned Transformers	MusicLM, Riffusion	Language driven waveform synthesis	Language controls generation
2024	Audio AI	Task oriented semantic modeling	SpeechSQLNet, MAM-BERT	Semantic prosodic integration	ASR/TTS linked to meaning
	Music AI	LLM diffusion pipelines	MusicGen, AudioLDM, T5	Cross cultural text to-music [342]	Linguistic emotion governs music
	Multimodal AI	Diffusion based alignment	CLAP, MuLan, AudioCLIP	Shared semantic spaces	Text-guided retrieval and remixing
2025	Audio AI	Symbolic neural integration	MR NLP, DAX™, Voice VR [343]	Human in the loop reasoning	Explainable systems for health [344]
	Music AI	Diffusion Transformer hybrids	Udio T5, Jukebox-XL	Instruction tuning & RLHF [345][346]	Co creative music with latent language
	Multimodal AI	Language-conditioned multistream	AudioLM, MusicLM v2, LoRA	Unified text audio music pipelines	Language as cognitive interface

The evolution of Audio AI architectures between 2021 and 2025 reflects a shift from signal based perceptual systems toward models capable of interpreting linguistic meaning and emotional nuance in sound. Early work relied on CNN, CRNN, and BiLSTM-based architectures [347], such as DeepSpeech 2 and PANNs, which processed MFCC [348] and spectrogram features for Automatic Speech Recognition (ASR) [349][350] and Speech Emotion Recognition (SER) [351]. A paradigmatic transition emerged in 2022 with self-supervised pretraining, redirecting Audio AI from pattern recognition to representation learning [352]. Models such as Wav2Vec 2.0 and HuBERT learned phonetic and linguistic structures directly from raw waveforms via masked-prediction mechanisms [353], improving cross-lingual generalization and reducing Word Error Rate (WER) by up to 20 %. Concurrently, Transformer architectures introduced attention-based alignment for speech translation and emotion recognition tasks [354]. By 2023, SpeechT5 proposed a multi-task encoder decoder Transformer [355] that unified ASR, Text-to Speech (TTS), and speaker identification within a shared latent space. In 2024, task-oriented semantic modeling became prominent. SpeechSQLNet enabled direct speech-to-structured-query conversion without an intermediate ASR stage by jointly modeling phonetic and syntactic features [356]. In TTS [357], while Stack Transformer architectures increased semantic parsing accuracy by up to 12 %. By 2025, Modified Rule-based NLP (MR NLP) framework combined linguistic rules with human-in-the-loop refinement to support ethically contextualized analysis of elderly speech in Smart

Voice Assistant systems [358]. Applications such as DAX™ achieved time efficiency gains of up to 30 % in automated medical documentation, while SVM-based voice biomarker detection for depression reached an AUC of 0.93 [359]-[361]. Extensions including the Voice VR Navigation Model and the GAD-7 Alexa Interface further expanded Audio AI into mental health monitoring and immersive interaction. A symbolic neural hybrid phase integrating linguistic rules, affective states, and human context within unified frameworks [362][363]. This progression not only improved technical performance but also expanded the cognitive capacity of Audio AI systems to interpret intent, emotion, and meaning in spoken communication.

Between 2021 and 2025, Music AI architectures evolved from sequential statistical models toward generative systems in which language functions as the primary mediator of musical meaning [364]. Early approaches relied on CNN [365] and LSTM [366] architectures with acoustic and lexical features to support basic lyric melody embedding [367] and emotion classification, enabling initial linguistic acoustic alignment. From 2022 onward, Transformer enhanced models introduced prosody syntax emotion alignment [368], allowing music representations to capture contextual and affective semantics [369]. By 2023, multimodal generative frameworks treated music as a form of linguistic expression through text-conditioned learning. This trajectory culminated in 2024-2025 with diffusion Transformer pipelines that support instruction driven, co-creative music generation [67], positioning Music AI as a semantic and affective extension of NLP rather than a purely acoustic modeling task.

Between 2021 and 2025, Multimodal AI architectures evolved from monomodal encoders toward language conditioned systems that unify audio, music, and text within shared semantic spaces [370]. This evolution culminated in 2024-2025 with diffusion based and multistreaming without crossBmodal awareness. From 2022, dual encoder contrastive architectures enabled joint latent representations, positioning language as a semantic mediator across modalities. By 2023, text-conditioned Transformer-based models supported direct control of audio and music generation through linguistic prompts [371]. This evolution culminated in 2024-2025 with diffusion-based and multistream architectures that integrate text, audio, and music within a single pipeline, establishing language as a cognitive interface that coordinates perception, generation, and cross-modal reasoning.

4.2.1.7. The Central Role of NLP as a Semantic Bridge

The findings of this review indicate that the most fundamental transformation in Audio, Music, and Multimodal AI between 2021 and 2025 is the elevation of Natural Language Processing (NLP) from a supporting component to a semantic core across domains. In early systems, language primarily served as a transcription or annotation layer. With the emergence of self-supervised and Transformer-based architectures, however, NLP became the central representational mechanism that mediates acoustic perception, contextual interpretation, and generative synthesis. In Audio AI, NLP links raw speech signals to intent, pragmatics, and affective meaning. In Music AI, language governs emotional and structural composition through text-conditioned modeling [39], positioning linguistic prompts as drivers of musical generation [59]. In Multimodal AI, NLP establishes shared semantic spaces that align text, sound, music, vision, and contextual signals within unified representations. Across domains, language no longer operates as metadata but as the organizing principle through which heterogeneous modalities are interpreted and integrated. This convergence reflects an epistemic transition from signal-level decoding to meaning-centered, multimodal intelligence. The core shift is not merely architectural but conceptual: sound is redefined as a linguistic and affective construct embedded in semantic space. NLP thus functions as the unifying explanatory axis that enables cross-modal reasoning, affective alignment, and co-creative generation in contemporary Audio and Music AI.

4.2.1.8. Increasing Multimodal Convergence

Another central finding of this review is the increasing convergence of modalities in Audio, Music, and Multimodal AI between 2021 and 2025. Early systems were largely unimodal or dual-modal, focusing on isolated processing of speech, music, or text. Audio models primarily analyzed acoustic signals, while Music AI systems processed lyrics and musical features separately. Cross-modal interaction was limited to descriptive alignment, such as text audio retrieval or captioning tasks. Over time, however, architectures evolved toward deeper multimodal integration. The emergence of contrastive learning, shared embedding spaces, and Transformer-based fusion mechanisms enabled joint modeling of text, audio, music, vision, and affect. Systems no longer treated modalities as parallel streams but as interconnected components within unified semantic frameworks. This shift is visible in the progression from dual-modality systems (text audio) to tri- and quad-modality integration (text audio music emotion), and ultimately toward penta- and hexa-modality architectures that incorporate visual context, social signals, and memory.

Importantly, this convergence is not merely technical but epistemic. Multimodal integration reflects a redefinition of intelligence in which meaning emerges through the interaction of heterogeneous signals rather than from a single modality. Language plays a stabilizing role within this convergence, providing structured symbolic grounding that aligns perception, emotion, and contextual reasoning across modalities. By 2025, multimodal systems increasingly operate as integrated ecosystems capable of cross-modal generation, adaptive dialogue, and context-aware synthesis. The trajectory indicates a move from modular AI pipelines toward cognitively inspired architectures in which modalities are dynamically coordinated within shared semantic spaces. This growing multimodal convergence thus reinforces the broader paradigm shift toward meaning-centered, language-mediated artificial intelligence.

4.2.2. Comparison with Previous Studies

Previous review studies in Audio and Music AI have typically concentrated on specific subdomains or technical perspectives. For example, surveys on music recommendation systems have focused on personalization pipelines and metadata integration, while reviews on sound event detection have emphasized recognition-oriented architectures and benchmark datasets. Other works have examined generative audio models primarily from ethical, governance, or application-driven perspectives [372]. Although these studies provide valuable domain-specific insights, their scope remains largely task-oriented and technically segmented. In contrast, the present review adopts a longitudinal and cross-domain perspective spanning 2021-2025. Rather than examining a single task (e.g., ASR, genre classification, or music generation) or a single paradigm (e.g., contrastive learning or diffusion models) [373], this study integrates Audio AI, Music AI, and Multimodal AI within a unified Domain-Modality-Technique-Task (D-M-T-T) framework. This structure enables systematic comparison across domains and reveals patterns of convergence that are not visible in isolated reviews. Moreover, while previous surveys tend to describe architectural advancements, this review advances a conceptual interpretation of the observed transformation. Specifically, it identifies the elevation of Natural Language Processing as a semantic bridge that mediates perception [280], expression [374], and generation across modalities [123]. By framing the evolution as an epistemic transition from signal-level recognition [375] to meaning-centered multimodal intelligence [376], this study extends beyond technical taxonomy toward theoretical synthesis. Therefore, compared to prior reviews, the contribution of this work lies not only in updated coverage of recent models and datasets, but in providing a unifying explanatory perspective that connects architectural evolution, modality expansion, and semantic integration within a single analytical framework.

4.2.3. Implications and Theoretical Interpretation

The findings of this review carry significant theoretical implications for the understanding of contemporary Artificial Intelligence in audio and music domains. First, the observed transition from signal-level processing to language-mediated semantic modeling suggests that intelligence in Audio and Music AI can no longer be interpreted solely through the lens of feature extraction and classification. Instead, intelligence increasingly emerges from the interaction between acoustic perception and symbolic linguistic representation. This reconfiguration positions language not as a peripheral annotation layer, but as a structural mechanism for meaning construction. Second, the increasing convergence across Audio AI, Music AI, and Multimodal AI implies a shift from domain-specific modeling toward integrative cognitive architectures. The expansion from unimodal to multi-, and eventually hyper-modal systems indicates that meaning is generated through coordinated cross-modal alignment rather than isolated signal decoding. Theoretically, this suggests that semantic coherence arises from shared representational spaces where linguistic, affective, and perceptual features are jointly embedded. Third, the elevation of NLP as a semantic bridge reframes the epistemic status of sound. Rather than being treated as a purely physical phenomenon, sound is increasingly modeled as a communicative and affective construct embedded within broader symbolic systems. This interpretation supports the view that Audio and Music AI are evolving toward cognitively inspired frameworks in which perception, reasoning, and generation are interlinked through structured language representations. Within the D-M-T-T framework, these developments demonstrate that domain evolution (Audio, Music, Multimodal), modality expansion, architectural innovation, and task diversification are not independent trajectories. Instead, they converge around a shared theoretical axis: the mediation of meaning through language. The principal theoretical implication of this study, therefore, is that the core transformation in Audio and Music AI is epistemic rather than merely technical. It reflects a redefinition of artificial intelligence from signal recognition systems to meaning-centered, multimodal semantic architectures.

4.2.4. Analysis of Gaps and Challenges

Despite substantial progress in Audio, Music, and Multimodal NLP AI between 2021 and 2025, this review identifies persistent gaps that constrain the development of meaning-centered intelligent systems at epistemic, methodological, computational, and evaluative levels. Conceptually, current architectures lack a unified framework explaining how linguistic meaning (semantic structure), affective meaning (emotion and valence), and pragmatic meaning (contextual intent) interact across modalities, as most models rely on correlation-based alignment in shared embedding spaces rather than explicitly modeling the cognitive mechanisms of meaning construction. This limitation manifests differently across domains, in Audio NLP, semantic inference remains weakly grounded in pragmatic intent; in Music NLP, affective expression is generated without explainable semantic causality; and in Multimodal AI, cross-modal alignment often obscures rather than clarifies meaning formation. Methodologically, claims of semantic or affective understanding are insufficiently operationalized, with meaning inferred from retrieval accuracy, embedding similarity, or generation fluency instead of explicitly defined semantic, affective, and contextual constructs; this issue is compounded by dataset bias and the dominance of Western linguistic and musical forms, limiting cross-cultural generalization. From a computational perspective, large-scale multimodal architectures remain resource-intensive and largely opaque, restricting interpretability, reproducibility, and ethical auditing, particularly in sensitive domains such as healthcare and education.

Evaluation practices pose a critical challenge, as widely used metrics such as BLEU, CLAP-Score, and MOS primarily assess surface alignment and perceptual quality, failing to evaluate contextual meaning preservation, emotional coherence, or pragmatic intent; consequently, evaluation remains performance-oriented rather than understanding-oriented. Importantly, these limitations are structural rather than incidental and cannot be resolved through model scaling alone. Advancing Audio, Music, and Multimodal NLP AI therefore requires a paradigm shift from optimization-driven modeling toward explicitly grounded semantic frameworks, interpretable and resource-aware architectures, culturally inclusive datasets, and validation methodologies capable of assessing genuine meaning comprehension rather than numerical correlation alone.

5. CONCLUSIONS

This study demonstrates a qualitative reorientation in Audio, Music, and Multimodal AI from signal-centered pattern recognition toward semantically mediated and generative architectures. Based on the analysis of 369 Q1-Q2 studies (2021-2025), three principal findings emerge. First, Natural Language Processing has evolved into a semantic bridge linking acoustic perception, musical expression, and cross-modal reasoning within shared representational spaces. Second, under the D-M-T-T framework, unimodal processing pipelines have progressively converged into language-conditioned multimodal systems. Third, research tasks have shifted from classification and detection toward generation, instruction-following, and contextual reasoning, indicating a structural transition from recognition-based models to meaning-oriented architectures.

The research questions are addressed as follows: (RQ1) NLP has transitioned from a transcription-support tool to a central semantic controller across domains; (RQ2) Transformer-based and self-supervised models now dominate contextual and affective representation learning; (RQ3) datasets have evolved from phonetic corpora to text-grounded and multimodal resources, although evaluation metrics remain largely performance-driven; (RQ4) applications increasingly extend to healthcare, creative industries, and human AI interaction, reflecting cross-domain semantic integration; and (RQ5) persistent challenges include interpretability limitations, cultural bias in datasets, and insufficient evaluation of semantic and affective coherence. In this context, interpretive structure refers specifically to semantic alignment, affective modeling, contextual reasoning, and cross-modal representational integration.

The principal theoretical contribution of this review lies in reframing recent developments as an epistemic transformation rather than merely an architectural progression. By positioning language as a unifying mechanism within the D-M-T-T framework, this study provides a structured cross-domain synthesis that advances conceptual clarity in Audio, Music, and Multimodal AI research. Nevertheless, several limitations must be acknowledged. The review is restricted to Scopus-indexed, English-language Q1-Q2 publications, which may exclude regionally diverse or emerging contributions. In addition, the rapid pace of multimodal model development may extend beyond the review timeframe. Ethical and legal concerns particularly copyright and artist rights in generative music systems between 2023 and 2025 were recognized as significant but fall outside the structural and technical scope of this analysis.

Looking forward, the next logical step beyond 2025 may involve real-time multimodal audio agents capable of adaptive reasoning, memory-augmented interaction, and instruction-driven generation. Advancing such systems will require interdisciplinary frameworks, culturally inclusive datasets, semantically grounded

evaluation metrics, and more interpretable multimodal architectures. By consolidating empirical trends and theoretical interpretation, this study offers a foundation for researchers seeking to design meaning-centered, context-aware, and cognitively informed AI systems.

DECLARATION

Acknowledgement

The author acknowledges the contributions of the scientific community through open-access publications that enabled this research. This study was conducted without external financial support.

REFERENCES

[1] K. Dabbabi and A. Mars, "Spoken Utterance Classification Task of Arabic Numerals and Selected Isolated Words," *Arab. J. Sci. Eng.*, vol. 47, no. 8, pp. 10731–10750, 2022, https://doi.org/10.1007/s13369-022-06649-0.

[2] L. Betti, C. Abrate, and A. Kaltenbrunner, "Large scale analysis of gender bias and sexism in song lyrics," *EPJ Data Sci.*, vol. 12, no. 1, 2023, https://doi.org/10.1140/epjds/s13688-023-00384-8.

[3] L. V Cuaya, R. Hernández-Pérez, M. Boros, A. Deme, and A. Andics, "Speech naturalness detection and language representation in the dog brain," *Neuroimage*, vol. 248, 2022, https://doi.org/10.1016/j.neuroimage.2021.118811.

[4] R. Gutierrez, J. C. Uhl, H. Schrom-Feiertag, and M. Tscheligi, "Integrating GPT-Based AI into Virtual Patients to Facilitate Communication Training Among Medical First Responders: Usability Study of Mixed Reality Simulation," *JMIR Form. Res.*, vol. 8, 2024, https://doi.org/10.2196/58623.

[5] T. Ariga and Y. Hirose, "Recognition of spoken words with mispronounced lexical prosody in Japanese," *J. Acoust. Soc. Am.*, vol. 157, no. 6, pp. 4102–4118, 2025, https://doi.org/10.1121/10.0036775.

[6] A. Derington, H. Wierstorf, A. G. Özkil, F. Eyben, F. Burkhardt, and B. W. Schuller, "Testing Correctness, Fairness, and Robustness of Speech Emotion Recognition Models," *IEEE Trans. Affect. Comput.*, vol. 16, no. 3, pp. 1929–1941, 2025, https://doi.org/10.1109/TAFFC.2025.3547218.

[7] S. Yoo, H. Lee, J. I. Song, and O. Jeong, "A Korean emotion-factor dataset for extracting emotion and factors in Korean conversations," *Sci. Rep.*, vol. 13, no. 1, 2023, https://doi.org/10.1038/s41598-023-45386-8.

[8] E. Jeong, G. Kim, and S. Kang, "Multimodal Prompt Learning in Emotion Recognition Using Context and Audio Information," *Mathematics*, vol. 11, no. 13, 2023, https://doi.org/10.3390/math11132908.

[9] Z. Mengesha, C. M. Heldreth, M. Lahav, J. Sublewski, and E. Tuennerman, "'I don't Think These Devices are Very Culturally Sensitive.'—Impact of Automated Speech Recognition Errors on African Americans," *Front. Artif. Intell.*, vol. 4, 2021, https://doi.org/10.3389/frai.2021.725911.

[10] C. Wu, S. Le Vine, E. Bengel, J. Czerwinski, and J. W. Polak, "Sentiment analysis of popular-music references to automobiles, 1950s to 2010s," *Transportation (Amst).*, vol. 49, no. 2, pp. 641–678, 2022, https://doi.org/10.1007/s11116-021-10189-1.

[11] P. M. Lindborg, L. H. Lam, Y. C. Kam, and R. Yue, "Sensory Heritage Is Vital for Sustainable Cities: A Case Study of Soundscape and Smellscape at Wong Tai Sin," *Sustain.*, vol. 17, no. 16, 2025, https://doi.org/10.3390/su17167564.

[12] Ö. Aydoğmuş, M. C. Bingol, G. Boztas, and T. Tuncer, "An automated voice command classification model based on an attention-deep convolutional neural network for industrial automation system," *Eng. Appl. Artif. Intell.*, vol. 126, 2023, https://doi.org/10.1016/j.engappai.2023.107120.

[13] A. Q. A. Hassan *et al.*, "Integrating Applied Linguistics With Artificial Intelligence-Enabled Arabic Text-To-Speech Synthesizer," *Fractals*, vol. 32, no. 9–10, 2024, https://doi.org/10.1142/S0218348X2540050X.

[14] F. Jalali-Najafabadi, C. Gadepalli, D. Jarchi, and B. M. G. Cheetham, "Acoustic analysis and digital signal processing for the assessment of voice quality," *Biomed. Signal Process. Control*, vol. 70, 2021, https://doi.org/10.1016/j.bspc.2021.103018.

[15] A. H. El Fawal, A. Mansour, and A. Nasser, "Markov-Modulated Poisson Process Modeling for Machine-to-Machine Heterogeneous Traffic," *Appl. Sci.*, vol. 14, no. 18, 2024, https://doi.org/10.3390/app14188561.

[16] N. Nixon, Y. Lin, and L. Snow, "Catalyzing Equity in STEM Teams: Harnessing Generative AI for Inclusion and Diversity," *Policy Insights from Behav. Brain Sci.*, vol. 11, no. 1, pp. 85–92, 2024, https://doi.org/10.1177/23727322231220356.

[17] C. M. G. Villame and S. A. Guirnaldo, "Design and implementation of voice-command controller for fixed-wing unmanned aerial vehicles using automatic speech recognition and natural language processing techniques," *Sustain. Eng. Innov.*, vol. 6, no. 2, pp. 199–212, 2024, https://doi.org/10.37868/sei.v6i2.id309.

[18] M. R. Islam, A. Ahmad, and M. S. Rahman, "Bangla text normalization for text-to-speech synthesizer using machine learning algorithms," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 1, 2024, https://doi.org/10.1016/j.jksuci.2023.101807.

[19] S. Baek, J. Kim, J. Lee, and M. Lee, "Implementation of a Virtual Assistant System Based on Deep Multi-modal Data Integration," *J. Signal Process. Syst.*, vol. 96, no. 3, pp. 179–189, 2024, https://doi.org/10.1007/s11265-022-01829-5.

[20] S. J. I. Sam and K. Mohamed Jasim, "Hybridization of metaheuristics and NLP approach to examine public opinion towards virtual voice assistants," *Ann. Oper. Res.*, pp. 1-32, 2024, https://doi.org/10.1007/s10479-024-06105-2.

[21] S. Dutta and J. H. Hansen, "Navigating the united states legislative landscape on voice privacy: Existing laws, proposed bills, protection for children, and synthetic data for ai," *arXiv preprint arXiv:2407.19677*, 2024, https://doi.org/10.48550/arXiv.2407.19677.
[22] Y. Wei *et al.*, "Acoustic-based machine learning approaches for depression detection in Chinese university students," *Front. Public Heal.*, vol. 13, 2025, https://doi.org/10.3389/fpubh.2025.1561332.
[23] D. Fernandes, S. Garg, M. Nikkel, and G. Guven, "A GPT-Powered Assistant for Real-Time Interaction with Building Information Models," *Buildings*, vol. 14, no. 8, 2024, https://doi.org/10.3390/buildings14082499.
[24] Z. Xu *et al.*, "Depression detection methods based on multimodal fusion of voice and text," *Sci. Rep.*, vol. 15, no. 1, 2025, https://doi.org/10.1038/s41598-025-03524-4.
[25] I. M. A. Shahin, A. B. Nassif, N. A. Al Hindawi, B. Alsabek, and N. A. AbuJabal, "Two-stage emotion recognition framework using CNN–transformer architecture and speaker cues," *Appl. Acoust.*, vol. 240, 2025, https://doi.org/10.1016/j.apacoust.2025.110963.
[26] G. Chen, Z. Qian, S. Qiu, D. Zhang, and R. Zhou, "A gated leaky integrate-and-fire spiking neural network based on attention mechanism for multi-modal emotion recognition," *Digit. Signal Process. A Rev. J.*, vol. 165, 2025, https://doi.org/10.1016/j.dsp.2025.105322.
[27] M. Alfaro-Contreras, J. M. Iñesta, and J. Calvo-Zaragoza, "Optical music recognition for homophonic scores with neural networks and synthetic music generation," *Int. J. Multimed. Inf. Retr.*, vol. 12, no. 1, 2023, https://doi.org/10.1007/s13735-023-00278-5.
[28] L. Gong and X. J. Li, "Deepfake Voice Detection: An Approach Using End-to-End Transformer with Acoustic Feature Fusion by Cross-Attention," *Electron.*, vol. 14, no. 10, 2025, https://doi.org/10.3390/electronics14102040.
[29] I. Carvalho, H. G. Gonçalo Oliveira and C. Silva, "The Importance of Context for Sentiment Analysis in Dialogues," *IEEE Access*, vol. 11, pp. 86088–86103, 2023, https://doi.org/10.1109/ACCESS.2023.3304633.
[30] M. Alfaro-Contreras and J. J. Valero-Mas, "Exploiting the two-dimensional nature of agnostic music notation for neural optical music recognition," *Appl. Sci.*, vol. 11, no. 8, 2021, https://doi.org/10.3390/app11083621.
[31] Z. Dai, H. Zhou, Q. Ba, Y. Zhou, L. Wang, and G. Li, "Improving depression prediction using a novel feature selection algorithm coupled with context-aware analysis," *J. Affect. Disord.*, vol. 295, pp. 1040–1048, 2021, https://doi.org/10.1016/j.jad.2021.09.001.
[32] N. Kumar and B. K. Baghel, "Intent Focused Semantic Parsing and Zero-Shot Learning for Out-of-Domain Detection in Spoken Language Understanding," *IEEE Access*, vol. 9, pp. 165786–165794, 2021, https://doi.org/10.1109/ACCESS.2021.3133657.
[33] A. Marijic and M. Bagic Babac, "Predicting song genre with deep learning," *Glob. Knowledge, Mem. Commun.*, vol. 74, no. 1–2, pp. 93–110, 2025, https://doi.org/10.1108/GKMC-08-2022-0187.
[34] G. Chen, Z. Qian, D. Zhang, S. Qiu, and R. Zhou, "Enhancing Robustness Against Adversarial Attacks in Multimodal Emotion Recognition with Spiking Transformers," *IEEE Access*, vol. 13, pp. 34584–34597, 2025, https://doi.org/10.1109/ACCESS.2025.3544086.
[35] Y. Li *et al.*, "Improving Text-Independent Forced Alignment to Support Speech-Language Pathologists with Phonetic Transcription," *Sensors*, vol. 23, no. 24, 2023, https://doi.org/10.3390/s23249650.
[36] M. M. Selim and M. S. Assiri, "Enhancing Arabic text-to-speech synthesis for emotional expression in visually impaired individuals using the artificial hummingbird and hybrid deep learning model," *Alexandria Eng. J.*, vol. 119, pp. 493–502, 2025, https://doi.org/10.1016/j.aej.2025.02.011.
[37] S. Li and Y. Sung, "MRBERT: Pre-Training of Melody and Rhythm for Automatic Music Generation," *Mathematics*, vol. 11, no. 4, 2023, https://doi.org/10.3390/math11040798.
[38] D. Yook, G. Han, H. Chang, and I. Yoo, "CycleDiffusion: Voice Conversion Using Cycle-Consistent Diffusion Models," *Appl. Sci.*, vol. 14, no. 20, 2024, https://doi.org/10.3390/app14209595.
[39] S. Jang and J. Lee, "User Intent-Based Music Generation Model Combining Actor-Critic Approach with MusicVAE," *IEEE Access*, vol. 13, pp. 141281–141294, 2025, https://doi.org/10.1109/ACCESS.2025.3597741.
[40] J. Min, Z. Liu, L. Wang, D. Li, M. Zhang, and Y. Huang, "Music Generation System for Adversarial Training Based on Deep Learning," *Processes*, vol. 10, no. 12, 2022, https://doi.org/10.3390/pr10122515.
[41] S. Lu and P. Wang, "Multi-dimensional fusion: transformer and GANs-based multimodal audiovisual perception robot for musical performance art," *Front. Neurorobot.*, vol. 17, 2023, https://doi.org/10.3389/fnbot.2023.1281944.
[42] L. Comanducci, P. Bestagini, and S. Tubaro, "FakeMusicCaps: A Dataset for Detection and Attribution of Synthetic Music Generated via Text-to-Music Models," *J. Imaging*, vol. 11, no. 7, 2025, https://doi.org/10.3390/jimaging11070242.
[43] S. S. Reddy, S. K. Ahmad Mnoj, and K. P. Prasada Rao, "DNNT (Deep Neural Network for Telugu): a framework for speech recognition of Telugu language with parallel computing approach," *Int. J. Speech Technol.*, vol. 28, no. 2, pp. 341–349, 2025, https://doi.org/10.1007/s10772-025-10186-0.
[44] Y. Zhao, Z. Xu, T. Zhang, M. Xie, B. Han, and Y. Liu, "Interactive Holographic Display System Based on Emotional Adaptability and CCNN-PCG," *Electron.*, vol. 14, no. 15, 2025, https://doi.org/10.3390/electronics14152981.
[45] W. Kai and K. Xing, "Video-driven musical composition using large language model with memory-augmented state space," *Vis. Comput.*, vol. 41, no. 5, pp. 3345–3357, 2025, https://doi.org/10.1007/s00371-024-03606-w.
[46] M. Ahmed, U. Rozario, M. M. Kabir, Z. Aung, J. Shin, and M. F. Mridha, "Musical Genre Classification Using Advanced Audio Analysis and Deep Learning Techniques," *IEEE Open J. Comput. Soc.*, vol. 5, pp. 457–467, 2024, https://doi.org/10.1109/OJCS.2024.3431229.

[47] F. Zeng, "Multimodal music emotion recognition method based on multi data fusion," *Int. J. Arts Technol.*, vol. 14, no. 4, pp. 271–282, 2023, https://doi.org/10.1504/IJART.2023.133662.
[48] V. S. G. S. P. Bottu and K. Ragavan, "Emotion-Based Music Recommendation System Integrating Facial Expression Recognition and Lyrics Sentiment Analysis," *IEEE Access*, vol. 13, pp. 87740–87752, 2025, https://doi.org/10.1109/ACCESS.2025.3570011.
[49] R. Phukan, N. Baruah, M. Neog, S. K. Sarma, and D. Konwar, "A Hybrid Neural-CRF Framework for Assamese Part-of-Speech Tagging," *IEEE Access*, vol. 13, pp. 160476–160489, 2025, https://doi.org/10.1109/ACCESS.2025.3609572.
[50] X. Liu *et al*., "Separate Anything You Describe," in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 458-471, 2025, https://doi.org/10.1109/TASLP.2024.3520017.
[51] A. Grgurevic and M. Bagic Babac, "Transformer-Based Approach for Solving Mathematical Problems Using Automatic Speech Recognition," *IEEE Access*, vol. 13, pp. 79845–79859, 2025, https://doi.org/10.1109/ACCESS.2025.3564121.
[52] D. Fernández-González, "Shift-Reduce Task-Oriented Semantic Parsing with Stack-Transformers," *Cognit. Comput.*, vol. 16, no. 6, pp. 2846–2862, 2024, https://doi.org/10.1007/s12559-024-10339-4.
[53] J. Zhang, Z. Wang, J. Lai, and H. Wang, "GPTArm: An Autonomous Task Planning Manipulator Grasping System Based on Vision–Language Models," *Machines*, vol. 13, no. 3, 2025, https://doi.org/10.3390/machines13030247.
[54] A. S. Khatouni, N. Seddigh, B. Nandy, and null null, "Machine Learning Based Classification Accuracy of Encrypted Service Channels: Analysis of Various Factors," *J. Netw. Syst. Manag.*, vol. 29, no. 1, 2021, https://doi.org/10.1007/s10922-020-09566-5.
[55] A. Sophia Koepke, A.-M. Oncescu, J. F. Henriques, Z. Akata, and S. Albanie, "Audio Retrieval with Natural Language Queries: A Benchmark Study," *IEEE Trans. Multimed.*, vol. 25, pp. 2675–2685, 2023, https://doi.org/10.1109/TMM.2022.3149712.
[56] B. D. Killeen, S. Chaudhary, G. M. Osgood, and M. Unberath, "Take a shot! Natural language control of intelligent robotic X-ray systems in surgery," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 19, no. 6, pp. 1165–1173, 2024, https://doi.org/10.1007/s11548-024-03120-3.
[57] X. Mu and J. He, "Virtual Teacher-Aided Learning System Based on Voice Operated Character Animation," *Appl. Sci.*, vol. 14, no. 18, 2024, https://doi.org/10.3390/app14188177.
[58] V. Brydinskyi, D. Sabodashko, Y. Khoma, M. A. Podpora, A. Konovalov, and V. V Khoma, "Enhancing Automatic Speech Recognition with Personalized Models: Improving Accuracy Through Individualized Fine-Tuning," *IEEE Access*, vol. 12, pp. 116649–116656, 2024, https://doi.org/10.1109/ACCESS.2024.3443811.
[59] S. Y. Ahn *et al.*, "How do AI and human users interact? Positioning of AI and human users in customer service," *Text Talk*, vol. 45, no. 3, pp. 301–318, 2025, https://doi.org/10.1515/text-2023-0116.
[60] M. Jelassi, K. Matteli, H. Ben Khalfallah, and J. F. Demongeot, "Enhancing Personalized Mental Health Support Through Artificial Intelligence: Advances in Speech and Text Analysis Within Online Therapy Platforms," *Inf.*, vol. 15, no. 12, 2024, https://doi.org/10.3390/info15120813.
[61] C. Qiang *et al*., "Learning Speech Representation from Contrastive Token-Acoustic Pretraining," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10196-10200, 2024, https://doi.org/10.1109/ICASSP48485.2024.10447797.
[62] K. Pothugunta, X. Liu, A. Susarla, and R. Padman, "Assessing inclusion and representativeness on digital platforms for health education: Evidence from YouTube," *J. Biomed. Inform.*, vol. 157, 2024, https://doi.org/10.1016/j.jbi.2024.104669.
[63] Y. Yao, Z. Dai, and M. Shahbaz, "Integrating international Chinese visualization teaching and vocational skills training: leveraging attention-connectionist temporal classification models," *PeerJ Comput. Sci.*, vol. 10, 2024, https://doi.org/10.7717/PEERJ-CS.2223.
[64] C. Gunasekara *et al.*, "Overview of the Ninth Dialog System Technology Challenge: DSTC9," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 4066–4076, 2024, https://doi.org/10.1109/TASLP.2024.3426331.
[65] T. HN, K. P. Gowda, R. J. R and A. J. L, "Empathy AI: Leveraging Emotion Recognition for Enhanced Human-AI Interaction," *2024 7th Asia Conference on Cognitive Engineering and Intelligent Interaction (CEII)*, pp. 233-237, 2024, https://doi.org/10.1109/CEII65291.2024.00053.
[66] K. Lim and J. Park, "Part-of-speech tagging using multiview learning," *IEEE Access*, vol. 8, pp. 185184–195196, 2020, https://doi.org/10.1109/ACCESS.2020.3033979.
[67] N. Ratnasari, A. P. Wibawa, and S. Patmanthara, "A digital hermeneutic analysis of linguistic musical meaning in generative AI using suno," *International Journal of Visual and Performing Arts*, vol. 7, no. 2, 2025, https://doi.org/10.31763/viperarts.v7i2.2326.
[68] S.-K. Choi, H.-C. Kwon, and M. Kim, "Combining Autoregressive Models and Phonological Knowledge Bases for Improved Accuracy in Korean Grapheme-to-Phoneme Conversion," *IEEE Access*, vol. 13, pp. 107678–107693, 2025, https://doi.org/10.1109/ACCESS.2025.3581981.
[69] . Sharma, A. Singh, S. Singh and G. Gupta, "AI-Powered Mock Interview Platform using Computer Vision, Natural Language Processing and Generative AI," *2025 3rd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, pp. 1258-1263, 2025, https://doi.org/10.1109/ICSSAS66150.2025.11080941.

[70] B. Crowell, "Blockchain-based Metaverse Platforms: Augmented Analytics Tools, Interconnected Decision-Making Processes, and Computer Vision Algorithms," *Linguist. Philos. Investig.*, vol. 21, pp. 121–136, 2022, https://doi.org/10.22381/lpi2120228.
[71] Supriyono, A. P. Wibawa, Suyono, and F. Kurniawan, "A survey of text summarization: Techniques, evaluation and challenges," *Nat. Lang. Process. J.*, vol. 7, no. April, p. 100070, 2024, https://doi.org/10.1016/j.nlp.2024.100070.
[72] G. Badh and T. Knowles, "Acoustic and perceptual impact of face masks on speech: A scoping review," *PLoS One*, vol. 18, no. 8 August, 2023, https://doi.org/10.1371/journal.pone.0285009.
[73] P. Visutsak, J. Loungna, S. Sopromrat, C. Jantip, P. Soponkittikunchai, and X. Liu, "Mood-Based Music Discovery: A System for Generating Personalized Thai Music Playlists Using Emotion Analysis," *Appl. Syst. Innov.*, vol. 8, no. 2, 2025, https://doi.org/10.3390/asi8020037.
[74] R. Griscom, J. A. Henry, D. Lee, R. P. Smiraglia, R. Szostak, and J. B. Young, "Classifying Musical Medium Of Performance: Object Or Property?," *Notes*, vol. 80, no. 3, pp. 455–472, 2024, https://doi.org/10.1353/not.2024.a919032.
[75] R. Flores, M. L. Tlachac, A. Shrestha, and E. A. Rundensteiner, "WavFace: A Multimodal Transformer-Based Model for Depression Screening," *IEEE J. Biomed. Heal. Informatics*, vol. 29, no. 5, pp. 3632–3641, 2025, https://doi.org/10.1109/JBHI.2025.3529348.
[76] Q. Wei, X. Huang, and Y. Zhang, "FV2ES: A Fully End2End Multimodal System for Fast Yet Effective Video Emotion Recognition Inference," *IEEE Trans. Broadcast.*, vol. 69, no. 1, pp. 10–20, 2023, https://doi.org/10.1109/TBC.2022.3215245.
[77] M. F. Naaz, K. K. Goyal, and D. K. Alwani, "Explore the Integration of Multimodal Inputs with Facial Expressions for More Comprehensive Emotion Recognition," *Commun. Appl. Nonlinear Anal.*, vol. 31, no. 8s, pp. 651–670, 2024, https://doi.org/10.52783/cana.v31.1576.
[78] B. Abibullaev, A. Keutayeva, and A. Zollanvari, "Deep Learning in EEG-Based BCIs: A Comprehensive Review of Transformer Models, Advantages, Challenges, and Applications," *IEEE Access*, vol. 11, pp. 127271–127301, 2023, https://doi.org/10.1109/ACCESS.2023.3329678.
[79] L. Pedrelli and X. Hinaut, "Hierarchical-Task Reservoir for Online Semantic Analysis from Continuous Speech," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 6, pp. 2654–2663, 2022, https://doi.org/10.1109/TNNLS.2021.3095140.
[80] M. K. A. Aljero and N. Dimililer, "Genetic Programming Approach to Detect Hate Speech in Social Media," *IEEE Access*, vol. 9, pp. 115115–115125, 2021, https://doi.org/10.1109/ACCESS.2021.3104535.
[81] J. Khan, A. Alam, and Y. Lee, "Intelligent Hybrid Feature Selection for Textual Sentiment Classification," *IEEE Access*, vol. 9, pp. 140590–140608, 2021, https://doi.org/10.1109/ACCESS.2021.3118982.
[82] V. Sornlertlamvanich and S. Yuenyong, "Thai Named Entity Recognition Using BiLSTM-CNN-CRF Enhanced by TCC," *IEEE Access*, vol. 10, pp. 53043–53052, 2022, https://doi.org/10.1109/ACCESS.2022.3175201.
[83] T. Ashihara, M. Delcroix, Y. Ijima, and M. Kashino, "Unveiling the Linguistic Capabilities of a Self-Supervised Speech Model Through Cross-Lingual Benchmark and Layer-Wise Similarity Analysis," *IEEE Access*, vol. 12, pp. 98835–98855, 2024, https://doi.org/10.1109/ACCESS.2024.3428364.
[84] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," *ACM Comput. Surv.*, vol. 54, no. 10, 2022, https://doi.org/10.1145/3505244.
[85] C.-C. Wang, M.-Y. Day, and C.-L. Wu, "Political Hate Speech Detection and Lexicon Building: A Study in Taiwan," *IEEE Access*, vol. 10, pp. 44337–44346, 2022, https://doi.org/10.1109/ACCESS.2022.3160712.
[86] M. H. Asnawi, A. A. Pravitasari, T. Herawan, and T. Hendrawati, "The Combination of Contextualized Topic Model and MPNet for User Feedback Topic Modeling," *IEEE Access*, vol. 11, pp. 130272–130286, 2023, https://doi.org/10.1109/ACCESS.2023.3332644.
[87] G. Naif Alwakid, M. Humayun, and Z. Ahmad, "Transforming Disability Into Ability: An Explainable Vision-to-Voice Image Captioning Framework Using Transformer Models and Edge Computing," *IEEE Access*, vol. 13, pp. 175212–175224, 2025, https://doi.org/10.1109/ACCESS.2025.3618646.
[88] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 3313–3332, 2023, https://doi.org/10.1109/TKDE.2021.3130191.
[89] T. Tambe *et al.*, "A 16-nm SoC for Noise-Robust Speech and NLP Edge AI Inference With Bayesian Sound Source Separation and Attention-Based DNNs," *IEEE J. Solid-State Circuits*, vol. 58, no. 2, pp. 569–581, 2023, https://doi.org/10.1109/JSSC.2022.3179303.
[90] M. R. Rajeswari and S. V Gangashetty, "Hybrid DNN-HMM-Based Approach for Telugu Language Speech Recognition," *IEEE Access*, vol. 13, pp. 122752–122768, 2025, https://doi.org/10.1109/ACCESS.2025.3588664.
[91] S. S. Malik *et al.*, "Multi-Modal Emotion Detection and Sentiment Analysis," *IEEE Access*, vol. 13, pp. 59790–59810, 2025, https://doi.org/10.1109/ACCESS.2025.3552475.
[92] K. Mao *et al.*, "Prediction of Depression Severity Based on the Prosodic and Semantic Features With Bidirectional LSTM and Time Distributed CNN," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2251–2265, 2023, https://doi.org/10.1109/TAFFC.2022.3154332.
[93] D.-W. Kim *et al.*, "Automatic Assessment of Upper Extremity Function and Mobile Application for Self-Administered Stroke Rehabilitation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 652–661, 2024, https://doi.org/10.1109/TNSRE.2024.3358497.

[94] J. Zheng, H. Wang, and J. Yao, "Building Lightweight Domain-Specific Consultation Systems via Inter-External Knowledge Fusion Contrastive Learning," *IEEE Access*, vol. 12, pp. 113244–113258, 2024, https://doi.org/10.1109/ACCESS.2024.3434648.
[95] I. Hussain, M. R. Rizvi, Z. Abbas, A. N. Cheema, and I. M. Almanjahie, "MUST: An explainable AI–based framework for MUltilingual hate Speech deTection," *IEEE Access*, 2025, https://doi.org/10.1109/ACCESS.2025.3629527.
[96] A. Kukkar, R. Mohana, A. Sharma, A. Nayyar, and M. A. Shah, "Improving Sentiment Analysis in Social Media by Handling Lengthened Words," *IEEE Access*, vol. 11, pp. 9775–9788, 2023, https://doi.org/10.1109/ACCESS.2023.3238366.
[97] J. M. Molero, J. Pérez-Martín, A. Rodrigo, and A. Peñas, "Offensive Language Detection in Spanish Social Media: Testing from Bag-of-Words to Transformers Models," *IEEE Access*, vol. 11, pp. 95639–95652, 2023, https://doi.org/10.1109/ACCESS.2023.3310244.
[98] D. Suhartono, W. Wongso, and A. Tri Handoyo, "IdSarcasm: Benchmarking and Evaluating Language Models for Indonesian Sarcasm Detection," *IEEE Access*, vol. 12, pp. 87323–87332, 2024, https://doi.org/10.1109/ACCESS.2024.3416955.
[99] I. Hussain, R. Ahmad, S. Muhammad, K. Ullah, H. Shah, and A. Namoun, "PHTI: Pashto Handwritten Text Imagebase for Deep Learning Applications," *IEEE Access*, vol. 10, pp. 113149–113157, 2022, https://doi.org/10.1109/ACCESS.2022.3216881.
[100] F. Hasnat *et al.*, "Understanding Sarcasm from Reddit texts using Supervised Algorithms," in *IEEE Region 10 Humanitarian Technology Conference, R10-HTC*, pp. 1–6, 2022, https://doi.org/10.1109/R10-HTC54060.2022.9929882.
[101] S. A. A. Ahmed, M. Awais, W. Wang, M. D. Plumbley, and J. Kittler, "ASiT: Local-Global Audio Spectrogram Vision Transformer for Event Classification," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 3684–3693, 2024, https://doi.org/10.1109/TASLP.2024.3428908.
[102] J. Wang and J. Liu, "Voice Adversarial Sample Generation Method for Ultrasonicization of Motion Noise," *IEEE Access*, vol. 12, pp. 177996–178009, 2024, https://doi.org/10.1109/ACCESS.2024.3506605.
[103] W. Chen, X. Xing, X. Xu, J. Pang, and L. Du, "SpeechFormer++: A Hierarchical Efficient Framework for Paralinguistic Speech Processing," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 775–788, 2023, https://doi.org/10.1109/TASLP.2023.3235194.
[104] M. A. P. Putra *et al.*, "Loss-Based Decentralized Federated Learning for Robust IoT Intrusion Detection System," *2024 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, pp. 119-123, 2024, https://doi.org/10.1109/IAICT62357.2024.10617474.
[105] E. Ayetiran and O. Özgöbek, "A Review of Deep Learning Techniques for Multimodal Fake News and Harmful Languages Detection," *IEEE Access*, vol. 12, pp. 76133–76153, 2024, https://doi.org/10.1109/ACCESS.2024.3406258.
[106] Y. Patel *et al.*, "Deepfake Generation and Detection: Case Study and Challenges," *IEEE Access*, vol. 11, pp. 143296–143323, 2023, https://doi.org/10.1109/ACCESS.2023.3342107.
[107] A. Hidayat, M. Sarifuddin, and null Hustinawaty, "Multi-Label Classification of Indonesian Voice Phishing Conversations: A Comparative Study of XLM-RoBERTa and ELECTRA," *J. Appl. Data Sci.*, vol. 6, no. 3, pp. 2177–2191, 2025, https://doi.org/10.47738/jads.v6i3.858.
[108] F. T. Johora, R. Hasan, S. F. Farabi, M. Z. Alam, I. Sarkar and A. A. Mahmud, "AI Advances: Enhancing Banking Security with Fraud Detection," *2024 First International Conference on Technological Innovations and Advance Computing (TIACOMP)*, pp. 289-294, 2024, https://doi.org/10.1109/TIACOMP64125.2024.00055.
[109] K. Sreelakshmi, B. Premjith, B. R. Chakravarthi, and K. P. Padannayil, "Detection of Hate Speech and Offensive Language CodeMix Text in Dravidian Languages Using Cost-Sensitive Learning Approach," *IEEE Access*, vol. 12, pp. 20064–20090, 2024, https://doi.org/10.1109/ACCESS.2024.3358811.
[110] K. Shuang, M. Gu, R. Li, J. Loo, and S. Su, "Interactive POS-aware network for aspect-level sentiment classification," *Neurocomputing*, vol. 420, pp. 181–196, 2021, https://doi.org/10.1016/j.neucom.2020.08.013.
[111] D. Aziz and D. Sztahó, "Multitask and Transfer Learning Approach for Joint Classification and Severity Estimation of Dysphonia," *IEEE J. Transl. Eng. Heal. Med.*, vol. 12, pp. 233–244, 2024, https://doi.org/10.1109/JTEHM.2023.3340345.
[112] A. Willis, W. Portlock, S. J. Lee and H. -Y. Chang, "Home Automation, Voice, and Entry Network," *2025 Systems and Information Engineering Design Symposium (SIEDS)*, pp. 191-196, 2025, https://doi.org/10.1109/SIEDS65500.2025.11021198.
[113] J. Liao, Y. Shi, and Y. Xu, "Automatic Speech Recognition Post-Processing for Readability: Task, Dataset and a Two-Stage Pre-Trained Approach," *IEEE Access*, vol. 10, pp. 117053–117066, 2022, https://doi.org/10.1109/ACCESS.2022.3219838.
[114] A. M. J. M. Z. Rahman, M. M. Kabir, M. F. Mridha, M. H. Alatiyyah, H. F. Alhasson, and S. S. Alharbi, "Arabic Speech Recognition: Advancement and Challenges," *IEEE Access*, vol. 12, pp. 39689–39716, 2024, https://doi.org/10.1109/ACCESS.2024.3376237.
[115] A.-H. Al-Ajmi and N. Al-Twairesh, "Building an Arabic Flight Booking Dialogue System Using a Hybrid Rule-Based and Data Driven Approach," *IEEE Access*, vol. 9, pp. 7043–7053, 2021, https://doi.org/10.1109/ACCESS.2021.3049732.

[116] Y.-C. Tsai and F.-C. Lin, "Paraphrase Generation Model Integrating Transformer Architecture, Part-of-Speech Features, and Pointer Generator Network," *IEEE Access*, vol. 11, pp. 30109–30117, 2023, https://doi.org/10.1109/ACCESS.2023.3260849.
[117] C. Tran and S. Sakti, "From Pixels to Voice: A Simple and Efficient End-to-End Spoken Image Description Approach via Vision Codec Language Models," *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1-5, 2025. https://doi.org/10.1109/ICASSP49660.2025.10890285.
[118] V. Hassija, A. Chakrabarti, A. Singh, V. Chamola, and B. Sikdar, "Unleashing the Potential of Conversational AI: Amplifying Chat-GPT's Capabilities and Tackling Technical Hurdles," *IEEE Access*, vol. 11, pp. 143657–143682, 2023, https://doi.org/10.1109/ACCESS.2023.3339553.
[119] S. Uruj, R. Goswami, S. D. Shetty, K. Kalaichelvi, and K. Karthikeyan, "Comparative Analysis of GPT-4 and LLaMA 3.2 Integration With Speech Processing Models for Enhancing Human–Robot Interaction and Motion Control in Real-World Applications," *IEEE Access*, vol. 13, pp. 127170–127182, 2025, https://doi.org/10.1109/ACCESS.2025.3590592.
[120] H. Ren *et al.*, "F-NIRS-Based Dynamic Functional Connectivity Reveals the Innate Musical Sensing Brain Networks in Preterm Infants," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1806–1816, 2022, https://doi.org/10.1109/TNSRE.2022.3178078.
[121] E. Pusateri *et al*., "Retrieval Augmented Correction of Named Entity Speech Recognition Errors," *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1-5, 2025, https://doi.org/10.1109/ICASSP49660.2025.10888936.
[122] K. Chauhan, K. K. Sharma, and T. Varma, "Multimodal Emotion Recognition Using Contextualized Audio Information and Ground Transcripts on Multiple Datasets," *Arab. J. Sci. Eng.*, vol. 49, no. 9, pp. 11871–11881, 2024, https://doi.org/10.1007/s13369-023-08395-3.
[123] Z. Li, Z. Li, J. Zhang, Y. Feng, and J. Zhou, "Bridging Text and Video: A Universal Multimodal Transformer for Audio-Visual Scene-Aware Dialog," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 2476–2483, 2021, https://doi.org/10.1109/TASLP.2021.3065823.
[124] Á. Gijón Flores, C. Bolaños Peño, H. Llumiguano Solano, J. Fernández-Bermejo Ruiz, F. Jesús Villanueva Molina and F. Rincón Calle, "From Voice to Shell: A SLM-Based Assistant for IoT Maintenance Tasks on the Edge," in *IEEE Internet of Things Journal*, vol. 13, no. 2, pp. 3000-3012, 2026, https://doi.org/10.1109/JIOT.2025.3632638.
[125] J. A. Ovi, M. A. Islam, and M. R. Karim, "BaNeP: An End-to-End Neural Network Based Model for Bangla Parts-of-Speech Tagging," *IEEE Access*, vol. 10, pp. 102753–102769, 2022, https://doi.org/10.1109/ACCESS.2022.3208269.
[126] S. Singh and A. Mahmood, "The NLP Cookbook: Modern Recipes for Transformer Based Deep Learning Architectures," *IEEE Access*, vol. 9, pp. 68675–68702, 2021, https://doi.org/10.1109/ACCESS.2021.3077350.
[127] S. Sevim, S. İLhan Omurca, and E. Ekinci, "ParallelCVAE: A Parallel CVAE Mechanism for Multi-Turn Dialog Response Generation Model," *IEEE Access*, vol. 13, pp. 189315–189328, 2025, https://doi.org/10.1109/ACCESS.2025.3628205.
[128] T. Gu, H. Chen, C. Bin, L. Chang, and W. Chen, "Neighborhood Attentional Memory Networks for Recommendation Systems," *Sci. Program.*, vol. 2021, 2021, https://doi.org/10.1155/2021/8880331.
[129] Z. Lin, J. He, Y. Zhao, R. Liang, H. Li, and Z. Wu, "EGRTE: adversarially training a self-explaining smoothed classifier for certified robustness," *Cybersecurity*, vol. 8, no. 1, 2025, https://doi.org/10.1186/s42400-025-00375-4.
[130] K. Yoshino *et al.*, "Overview of the Tenth Dialog System Technology Challenge: DSTC10," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 765–778, 2024, https://doi.org/10.1109/TASLP.2023.3293030.
[131] Á. L. Murciego, D. M. Jim, and N. Moreno-garc, "Context-Aware Recommender Systems in the Music Domain : A Systematic Literature Review," *Electronics*, vol. 10, no. 13, p. 1555, 2021, https://doi.org/10.3390/electronics10131555.
[132] J. Barnett, "The Ethical Implications of Generative Audio Models : A Systematic The Ethical Implications of Generative Audio Models : A Systematic Literature Review," *ACM Int. Conf. Proceeding Ser.*, pp. 146-161, 2026, https://doi.org/10.1145/3600211.3604686.
[133] M. N. Hossain Khan, J. Li, N. L. McElwain, M. Hasegawa–Johnson and B. Islam, "Sound Tagging in Infant-centric Home Soundscapes," *2024 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pp. 142-146, 2024, https://doi.org/10.1109/CHASE60773.2024.00023.
[134] W. Budiharto *et al.*, "Systematic Literature Review of The Use of Music Information Retrieval in Music Genre Classification," *Int. J. Comput. Sci. Humanit. AI*, vol. 2, no. 1, pp. 2–4, 2025, https://doi.org/10.21512/ijcshai.v2i1.13019.
[135] M. Shao, A. Basit, R. Karri, and M. Shafique, "Survey of Different Large Language Model Architectures: Trends, Benchmarks, and Challenges," *IEEE Access*, vol. 12, pp. 188664–188706, 2024, https://doi.org/10.1109/ACCESS.2024.3482107.
[136] J. Jiang, N. A. Teo, H. Pen, S. Ho, and Z. Wang, "Converting Vocal Performances into Sheet Music Leveraging Large Language Models," in *IEEE International Conference on Data Mining Workshops, ICDMW*, pp. 445–452, 2024, https://doi.org/10.1109/ICDMW65004.2024.00063.
[137] D. Pena, A. Aguilera, I. Dongo, J. Heredia, and Y. Cardinale, "A Framework to Evaluate Fusion Methods for Multimodal Emotion Recognition," *IEEE Access*, vol. 11, pp. 10218–10237, 2023, https://doi.org/10.1109/ACCESS.2023.3240420.

[138] J. Xue, Y. Deng, Y. Gao, and Y. Li, "Auffusion: Leveraging the Power of Diffusion and Large Language Models for Text-to-Audio Generation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 4700–4712, 2024, https://doi.org/10.1109/TASLP.2024.3485485.
[139] S. Luz, F. Haider, D. Fromm, I. Lazarou, I. Kompatsiaris, and B. MacWhinney, "An Overview of the ADReSS-M Signal Processing Grand Challenge on Multilingual Alzheimer's Dementia Recognition Through Spontaneous Speech," *IEEE Open J. Signal Process.*, vol. 5, pp. 738–749, 2024, https://doi.org/10.1109/OJSP.2024.3378595.
[140] V. Ponzi and C. Napoli, "Graph Neural Networks: Architectures, Applications, and Future Directions," *IEEE Access*, vol. 13, pp. 62870–62891, 2025, https://doi.org/10.1109/ACCESS.2025.3558752.
[141] S. Biswas and G. Poornalatha, "Opinion Mining Using Multi-Dimensional Analysis," *IEEE Access*, vol. 11, pp. 25906–25916, 2023, https://doi.org/10.1109/ACCESS.2023.3256521.
[142] B. Wilkes, I. Vatolkin, and H. Müller, "Statistical and visual analysis of audio, text, and image features for multi-modal music genre recognition," *Entropy*, vol. 23, no. 11, 2021, https://doi.org/10.3390/e23111502.
[143] O. Sen *et al.*, "Bangla natural language processing: A comprehensive analysis of classical, machine learning, and deep learning-based methods," *IEEE Access*, vol. 10, pp. 38999–39044, 2022, https://doi.org/10.1109/ACCESS.2022.3165563.
[144] M. Zulqarnain, R. Ghazali, M. G. Ghouse, N. A. Husaini, A. K. Z. Al-Saedi, and W. Sharif, "A comparative analysis on question classification task based on deep learning approaches," *PeerJ Comput. Sci.*, vol. 7, pp. 1–27, 2021, https://doi.org/10.7717/PEERJ-CS.570.
[145] P. Bell, J. Fainberg, O. Klejch, J. Li, S. Renals, and P. Swietojanski, "Adaptation Algorithms for Neural Network-Based Speech Recognition: An Overview," *IEEE Open J. Signal Process.*, vol. 2, pp. 33–66, 2021, https://doi.org/10.1109/OJSP.2020.3045349.
[146] M. Du, C. Liu and J. Lai, "InstantSpeech: Instant Synchronous Text-to-Speech Synthesis for LLM-driven Voice Chatbots," *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1-5, 2025. https://doi.org/10.1109/ICASSP49660.2025.10890120.
[147] S.-Y. Tseng, S. Narayanan, and P. Georgiou, "Multimodal Embeddings from Language Models for Emotion Recognition in the Wild," *IEEE Signal Process. Lett.*, vol. 28, pp. 608–612, 2021, https://doi.org/10.1109/LSP.2021.3065598.
[148] R. Yoshida, S. Yoshida, and M. MUNEYASU, "MAHGA: Multi-Aspect Heterogeneous Graph Analysis for Harmful Speech Detection on Social Networks," *IEEE Access*, vol. 13, pp. 106673–106689, 2025, https://doi.org/10.1109/ACCESS.2025.3581214.
[149] M. J. Dileep Kumar, M. Sukesh Rao and K. C. Narendra, "Multimodal Emotion Recognition: A Comprehensive Survey of Datasets, Methods, and Applications," in *IEEE Access*, vol. 13, pp. 201067-201097, 2025, https://doi.org/10.1109/ACCESS.2025.3636186.
[150] F. Casu, A. Lagorio, P. Ruiu, G. A. Trunfio, and E. Grosso, "Integrating Fine-Tuned LLM with Acoustic Features for Enhanced Detection of Alzheimer's Disease," *IEEE J. Biomed. Heal. Informatics*, 2025, https://doi.org/10.1109/JBHI.2025.3566615.
[151] Y. Sermet and I. Demir, "A semantic web framework for automated smart assistants: A case study for public health," *Big Data Cogn. Comput.*, vol. 5, no. 4, 2021, https://doi.org/10.3390/bdcc5040057.
[152] S. Shin and R. R. A. Issa, "BIMASR: Framework for Voice-Based BIM Information Retrieval," *J. Constr. Eng. Manag.*, vol. 147, no. 10, 2021, https://doi.org/10.1061/(ASCE)CO.1943-7862.0002138.
[153] M. Wairagkar *et al.*, "Conversational artificial intelligence and affective social robot for monitoring health and well-being of people with dementia," *Alzheimer's Dement.*, vol. 17, p. e053276, 2021, https://doi.org/10.1002/alz.053276.
[154] J. Park and M. Nammee, "Design and Implementation of Attention Depression Detection Model Based on Multimodal Analysis," *Sustain.*, vol. 14, no. 6, 2022, https://doi.org/10.3390/su14063569.
[155] K. Mori, "Decoding peak emotional responses to music from computational acoustic and lyrical features," *Cognition*, vol. 222, 2022, https://doi.org/10.1016/j.cognition.2021.105010.
[156] S. Salmi, S. Y. M. Mérelle, R. Gilissen, R. D. van der Mei, and S. Bhulai, "Detecting changes in help seeker conversations on a suicide prevention helpline during the COVID− 19 pandemic: in-depth analysis using encoder representations from transformers," *BMC Public Health*, vol. 22, no. 1, 2022, https://doi.org/10.1186/s12889-022-12926-2.
[157] M. R. Lima *et al.*, "Conversational Affective Social Robots for Ageing and Dementia Support," *IEEE Trans. Cogn. Dev. Syst.*, vol. 14, no. 4, pp. 1378–1397, 2022, https://doi.org/10.1109/TCDS.2021.3115228.
[158] E. J. Tan, E. Neill, J. L. Kleiner, and S. L. Rossell, "Depressive symptoms are specifically related to speech pauses in schizophrenia spectrum disorders," *Psychiatry Res.*, vol. 321, 2023, https://doi.org/10.1016/j.psychres.2023.115079.
[159] J. J. Huallpa *et al.*, "Exploring the ethical considerations of using Chat GPT in university education," *Period. Eng. Nat. Sci.*, vol. 11, no. 4, pp. 105–115, 2023, https://doi.org/10.21533/pen.v11i4.3770.
[160] M. Qasim, T. Habib, S. Urooj, and B. Mumtaz, "DESCU: Dyadic emotional speech corpus and recognition system for Urdu language," *Speech Commun.*, vol. 148, pp. 40–52, 2023, https://doi.org/10.1016/j.specom.2023.02.002.
[161] Y. Song, R. C. W. Wong, and X. Zhao, "Speech-to-SQL: toward speech-driven SQL query generation from natural language question," *VLDB J.*, vol. 33, no. 4, pp. 1179–1201, 2024, https://doi.org/10.1007/s00778-024-00837-0.

[162] R. K. Chakrawarti, J. Bansal, and P. Bansal, "Machine translation model for effective translation of Hindi poetries into English," *J. Exp. Theor. Artif. Intell.*, vol. 34, no. 1, pp. 95–109, 2022, https://doi.org/10.1080/0952813X.2020.1836033.
[163] D. Kim, W. H. Son, S. S. Kwak, T. Yun, J. Park, and J. Lee, "A Hybrid Deep Learning Emotion Classification System Using Multimodal Data," *Sensors*, vol. 23, no. 23, 2023, https://doi.org/10.3390/s23239333.
[164] R. Zheng and R. Zhang, "Classification of intelligent speech system and education method based on improved multi label transfer learning model," *Int. J. Syst. Assur. Eng. Manag.*, 2023, https://doi.org/10.1007/s13198-023-02056-2.
[165] P. Zhang, X. Huang, Y. Wang, C. Jiang, S. He, and H. Wang, "Semantic Similarity Computing Model Based on Multi Model Fine-Grained Nonlinear Fusion," *IEEE Access*, vol. 9, pp. 8433–8443, 2021, https://doi.org/10.1109/ACCESS.2021.3049378.
[166] Y. Guo and Z. Yan, "Recommended System: Attentive Neural Collaborative Filtering," *IEEE Access*, vol. 8, pp. 125953–125960, 2020, https://doi.org/10.1109/ACCESS.2020.3006141.
[167] N. Amangeldy, A. Ukenova, G. T. Bekmanova, B. Razakhova, M. Milosz, and S. A. Kudubayeva, "Continuous Sign Language Recognition and Its Translation into Intonation-Colored Speech," *Sensors*, vol. 23, no. 14, 2023, https://doi.org/10.3390/s23146383.
[168] A. J. Moshayedi, A. S. Roy, A. Kolahdooz, and S. Yang, "Deep Learning Application Pros and Cons Over Algorithm," *EAI Endorsed Trans. AI Robot.*, vol. 1, 2022, https://doi.org/10.4108/airo.v1i.19.
[169] A. Ghosh and K. Deepa, "QueryMintAI: Multipurpose Multimodal Large Language Models for Personal Data," *IEEE Access*, vol. 12, pp. 144631–144651, 2024, https://doi.org/10.1109/ACCESS.2024.3468996.
[170] T. Ben Moshe, I. Ziv, N. Dershowitz, and K. Bar, "The contribution of prosody to machine classification of schizophrenia," *Schizophrenia*, vol. 10, no. 1, 2024, https://doi.org/10.1038/s41537-024-00463-3.
[171] D. Thompson, H. Leuthold, and R. Filik, "Examining the influence of perspective and prosody on expected emotional responses to irony: Evidence from event-related brain potentials.," *Can. J. Exp. Psychol.*, vol. 75, no. 2, pp. 107–113, 2021, https://doi.org/10.1037/cep0000249.
[172] R. Hou, "Music content personalized recommendation system based on a convolutional neural network," *Soft Comput.*, vol. 28, no. 2, pp. 1785–1802, 2024, https://doi.org/10.1007/s00500-023-09457-2.
[173] T. Khan, M. Saif, and A. F. Mollah, "MuSIC: A Novel Multi-Scale Deep Neural Framework for Script Identification in the Wild," *IEEE Access*, vol. 12, pp. 166955–166976, 2024, https://doi.org/10.1109/ACCESS.2024.3494023.
[174] M. Rospocher and S. Eksir, "Assessing Fine-Grained Explicitness of Song Lyrics," *Inf.*, vol. 14, no. 3, 2023, https://doi.org/10.3390/info14030159.
[175] M. Martinc, F. Haider, S. Pollak, and S. F. Luz, "Temporal Integration of Text Transcripts and Acoustic Features for Alzheimer's Diagnosis Based on Spontaneous Speech," *Front. Aging Neurosci.*, vol. 13, 2021, https://doi.org/10.3389/fnagi.2021.642647.
[176] A. Lücking and J. Ginzburg, "Leading voices: Dialogue semantics, cognitive science and the polyphonic structure of multimodal interaction," *Lang. Cogn.*, vol. 15, no. 1, pp. 148–172, 2023, https://doi.org/10.1017/langcog.2022.30.
[177] J. Liu, C. Li, Y. Huang, and J. Han, "An intelligent medical guidance and recommendation model driven by patient-physician communication data," *Front. Public Heal.*, vol. 11, 2023, https://doi.org/10.3389/fpubh.2023.1098206.
[178] S. N., S. Wagle, P. Ghosh and K. Kishore, "Sentiment Classification of English and Hindi Music Lyrics Using Supervised Machine Learning Algorithms," *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, pp. 1-6, 2022, https://doi.org/10.1109/ASIANCON55314.2022.9908688.
[179] V. Rajan, A. Brutti and A. Cavallaro, "Is Cross-Attention Preferable to Self-Attention for Multi-Modal Emotion Recognition?," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4693-4697, 2022, https://doi.org/10.1109/ICASSP43922.2022.9746924.
[180] M. A. Azim, W. Hussein, and N. L. Badr, "Using Character-Level Sequence-to-Sequence Model for Word Level Text Generation to Enhance Arabic Speech Recognition," *IEEE Access*, vol. 11, pp. 91173–91183, 2023, https://doi.org/10.1109/ACCESS.2023.3302257.
[181] A. E. A. Kaneho, N. Zrira, K. Ouazzani-Touhami, H. A. Khan, and S. Nawaz, "Development of a bilingual healthcare chatbot for pregnant women: A comparative study of deep learning models with BiGRU optimization," *Intell. Med.*, vol. 12, 2025, https://doi.org/10.1016/j.ibmed.2025.100261.
[182] G. Ahmed and A. A. Lawaye, "CNN-based speech segments endpoints detection framework using short-time signal energy features," *Int. J. Inf. Technol.*, vol. 15, no. 8, pp. 4179–4191, 2023, https://doi.org/10.1007/s41870-023-01466-6.
[183] B. Zhang, H. Cui, V. A. Nguyen, and M. T. Whitty, "Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead," *Sensors*, vol. 25, no. 7, 2025, https://doi.org/10.3390/s25071989.
[184] S. Wu and M. Sun, "Exploring the efficacy of pre-trained checkpoints in text-to-music generation taskm," *arXiv preprint arXiv:2211.11216*, 2022, https://doi.org/10.48550/arXiv.2211.11216.
[185] Z. Wang, D. Li, R. Jiang, and M. Okumura, "Continuous Sign Language Recognition with Multi-Scale Spatial-Temporal Feature Enhancement," *IEEE Access*, vol. 13, pp. 5491–5506, 2025, https://doi.org/10.1109/ACCESS.2025.3526330.
[186] C. Singla *et al.*, "Utilizing Convolutional Neural Networks To Comprehend Sign Language And Recognize Emotions," *Fractals*, vol. 32, no. 9–10, 2024, https://doi.org/10.1142/S0218348X2540016X.

[187] A. Casas-Mas, J. I. Pozo, and I. Montero, "Oral Tradition as Context for Learning Music From 4E Cognition Compared With Literacy Cultures. Case Studies of Flamenco Guitar Apprenticeship," *Front. Psychol.*, vol. 13, 2022, https://doi.org/10.3389/fpsyg.2022.733615.
[188] W. Chen, "Deep Adversarial Neural Network Model Based on Information Fusion for Music Sentiment Analysis," *Comput. Sci. Inf. Syst.*, vol. 20, no. 4, pp. 1797–1817, 2023, https://doi.org/10.2298/CSIS221212031C.
[189] S. MacNiven, J. J. Lennon, J. Roberts, and M. MacNiven, "The language of marketing hyperbole and consumer perception–The case of Glasgow," *PLoS One*, vol. 18, no. 12, 2023, https://doi.org/10.1371/journal.pone.0295132.
[190] Y. Wang, L. Yang, and Z. Lun, "Big Data Mining Analysis Technology For Natural Language Processing Robot Design," *J. Appl. Sci. Eng.*, vol. 27, no. 12, pp. 3677–3686, 2024, https://doi.org/10.6180/jase.202412_27(12).0008.
[191] M. K. Singh, "A text independent speaker identification system using ANN, RNN, and CNN classification technique," *Multimed. Tools Appl.*, vol. 83, no. 16, pp. 48105–48117, 2024, https://doi.org/10.1007/s11042-023-17573-2.
[192] T. Kim, J. Yang, and E. Park, "MSDLF-K: A Multimodal Feature Learning Approach for Sentiment Analysis in Korean Incorporating Text and Speech," *IEEE Trans. Multimed.*, vol. 27, pp. 1266–1276, 2025, https://doi.org/10.1109/TMM.2024.3521707.
[193] Z. Shibo, H. Danke, H. Feifei, L. Liu, and X. Fei, "Application of intelligent speech analysis based on BiLSTM and CNN dual attention model in power dispatching," *Nanotechnol. Environ. Eng.*, vol. 6, no. 3, 2021, https://doi.org/10.1007/s41204-021-00148-7.
[194] A. B. Nassif, I. M. A. Shahin, S. Hamsa, N. Nemmour, and K. Hirose, "CASA-based speaker identification using cascaded GMM-CNN classifier in noisy and emotional talking conditions," *Appl. Soft Comput.*, vol. 103, 2021, https://doi.org/10.1016/j.asoc.2021.107141.
[195] S. Ayadi and Z. Lachiri, "Deep neural network architectures for audio emotion recognition performed on song and speech modalities," *Int. J. Speech Technol.*, vol. 26, no. 4, pp. 1165–1181, 2023, https://doi.org/10.1007/s10772-023-10079-0.
[196] M. Dongbo, S. Miniaoui, L. Fen, S. A. Althubiti, and T. R. Alsenani, "Intelligent chatbot interaction system capable for sentimental analysis using hybrid machine learning algorithms," *Inf. Process. Manag.*, vol. 60, no. 5, 2023, https://doi.org/10.1016/j.ipm.2023.103440.
[197] S. A. Just *et al.*, "Moving beyond word error rate to evaluate automatic speech recognition in clinical samples: Lessons from research into schizophrenia-spectrum disorders," *Psychiatry Res.*, vol. 352, 2025, https://doi.org/10.1016/j.psychres.2025.116690.
[198] A. Valladares-Poncela, P. Fraga-Lamas, and T. M. Fernández-Caramés, "On-Device Automatic Speech Recognition for Low-Resource Languages in Mixed Reality Industrial Metaverse Applications: Practical Guidelines and Evaluation of a Shipbuilding Application in Galician," *IEEE Access*, vol. 13, pp. 77017–77038, 2025, https://doi.org/10.1109/ACCESS.2025.3564137.
[199] R. G. Al-Anazi *et al.*, "Multi-Class Automated Speech Language Recognition Using Natural Language Processing With Optimal Deep Learning Model," *Fractals*, vol. 33, no. 2, 2025, https://doi.org/10.1142/S0218348X25400213.
[200] X. Wu and Q. Zhang, "Intelligent Aging Home Control Method and System for Internet of Things Emotion Recognition," *Front. Psychol.*, vol. 13, 2022, https://doi.org/10.3389/fpsyg.2022.882699.
[201] F. S. Al-Anzi, "Improved Noise-Resilient Isolated Words Speech Recognition Using Piecewise Differentiation," *Fractals*, vol. 30, no. 8, 2022, https://doi.org/10.1142/S0218348X22402277.
[202] J. Oh and J. Lee, "Multilingual Mobility: Audio-Based Language ID for Automotive Systems," *Appl. Sci.*, vol. 15, no. 16, 2025, https://doi.org/10.3390/app15169209.
[203] K. Keyvan and J. X. Huang, "How to Approach Ambiguous Queries in Conversational Search: A Survey of Techniques, Approaches, Tools, and Challenges," *ACM Comput. Surv.*, vol. 55, no. 6, 2023, https://doi.org/10.1145/3534965.
[204] S. Shah, H. Ghomeshi, E. Vakaj, E. Cooper, and S. A. Fouad, "A review of natural language processing in contact centre automation," *Pattern Anal. Appl.*, vol. 26, no. 3, pp. 823–846, 2023, https://doi.org/10.1007/s10044-023-01182-8.
[205] M. Chiu, C. Tsai, and Y. Huang, "Integrating object detection and natural language processing models to build a personalized attraction recommendation agent in a smart product service system," *Adv. Eng. Informatics*, vol. 61, 2024, https://doi.org/10.1016/j.aei.2024.102484.
[206] M. Han, D. Zhu, X. Wen, L. Shu, and Z. Yao, "Research on Dialect Protection: Interaction Design of Chinese Dialects Based on BLSTM-CRF and FBM Theories," *IEEE Access*, vol. 12, pp. 22059–22071, 2024, https://doi.org/10.1109/ACCESS.2024.3364098.
[207] M. N. Islam, S. T. Mim, T. Tasfia, and M. M. N. Hossain, "Enhancing patient treatment through automation: The development of an efficient scribe and prescribe system," *Informatics Med. Unlocked*, vol. 45, 2024, https://doi.org/10.1016/j.imu.2024.101456.
[208] S. Chang and D. Kim, "Scalable Transformer Accelerator with Variable Systolic Array for Multiple Models in Voice Assistant Applications," *Electron.*, vol. 13, no. 23, 2024, https://doi.org/10.3390/electronics13234683.
[209] C. Kooli and R. Chakraoui, "AI-driven assistive technologies in inclusive education: benefits, challenges, and policy recommendations," *Sustain. Futur.*, vol. 10, 2025, https://doi.org/10.1016/j.sftr.2025.101042.

[210] J. Liu, X. Bao, and L. Chen, "Artificial intelligence in educational technology and transformative approaches to English language using fuzzy framework with CRITIC-TOPSIS method," *Sci. Rep.*, vol. 15, no. 1, 2025, https://doi.org/10.1038/s41598-025-09844-9.
[211] A. Al Tarabsheh *et al.*, "Towards contactless learning activities during pandemics using autonomous service robots," *Appl. Sci.*, vol. 11, no. 21, 2021, https://doi.org/10.3390/app112110449.
[212] T. Mavropoulos *et al.*, "Smart integration of sensors, computer vision and knowledge representation for intelligent monitoring and verbal human-computer interaction," *J. Intell. Inf. Syst.*, vol. 57, no. 2, pp. 321–345, 2021, https://doi.org/10.1007/s10844-021-00648-7.
[213] F. Rustam *et al.*, "Automated disease diagnosis and precaution recommender system using supervised machine learning," *Multimed. Tools Appl.*, vol. 81, no. 22, pp. 31929–31952, 2022, https://doi.org/10.1007/s11042-022-12897-x.
[214] T. P. Li *et al.*, "Socratic Artificial Intelligence Learning (SAIL): The Role of a Virtual Voice Assistant in Learning Orthopedic Knowledge," *J. Surg. Educ.*, vol. 81, no. 11, pp. 1655–1666, 2024, https://doi.org/10.1016/j.jsurg.2024.08.006.
[215] L. M. Owens, J. J. Wilda, P. Y. Hahn, T. Koehler, and J. J. Fletcher, "The association between use of ambient voice technology documentation during primary care patient encounters, documentation burden, and provider burnout," *Fam. Pract.*, vol. 41, no. 2, pp. 86–91, 2024, https://doi.org/10.1093/fampra/cmad092.
[216] L. M. Owens, J. J. Wilda, R. G. Grifka, J. Westendorp, and J. J. Fletcher, "Effect of Ambient Voice Technology, Natural Language Processing, and Artificial Intelligence on the Patient-Physician Relationship," *Appl. Clin. Inform.*, vol. 15, no. 4, pp. 660–667, 2024, https://doi.org/10.1055/a-2337-4739.
[217] T. Desot, F. Portet, and M. Vacher, "End-to-End Spoken Language Understanding: Performance analyses of a voice command task in a low resource setting," *Comput. Speech Lang.*, vol. 75, 2022, https://doi.org/10.1016/j.csl.2022.101369.
[218] D. Khalil *et al.*, "An Automatic Speaker Clustering Pipeline for the Air Traffic Communication Domain," *Aerospace*, vol. 10, no. 10, 2023, https://doi.org/10.3390/aerospace10100876.
[219] J. Lee, Y. Sim, J. Kim, and Y. Suh, "EmoSDS: Unified Emotionally Adaptive Spoken Dialogue System Using Self-Supervised Speech Representations," *Futur. Internet*, vol. 17, no. 4, 2025, https://doi.org/10.3390/fi17040143.
[220] Y. Guo, Y. Liu, T. Zhou, L. Xu, and Q. Zhang, "An automatic music generation and evaluation method based on transfer learning," *PLoS One*, vol. 18, no. 5, 2023, https://doi.org/10.1371/journal.pone.0283103.
[221] L. Psyche, B. M. Kubit, Y. Ou, and E. H. Margulis, "Imaginings From an Unfamiliar World: Narrative Engagement With a New Musical System," *Psychol. Aesthetics, Creat. Arts*, 2023, https://doi.org/10.1037/aca0000629.
[222] J. Chang, J. C. S. Hung, and K. Lin, "Singability-enhanced lyric generator with music style transfer," *Comput. Commun.*, vol. 168, pp. 33–53, 2021, https://doi.org/10.1016/j.comcom.2021.01.002.
[223] Y. Li, X. Li, Z. Lou, and C. Chen, "Long Short-Term Memory-Based Music Analysis System for Music Therapy," *Front. Psychol.*, vol. 13, 2022, https://doi.org/10.3389/fpsyg.2022.928048.
[224] S. Radhakrishnan, J. M. Chatterjee, B. Pathy, and Y. Hu, "Automatic emotion recognition using deep neural network," *Multimed. Tools Appl.*, vol. 84, no. 28, pp. 33633–33662, 2025, https://doi.org/10.1007/s11042-024-20590-4.
[225] R. He, R. Liu, T. Peng, and X. Hu, "MelodyTransformer: Improving lyric-to-melody generation by considering melodic features," *Neurocomputing*, vol. 638, 2025, https://doi.org/10.1016/j.neucom.2025.130166.
[226] M. Rospocher, "Explicit song lyrics detection with subword-enriched word embeddings," *Expert Syst. Appl.*, vol. 163, 2021, https://doi.org/10.1016/j.eswa.2020.113749.
[227] S. Deepaisarn, S. Chokphantavee, S. Chokphantavee, P. Prathipasen, S. Buaruk, and V. Sornlertlamvanich, "NLP-based music processing for composer classification," *Sci. Rep.*, vol. 13, no. 1, 2023, https://doi.org/10.1038/s41598-023-40332-0.
[228] U. Singh, A. Saraswat, H. K. Azad, K. Abhishek, and S. Selvarajan, "Towards improving e-commerce customer review analysis for sentiment detection," *Sci. Rep.*, vol. 12, no. 1, 2022, https://doi.org/10.1038/s41598-022-26432-3.
[229] X. Wang, Y. Mao, X. Wu, Q. Xu, W. Jiang, and S. Yin, "An ATC instruction processing-based trajectory prediction algorithm designing," *Neural Comput. Appl.*, vol. 35, no. 32, pp. 23477–23490, 2023, https://doi.org/10.1007/s00521-021-05713-4.
[230] A. K. Alshammari *et al.*, "Applied Linguistics With Deep Learning-Based Data-Driven Text-To-Speech Synthesizer For Arabic Corpus," *Fractals*, vol. 32, no. 9–10, 2024, https://doi.org/10.1142/S0218348X25400249.
[231] D. Jia *et al.*, "VOICE: Visual Oracle for Interaction, Conversation, and Explanation," *IEEE Trans. Vis. Comput. Graph.*, vol. 31, no. 10, pp. 8828–8845, 2025, https://doi.org/10.1109/TVCG.2025.3579956.
[232] P. K. Adhikary *et al.*, "Menstrual Health Education Using a Specialized Large Language Model in India: Development and Evaluation Study of MenstLLaMA," *J. Med. Internet Res.*, vol. 27, 2025, https://doi.org/10.2196/71977.
[233] S. Li and Y. Sung, "Type-based mixture of experts and semi-supervised multi-task pre-training for symbolic music," *Expert Syst. Appl.*, vol. 292, 2025, https://doi.org/10.1016/j.eswa.2025.128613.
[234] A. Amiri, A. Ghaffarnia, N. Ghaffar Nia, D. Wu, and Y. Liang, "Harmonizer: A Universal Signal Tokenization Framework for Multimodal Large Language Models," *Mathematics*, vol. 13, no. 11, 2025, https://doi.org/10.3390/math13111819.

[235] J. Kane, M. N. Johnstone, and P. Szewczyk, "Voice Synthesis Improvement by Machine Learning of Natural Prosody," *Sensors*, vol. 24, no. 5, 2024, https://doi.org/10.3390/s24051624.
[236] L. Chen, "Visual language transformer framework for multimodal dance performance evaluation and progression monitoring," *Sci. Rep.*, vol. 15, no. 1, 2025, https://doi.org/10.1038/s41598-025-16345-2.
[237] S. García-Méndez, F. de Arriba-Pérez, F. Javier González-Castaño, J. A. Regueiro-Janeiro, and F. J. Gil-Castiñeira, "Entertainment Chatbot for the Digital Inclusion of Elderly People without Abstraction Capabilities," *IEEE Access*, vol. 9, pp. 75878–75891, 2021, https://doi.org/10.1109/ACCESS.2021.3080837.
[238] F. Ke, "Intelligent Classification Model of Music Emotional Environment Using Convolutional Neural Networks," *J. Environ. Public Health*, vol. 2022, 2022, https://doi.org/10.1155/2022/7221064.
[239] J. G. Yu, J. Zhao, L. F. Miranda-Moreno, and M. Korp, "Modular AI agents for transportation surveys and interviews: Advancing engagement, transparency, and cost efficiency," *Commun. Transp. Res.*, vol. 5, 2025, https://doi.org/10.1016/j.commtr.2025.100172.
[240] S. Yawen, "The Educational Potential Of F. Chopin's Metodological Principles Of In The Piano Training Of Chinese Students; Образовательный Потенциал Методических Принципов Ф. Шопена В Фортепианной Подготовке Китайских Студентов," *Music. Art Educ.*, vol. 11, no. 2, pp. 81–90, 2023, https://doi.org/10.31862/2309-1428-2023-11-2-81-90.
[241] R. Pugalenthi, P. A. Chakkaravarthy, J. Ramya, S. Babu, and R. Rasika Krishnan, "Artificial learning companionusing machine learning and natural language processing," *Int. J. Speech Technol.*, vol. 24, no. 3, pp. 553–560, 2021, https://doi.org/10.1007/s10772-020-09773-0.
[242] L. Chikwetu, S. B. Daily, B. J. Mortazavi, and J. P. Dunn, "Automated Diet Capture Using Voice Alerts and Speech Recognition on Smartphones: Pilot Usability and Acceptability Study," *JMIR Form. Res.*, vol. 7, 2023, https://doi.org/10.2196/46659.
[243] D. Yang, K. Ji, and T. J. Tsai, "A deeper look at sheet music composer classification using self-supervised pretraining," *Appl. Sci.*, vol. 11, no. 4, pp. 1–16, 2021, https://doi.org/10.3390/app11041387.
[244] D. Paul, A. Jain, S. Saha, and J. Mathew, "Multi-objective PSO based online feature selection for multi-label classification," *Knowledge-Based Syst.*, vol. 222, 2021, https://doi.org/10.1016/j.knosys.2021.106966.
[245] T. V Rathcke and C. Y. Lin, "Towards a comprehensive account of rhythm processing issues in developmental dyslexia," *Brain Sci.*, vol. 11, no. 10, 2021, https://doi.org/10.3390/brainsci11101303.
[246] E. Sezgin, S. A. Hussain, S. W. Rust, and Y. Huang, "Extracting Medical Information From Free-Text and Unstructured Patient-Generated Health Data Using Natural Language Processing Methods: Feasibility Study With Real-world Data," *JMIR Form. Res.*, vol. 7, 2023, https://doi.org/10.2196/43014.
[247] Y. Huang and K. C. You, "Automated Generation of Chinese Lyrics Based on Melody Emotions," *IEEE Access*, vol. 9, pp. 98060–98071, 2021, https://doi.org/10.1109/ACCESS.2021.3095964.
[248] C. Gallezot *et al.*, "Emotion expression through spoken language in Huntington disease," *Cortex*, vol. 155, pp. 150–161, 2022, https://doi.org/10.1016/j.cortex.2022.05.024.
[249] A. Osmanovic-Thunström, H. K. Carlsen, L. Ali, T. Larson, A. Hellström, and S. Steingrímsson, "Usability Comparison Among Healthy Participants of an Anthropomorphic Digital Human and a Text-Based Chatbot as a Responder to Questions on Mental Health: Randomized Controlled Trial," *JMIR Hum. Factors*, vol. 11, 2024, https://doi.org/10.2196/54581.
[250] J. Chang, "Enabling progressive system integration for AIoT and speech-based HCI through semantic-aware computing," *J. Supercomput.*, vol. 78, no. 3, pp. 3288–3324, 2022, https://doi.org/10.1007/s11227-021-03996-x.
[251] N. J. Lahiff, K. E. Slocombe, J. P. Taglialatela, V. Dellwo, and S. W. Townsend, "Degraded and computer-generated speech processing in a bonobo," *Anim. Cogn.*, vol. 25, no. 6, pp. 1393–1398, 2022, https://doi.org/10.1007/s10071-022-01621-9.
[252] S. Badrinath and H. Balakrishnan, "Automatic Speech Recognition for Air Traffic Control Communications," *Transp. Res. Rec.*, vol. 2676, no. 1, pp. 798–810, 2022, https://doi.org/10.1177/03611981211036359.
[253] Y. Taniguchi *et al.*, "Counseling (ro)bot as a use case for 5G/6G," *Complex Intell. Syst.*, vol. 8, no. 5, pp. 3899–3917, 2022, https://doi.org/10.1007/s40747-022-00664-2.
[254] H. MacFarlane, A. C. Salem, L. Chen, M. Asgari, and E. J. Fombonne, "Combining voice and language features improves automated autism detection," *Autism Res.*, vol. 15, no. 7, pp. 1288–1300, 2022, https://doi.org/10.1002/aur.2733.
[255] S. Zhang, F. Fernicola, F. Garcea, P. Bonora, and A. Barrón-Cedeño, "AriEmozione 2.0: Identifying Emotions in Opera Verses and Arias," *Ital. J. Comput. Linguist.*, vol. 8, no. 2, pp. 7–26, 2022, https://doi.org/10.4000/ijcol.1039.
[256] J. N. de Boer, H. Corona Hernández, F. Gerritse, S. G. Brederoo, F. N. K. Wijnen, and I. E. Sommer, "Negative content in auditory verbal hallucinations: a natural language processing approach," *Cogn. Neuropsychiatry*, vol. 27, no. 2–3, pp. 139–149, 2022, https://doi.org/10.1080/13546805.2021.1941831.
[257] Y. Sprotte, "Computerized text and voice analysis of patients with chronic schizophrenia in art therapy," *Sci. Rep.*, vol. 13, no. 1, 2023, https://doi.org/10.1038/s41598-023-43069-y.
[258] I. M. Hajjar *et al.*, "Development of digital voice biomarkers and associations with cognition, cerebrospinal biomarkers, and neural representation in early Alzheimer's disease," *Alzheimer's Dement. Diagnosis, Assess. Dis. Monit.*, vol. 15, no. 1, 2023, https://doi.org/10.1002/dad2.12393.
[259] H. Ning and Z. Chen, "Fusion of the word2vec word embedding model and cluster analysis for the communication of music intangible cultural heritage," *Sci. Rep.*, vol. 13, no. 1, 2023, https://doi.org/10.1038/s41598-023-49619-8.

[260] L. Jaff, S. Garg, and G. Guven, "A Novel Framework for Natural Language Interaction with 4D BIM †," *Buildings*, vol. 15, no. 11, 2025, https://doi.org/10.3390/buildings15111840.
[261] R. Trivedi, T. J. Shaw, B. L. Sheahen, C. K. Chow, and L. Laranjo, "Patient Perspectives on Conversational Artificial Intelligence for Atrial Fibrillation Self-Management: Qualitative Analysis," *J. Med. Internet Res.*, vol. 27, 2025, https://doi.org/10.2196/64325.
[262] F. Barravecchia, L. Mastrogiacomo, and F. Franceschini, "Detecting digital voice of customer anomalies to improve product quality tracking," *Int. J. Qual. Reliab. Manag.*, 2025, https://doi.org/10.1108/IJQRM-07-2024-0229.
[263] L. Chegrani, M. Guerti, and B. Bachir, "The symmetric technique of formant transition generation for use in speech synthesis in Arabic," *Int. J. Inf. Technol.*, vol. 17, no. 2, pp. 1235–1245, 2025, https://doi.org/10.1007/s41870-024-01988-7.
[264] A. Ghajari, A. Benito-Santos, S. C. Ros, V. Fresno, and E. Gonzalez-Blanco, "Test-driving information theory-based compositional distributional semantics: A case study on Spanish song lyrics," *Knowledge-Based Syst.*, vol. 319, 2025, https://doi.org/10.1016/j.knosys.2025.113549.
[265] T. T. Hien Nguyen, T. Binh Nguyen, N. Phuong Pham, Q. Do, T. Luc Le, and C. Mai Luong, "Toward human-friendly ASR systems: Recovering capitalization and punctuation for Vietnamese text," *IEICE Trans. Inf. Syst.*, vol. E104D, no. 8, pp. 1195–1203, 2021, https://doi.org/10.1587/transinf.2020BDP0005.
[266] D. Seo, H.-S. Oh, and Y. Jung, "Wav2KWS: Transfer Learning from Speech Representations for Keyword Spotting," *IEEE Access*, vol. 9, pp. 80682–80691, 2021, https://doi.org/10.1109/ACCESS.2021.3078715.
[267] M. Chen, "A Deep Learning-Based Intelligent Quality Detection Model for Machine Translation," *IEEE Access*, vol. 11, pp. 89469–89477, 2023, https://doi.org/10.1109/ACCESS.2023.3305397.
[268] W. Sharif, S. Abdullah, S. Iftikhar, D. Al-Madani, and S. Mumtaz, "Enhancing Hate Speech Detection in the Digital Age: A Novel Model Fusion Approach Leveraging a Comprehensive Dataset," *IEEE Access*, vol. 12, pp. 27225–27236, 2024, https://doi.org/10.1109/ACCESS.2024.3367281.
[269] I. Haq, W. Qiu, J. Guo, and P. Tang, "NLPashto: NLP Toolkit for Low-resource Pashto Language," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 6, pp. 1344–1352, 2023, https://doi.org/10.14569/IJACSA.2023.01406142.
[270] S. Bhat, P. Bhat, and S. V Kolekar, "From Vision to Voice: A Multi-Modal Assistive Framework for the Physically Impaired," *IEEE Access*, vol. 13, pp. 128106–128121, 2025, https://doi.org/10.1109/ACCESS.2025.3590237.
[271] C. Liu and J. Zhao, "Resource Allocation in Large Language Model Integrated 6G Vehicular Networks," *2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring)*, pp. 1-6, 2024. https://doi.org/10.1109/VTC2024-Spring62846.2024.10683673.
[272] D. Harihar, V. Shrivastava, P. Talele, and A. S. Jahagirdar, "Voice-based user interface for hands-free data entry and automation at workplaces," *MethodsX*, vol. 15, 2025, https://doi.org/10.1016/j.mex.2025.103596.
[273] V. Kadić, S. Milanović and V. Batanović, "Classification of Lyric Poetry Written in Serbian," *2024 32nd Telecommunications Forum (TELFOR)*, pp. 1-4, 2024, https://doi.org/10.1109/TELFOR63250.2024.10819147.
[274] H. Zhang, H. Huang, and H. Han, "Attention-Based Convolution Skip Bidirectional Long Short-Term Memory Network for Speech Emotion Recognition," *IEEE Access*, vol. 9, pp. 5332–5342, 2021, https://doi.org/10.1109/ACCESS.2020.3047395.
[275] D. P. Panagoulias *et al.*, "LYRICEL: Knowledge Graphs Combined With Large Language Models and Machine Learning for Cross-Cultural Analysis of Lyrics—The Case of Greek Songs," *IEEE Access*, vol. 13, pp. 141985–142006, 2025, https://doi.org/10.1109/ACCESS.2025.3597213.
[276] P. V. Terlapu, "Drinkers Voice Recognition Intelligent System: An Ensemble Stacking Machine Learning Approach," *Ann. Data Sci.*, vol. 12, no. 4, pp. 1157–1187, 2025, https://doi.org/10.1007/s40745-024-00559-8.
[277] I. Martínez-Nicolás, F. Martínez-Sánchez, O. Ivanova, and J. J. G. Meilán, "Reading and lexical–semantic retrieval tasks outperforms single task speech analysis in the screening of mild cognitive impairment and Alzheimer's disease," *Sci. Rep.*, vol. 13, no. 1, pp. 1–9, 2023, https://doi.org/10.1038/s41598-023-36804-y.
[278] G. Demiris *et al.*, "Examining spoken words and acoustic features of therapy sessions to understand family caregivers' anxiety and quality of life," *Int. J. Med. Inform.*, vol. 160, 2022, https://doi.org/10.1016/j.ijmedinf.2022.104716.
[279] O. H. Anidjar, A. Barak, B. Ben-Moshe, E. Hagai, and S. Tuvyahu, "A Stethoscope for Drones: Transformers-Based Methods for UAVs Acoustic Anomaly Detection," *IEEE Access*, vol. 11, pp. 33336–33353, 2023, https://doi.org/10.1109/ACCESS.2023.3262702.
[280] C. Yu, X. Su, and Z. Qian, "Multi-Stage Audio-Visual Fusion for Dysarthric Speech Recognition With Pre-Trained Models," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1912–1921, 2023, https://doi.org/10.1109/TNSRE.2023.3262001.
[281] C. S. C S and A. G. Ramakrishnan, "Meta-Learning for Indian Languages: Performance Analysis and Improvements with Linguistic Similarity Measures," *IEEE Access*, vol. 11, pp. 82050–82064, 2023, https://doi.org/10.1109/ACCESS.2023.3300790.
[282] J. E. Dominguez-Vidal and A. Sanfeliu, "Voice Command Recognition for Explicit Intent Elicitation in Collaborative Object Transportation Tasks: a ROS-based Implementation," in *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 412–416, 2024, https://doi.org/10.1145/3610978.3640749.
[283] C. Crocamo *et al.*, "Acoustic and Natural Language Markers for Bipolar Disorder: A Pilot, mHealth Cross-Sectional Study," *JMIR Form. Res.*, vol. 9, 2025, https://doi.org/10.2196/65555.

[284] K. Pyrovolakis, P. K. Tzouveli, and G. B. Stamou, "Multi-Modal Song Mood Detection with Deep Learning†," *Sensors*, vol. 22, no. 3, 2022, https://doi.org/10.3390/s22031065.
[285] Y. A. Kumah-Crystal *et al.*, "Vanderbilt Electronic Health Record Voice Assistant Supports Clinicians," *Appl. Clin. Inform.*, vol. 15, no. 2, pp. 199–203, 2023, https://doi.org/10.1055/a-2177-4420.
[286] F. Balabdaoui and Y. Kulagina, "Randomly generated lyrics using mixture models: the poem that Frank Sinatra never sang," *J. Math. Arts*, vol. 17, no. 3–4, pp. 391–404, 2023, https://doi.org/10.1080/17513472.2023.2186727.
[287] A. Albladi *et al.*, "Hate Speech Detection Using Large Language Models: A Comprehensive Review," *IEEE Access*, vol. 13, pp. 20871–20892, 2025, https://doi.org/10.1109/ACCESS.2025.3532397.
[288] C. Heaton and P. Mitra, "Embedding and Clustering Multi-Entity Sequences," *IEEE Access*, vol. 12, pp. 57492–57503, 2024, https://doi.org/10.1109/ACCESS.2024.3391820.
[289] Y. -N. Hung, J. -C. Wang, X. Song, W. -T. Lu and M. Won, "Modeling Beats and Downbeats with a Time-Frequency Transformer," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 401-405, 2022, https://doi.org/10.1109/ICASSP43922.2022.9747048.
[290] E. H. Law, M. J. Auil, P. A. Spears, K. Berg, and R. Winnette, "Voice Analysis of Cancer Experiences Among Patients With Breast Cancer: VOICE-BC," *J. Patient Exp.*, vol. 8, 2021, https://doi.org/10.1177/23743735211048058.
[291] S. Amini *et al.*, "Automated detection of mild cognitive impairment and dementia from voice recordings: A natural language processing approach," *Alzheimer's Dement.*, vol. 19, no. 3, pp. 946–955, 2023, https://doi.org/10.1002/alz.12721.
[292] H. Zhou *et al.*, "Extracting Complementary and Integrative Health Approaches in Electronic Health Records," *J. Healthc. Informatics Res.*, vol. 7, no. 3, pp. 277–290, 2023, https://doi.org/10.1007/s41666-023-00137-2.
[293] I. M. Oraif, "Natural Language Processing (NLP) and EFL Learning: A Case Study Based on Deep Learning," *J. Lang. Teach. Res.*, vol. 15, no. 1, pp. 201–208, 2024, https://doi.org/10.17507/jltr.1501.22.
[294] C. M. Peterson, M. Moore, N. E. Sarwani, E. Gagnon, M. Å. Bruno, and S. G. Kanekar, "Resident-faculty overnight discrepancy rates as a function of number of consecutive nights during a week of night float," *Diagnosis*, vol. 8, no. 3, pp. 368–372, 2021, https://doi.org/10.1515/dx-2020-0092.
[295] A. C. Floriano, S. L. Avila, and R. C. Fernandes, "Structured vocabulary specific to power operation control centers," *Energy Syst.*, vol. 15, no. 3, pp. 1081–1104, 2024, https://doi.org/10.1007/s12667-022-00529-0.
[296] S. Colabianchi, A. Tedeschi, and F. Costantino, "Human-technology integration with industrial conversational agents: A conceptual architecture and a taxonomy for manufacturing," *J. Ind. Inf. Integr.*, vol. 35, 2023, https://doi.org/10.1016/j.jii.2023.100510.
[297] M. Shahin, Ff. F. Chen, A. R. Hosseinzadeh, M. Maghanaki, and A. Eghbalian, "A novel approach to voice of customer extraction using GPT-3.5 Turbo: linking advanced NLP and Lean Six Sigma 4.0," *Int. J. Adv. Manuf. Technol.*, vol. 131, no. 7–8, pp. 3615–3630, 2024, https://doi.org/10.1007/s00170-024-13167-w.
[298] N. Elyamany, Y. Omar Youssef, and M. M. Hafez, "Unheard melodies and emotional peaks in Let It Go and Show Yourself: a multimodal sentiment analysis," *Lang. Semiot. Stud.*, 2025, https://doi.org/10.1515/lass-2025-0032.
[299] R. Ahamad and K. N. Mishra, "Exploring sentiment analysis in handwritten and E-text documents using advanced machine learning techniques: a novel approach," *J. Big Data*, vol. 12, no. 1, 2025, https://doi.org/10.1186/s40537-025-01064-2.
[300] S. A. Sarbadhikary, "Unravelling of the Number 16 in Corporeality, Percussion, and the Bengali Hindu Cosmos: The Experience of the Body/Mardanga," *J. Hindu Stud.*, vol. 15, no. 2, pp. 168–190, 2022, https://doi.org/10.1093/jhs/hiac006.
[301] L. H. Hansen *et al.*, "Speech- and text-based classification of neuropsychiatric conditions in a multidiagnostic setting," *Nat. Ment. Heal.*, vol. 1, no. 12, pp. 971–981, 2023, https://doi.org/10.1038/s44220-023-00152-7.
[302] N. Laaidi, A. Ezzine, M. Telmem, M. Lamrini and H. Satori, "Building a Contextualized Arabic Voice Command Corpus for Industrial Automation Systems," *2025 5th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, pp. 1-6, 2025, https://doi.org/10.1109/IRASET64571.2025.11008248.
[303] A. M. Idrees and A. L. M. Al-Solami, "A Weighted Multi-Layer Analytics Based Model for Emoji Recommendation," *Comput. Mater. Contin.*, vol. 78, no. 1, pp. 1115–1133, 2024, https://doi.org/10.32604/cmc.2023.046457.
[304] S. Saunders, F. Haider, C. W. Ritchie, G. Muniz-Terrera, and S. F. Luz, "Longitudinal observational cohort study: Speech for Intelligent cognition change tracking and DEtection of Alzheimer's Disease (SIDE-AD)," *BMJ Open*, vol. 14, no. 3, 2024, https://doi.org/10.1136/bmjopen-2023-082388.
[305] G. Kumar and S. Bhardwaj, "Biomimetic Computing for Efficient Spoken Language Identification," *Biomimetics*, vol. 10, no. 5, 2025, https://doi.org/10.3390/biomimetics10050316.
[306] A. Belouali *et al.*, "Acoustic and language analysis of speech for suicidal ideation among US veterans," *BioData Min.*, vol. 14, no. 1, 2021, https://doi.org/10.1186/s13040-021-00245-y.
[307] L. Sarı, M. A. Hasegawa-Johnson, and C. D. Yoo, "Counterfactually Fair Automatic Speech Recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3515–3525, 2021, https://doi.org/10.1109/TASLP.2021.3126949.
[308] M. H. Hsih, J. X. Yang, and C. Hsieh, "Design and Implementation of an Intelligent Web Service Agent Based on Seq2Seq and Website Crawler," *Inf.*, vol. 15, no. 12, 2024, https://doi.org/10.3390/info15120818.

[309] M. Fell, Y. Nechaev, G. Meseguer-Brocal, E. Cabrio, F. L. Gandon, and G. G. F. Peeters, "Lyrics segmentation via bimodal text-audio representation," *Nat. Lang. Eng.*, vol. 28, no. 3, pp. 317–336, 2022, https://doi.org/10.1017/S1351324921000024.
[310] S. Park and K. Chung, "MMCNet: deep learning–based multimodal classification model using dynamic knowledge," *Pers. Ubiquitous Comput.*, vol. 26, no. 2, pp. 355–364, 2022, https://doi.org/10.1007/s00779-019-01261-w.
[311] N. Herbaz, H. El Idrissi, and B. Abdelmajid, "Sign Language Recognition for Deaf and Dumb People Using Convolution Neural Network," *J. ICT Stand.*, vol. 10, no. 3, pp. 411–426, 2022, https://doi.org/10.13052/jicts2245-800X.1033.
[312] O. J. M. Peña-Cáceres, H. Silva-Marchan, M. Albert, and M. Gil, "Recognition of Human Actions through Speech or Voice Using Machine Learning Techniques," *Comput. Mater. Contin.*, vol. 77, pp. 1873–1891, 2023, https://doi.org/10.32604/cmc.2023.043176.
[313] O. Sahin, "Generation of black-box adversarial attacks using many independent objective-based algorithm for testing the robustness of deep neural networks," *Appl. Soft Comput.*, vol. 164, 2024, https://doi.org/10.1016/j.asoc.2024.111969.
[314] Y. Frommherz and A. Zarcone, "Crowdsourcing Ecologically-Valid Dialogue Data for German," *Front. Comput. Sci.*, vol. 3, 2021, https://doi.org/10.3389/fcomp.2021.686050.
[315] M. Anjum and S. Shahab, "Improving Autonomous Vehicle Controls and Quality Using Natural Language Processing-Based Input Recognition Model," *Sustain.*, vol. 15, no. 7, 2023, https://doi.org/10.3390/su15075749.
[316] S. Sivakanthan *et al.*, "Accessible autonomous transportation and services: voice of the consumer–understanding end-user priorities," *Disabil. Rehabil. Assist. Technol.*, vol. 19, no. 6, pp. 2285–2297, 2024, https://doi.org/10.1080/17483107.2023.2283066.
[317] X. Zhu, X. Qian, and D. Liang, "SSDQ: Target Speaker Extraction via Semantic and Spatial Dual Querying," *IEEE Signal Process. Lett.*, vol. 32, pp. 3167–3171, 2025, https://doi.org/10.1109/LSP.2025.3591408.
[318] J. Tobolewski, M. Sakowicz, J. Turmo, and B. Kostek, "A Bimodal Deep Model to Capture Emotions from Music Tracks," *J. Artif. Intell. Soft Comput. Res.*, vol. 15, no. 3, pp. 215–235, 2025, https://doi.org/10.2478/jaiscr-2025-0011.
[319] A. Pajon-Sanmartin, F. de Arriba-Pérez, S. García-Méndez, F. Leal, B. Malheiro, and J. Carlos Burguillo-Rial, "Unraveling Emotions With Pre-Trained Models," *IEEE Access*, vol. 13, pp. 182458–182473, 2025, https://doi.org/10.1109/ACCESS.2025.3623877.
[320] M. R. Jawad *et al.*, "Advancement of artificial intelligence techniques based lexicon emotion analysis for vaccine of COVID-19," *Period. Eng. Nat. Sci.*, vol. 9, no. 4, pp. 580–588, 2021, https://doi.org/10.21533/pen.v9i4.2383.
[321] D. T. Goomas and T. D. Ludwig, "Ergonomic improvement using natural language processing for voice-directed order selection in large industrial settings," *Hum. Factors Ergon. Manuf. Serv. Ind.*, vol. 33, no. 6, pp. 537–544, 2023, https://doi.org/10.1002/hfm.21009.
[322] A. Samojluk and P. Artiemjew, "Prototype System for Supporting Medical Diagnosis Based on Voice Interviewing," *Appl. Sci.*, vol. 15, no. 1, 2025, https://doi.org/10.3390/app15010440.
[323] R. He *et al.*, "Automated Classification of Cognitive Decline and Probable Alzheimer's Dementia Across Multiple Speech and Language Domains," *Am. J. Speech-Language Pathol.*, vol. 32, no. 5, pp. 2075–2086, 2023, https://doi.org/10.1044/2023_AJSLP-22-00403.
[324] A. L. Hartzler *et al.*, "Integrating patient voices into the extraction of social determinants of health from clinical notes: ethical considerations and recommendations," *J. Am. Med. Informatics Assoc.*, vol. 30, no. 8, pp. 1456–1462, 2023, https://doi.org/10.1093/jamia/ocad043.
[325] N. Sanguansub, P. Kamolrungwarakul, S. Poopair, K. Techaphonprasit, and T. Siriborvornratanakul, "Song lyrics recommendation for social media captions using image captioning, image emotion, and caption-lyric matching via universal sentence embedding," *Soc. Netw. Anal. Min.*, vol. 13, no. 1, 2023, https://doi.org/10.1007/s13278-023-01097-6.
[326] S. Mathur, N. Van der Vleuten, K. G. Yager, and E. H. R. Tsai, "VISION: a modular AI assistant for natural human-instrument interaction at scientific user facilities," *Mach. Learn. Sci. Technol.*, vol. 6, no. 2, 2025, https://doi.org/10.1088/2632-2153/add9e4.
[327] L. Lawson, J. W. Beaman, and M. L. Mathews, "Within Clinic Reliability and Usability of a Voice-Based Amazon Alexa Administration of the General Anxiety Disorder 7 (GAD 7)," *J. Med. Syst.*, vol. 48, no. 1, 2024, https://doi.org/10.1007/s10916-024-02086-8.
[328] D. Zhang, Z. Hu, S. Li, H. Wu, Q. M. Zhu, and G. Zhou, "More than Text: Multi-modal ChineseWord Segmentation," *Association for Computational Linguistics (ACL)*, pp. 550–557, 2021, https://doi.org/10.18653/v1/2021.acl-short.70.
[329] L. J. Chaparro-Moreno, H. Gonzalez Villasanti, L. M. Justice, J. Sun, and M. B. Schmitt, "Accuracy of automatic processing of speech-language pathologist and child talk during school-based therapy sessions," *J. Speech, Lang. Hear. Res.*, vol. 67, no. 8, pp. 2669–2684, 2024, https://doi.org/10.1044/2024_JSLHR-23-00310.
[330] L. Liu, R. Ibrahim, N. A. Ismail and W. H. Chan, "Humor Style Transfer for Virtual AI Teachers Using StyleGPT4o for Comedian Guided Script Rewriting in Virtual Educational Environments," *2025 5th International Conference on Computer Graphics, Image and Virtualization (ICCGIV)*, pp. 166-170, 2025, https://doi.org/10.1109/ICCGIV65419.2025.11085094.

[331] L. Carbone and J. J. B. Mijs, "Sounds like meritocracy to my ears: exploring the link between inequality in popular music and personal culture," *Inf. Commun. Soc.*, vol. 25, no. 5, pp. 707–725, 2022, https://doi.org/10.1080/1369118X.2021.2020870.
[332] R. G. Al-Anazi *et al.*, "An Electronic Prescribing System For Teleconsultation Using Healthcare 5.0 Innovations," *Fractals*, vol. 32, no. 9–10, 2024, https://doi.org/10.1142/S0218348X25400559.
[333] M. Asif, T. A. Khan, and W. C. Song, "Leveraging Cognitive Machine Reasoning and NLP for Automated Intent-Based Networking and e2e Service Orchestration," *IEEE Access*, vol. 13, pp. 19456–19468, 2025, https://doi.org/10.1109/ACCESS.2025.3534282.
[334] H. Wu, W. Der Jeng, L. Chen, and C. C. Ho, "Developing the NLP-QFD Model to Discover Key Success Factors of Short Videos on Social Media," *Appl. Sci.*, vol. 14, no. 11, 2024, https://doi.org/10.3390/app14114870.
[335] S. Alijani, J. Tanha, and L. Mohammadkhanli, "An ensemble of deep learning algorithms for popularity prediction of flickr images," *Multimed. Tools Appl.*, vol. 81, no. 3, pp. 3253–3274, 2022, https://doi.org/10.1007/s11042-021-11517-4.
[336] Z. Jiang and H. N. Huynh, "Unveiling music genre structure through common-interest communities," *Soc. Netw. Anal. Min.*, vol. 12, no. 1, 2025, https://doi.org/10.1007/s13278-022-00863-2.
[337] F. Almeida Do Carmo, J. L. Figueira Da Silva Junior, R. Geraldeli Rossi, and F. M. Franca Lobato, "Text representations for lyric-based identification of musical subgenres," *IEEE Lat. Am. Trans.*, vol. 21, no. 6, pp. 737–744, 2023, https://doi.org/10.1109/TLA.2023.10172139.
[338] J. T. Anibal *et al.*, "Voice EHR: introducing multimodal audio data for health," *Front. Digit. Heal.*, vol. 6, 2024, https://doi.org/10.3389/fdgth.2024.1448351.
[339] M. Xuanyuan *et al*., "Creating Multimodal Interactive Digital Twin Characters From Videos: A Dataset and Baseline," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 48, no. 1, pp. 92-108, 2026, https://doi.org/10.1109/TPAMI.2025.3603653.
[340] L. Ying *et al.*, "Reviving Static Charts Into Live Charts," *IEEE Trans. Vis. Comput. Graph.*, vol. 31, no. 8, pp. 4314–4328, 2025, https://doi.org/10.1109/TVCG.2024.3397004.
[341] M. Li, W. Ma, and Z. Chu, "User preference interaction fusion and swap attention graph neural network for recommender system," *Neural Networks*, vol. 184, 2025, https://doi.org/10.1016/j.neunet.2024.107116.
[342] E. D. Hannaford *et al.*, "Our Heritage, Our Stories: developing AI tools to link and support community-generated digital cultural heritage," *J. Doc.*, vol. 80, no. 5, pp. 1133–1147, 2024, https://doi.org/10.1108/JD-03-2024-0057.
[343] D. Kitapçıoğlu, E. E. Aksoy, A. E. Ozkan, T. Usseli, D. Çabuk, and T. Torun, "Enhancing Immersion in Virtual Reality–Based Advanced Life Support Training: Randomized Controlled Trial," *JMIR Serious Games*, vol. 13, 2025, https://doi.org/10.2196/68272.
[344] R. Bindal, P. Dewangan, R. A. Walambe, S. Ramanna, and K. V Kotecha, "Framework for Digital Medical Scribe Technology for Healthcare Documentation," *Eng. Sci.*, vol. 33, 2025, https://doi.org/10.30919/es1377.
[345] H. A. Alaqel and K. M. El-Hindi, "Improving Diacritical Arabic Speech Recognition: Transformer-Based Models with Transfer Learning and Hybrid Data Augmentation," *Inf.*, vol. 16, no. 3, 2025, https://doi.org/10.3390/info16030161.
[346] W. Cui, B. Chen, and W. Ge, "A study on in-depth perception of customer behavior in digital business halls based on multimodal data integration," *Discov. Appl. Sci.*, vol. 7, no. 4, 2025, https://doi.org/10.1007/s42452-025-06717-8.
[347] Y. A. Wubet and K. Y. Lian, "How can we detect news surrounding community safety crisis incidents in the internet? Experiments using attention-based Bi-LSTM models," *Int. J. Inf. Manag. Data Insights*, vol. 4, no. 1, 2024, https://doi.org/10.1016/j.jjimei.2024.100227.
[348] S. Ramesh, G. Swaminathan, S. Sasikala, and T. R. Saravanan, "Automatic speech emotion detection using hybrid of gray wolf optimizer and naïve Bayes," *Int. J. Speech Technol.*, vol. 26, no. 3, pp. 571–578, 2023, https://doi.org/10.1007/s10772-021-09870-8.
[349] A. Bekarystankyzy, M. Z. Orken, and T. Anarbekova, "Integrated End-to-End Automatic Speech Recognition for Languages for Agglutinative Languages," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 23, no. 6, 2024, https://doi.org/10.1145/3663568.
[350] M. Yildiz, H. K. Keskin, S. Oyucu, D. K. Hartman, M. Temur, and M. Aydoğmuş, "Can Artificial Intelligence Identify Reading Fluency and Level? Comparison of Human and Machine Performance," *Read. Writ. Q.*, vol. 41, no. 1, pp. 66–83, 2025, https://doi.org/10.1080/10573569.2024.2345593.
[351] E. Mancini, A. Galassi, F. Ruggeri, and P. Torroni, "Disruptive situation detection on public transport through speech emotion recognition," *Intell. Syst. with Appl.*, vol. 21, 2024, https://doi.org/10.1016/j.iswa.2023.200305.
[352] I. A. Al-Omari, A. A. Al-Shargabi, and M. Hadwan, "Techniques of Quran reciters recognition: a review," *IAES Int. J. Artif. Intell.*, vol. 14, no. 3, pp. 1683–1695, 2025, https://doi.org/10.11591/ijai.v14.i3.pp1683-1695.
[353] Y. Xu, F. Farha, Y. Wan, J. Xu, H. Liu, and H. Ning, "Improving completeness and consistency of co-reference annotation standard," *Wirel. Networks*, vol. 30, no. 5, pp. 4581–4590, 2024, https://doi.org/10.1007/s11276-022-03077-8.
[354] I. Attri, L. K. Kumar Awasthi, and T. P. Sharma, "Machine learning in agriculture: a review of crop management applications," *Multimed. Tools Appl.*, vol. 83, no. 5, pp. 12875–12915, 2024, https://doi.org/10.1007/s11042-023-16105-2.

[355] Y. Qin and F. Yu, "An End-To-End Speech Recognition Model for the North Shaanxi Dialect: Design and Evaluation," *Sensors*, vol. 25, no. 2, 2025, https://doi.org/10.3390/s25020341.
[356] M. Guo and L. Han, "From manual to machine," *Interpreting*, vol. 26, no. 1, pp. 24–54, 2024, https://doi.org/10.1075/intp.00100.guo.
[357] S. R. Kapse *et al.*, "MediServe: An IoT-Enhanced Deep Learning Framework for Personalized Medication Management for Elderly Care," *Comput. Mater. Contin.*, vol. 83, no. 1, pp. 935–976, 2025, https://doi.org/10.32604/cmc.2025.061981.
[358] Z. Yan *et al.*, "Understanding older people's voice interactions with smart voice assistants: a new modified rule-based natural language processing model with human input," *Front. Digit. Heal.*, vol. 6, 2024, https://doi.org/10.3389/fdgth.2024.1329910.
[359] M. Shimamoto *et al.*, "Machine learning algorithm-based estimation model for the severity of depression assessed using Montgomery-Asberg depression rating scale," *Neuropsychopharmacol. Reports*, vol. 44, no. 1, pp. 115–120, 2024, https://doi.org/10.1002/npr2.12404.
[360] F. Menne *et al.*, "The voice of depression: speech features as biomarkers for major depressive disorder," *BMC Psychiatry*, vol. 24, no. 1, 2024, https://doi.org/10.1186/s12888-024-06253-6.
[361] J. Chen *et al.*, "Multimodal digital assessment of depression with actigraphy and app in Hong Kong Chinese," *Transl. Psychiatry*, vol. 14, no. 1, 2024, https://doi.org/10.1038/s41398-024-02873-4.
[362] A. Nandal and M. Dua, "A hybrid approach to secure automatic speaker verification: integrating clone detection and speaker identification," *Int. J. Speech Technol.*, vol. 28, no. 2, pp. 411–429, 2025, https://doi.org/10.1007/s10772-025-10195-z.
[363] D. A. P. Alvarez, A. Gelbukh, and G. Sidorov, "Composer classification using melodic combinatorial n-grams[Formula presented]," *Expert Syst. Appl.*, vol. 249, 2024, https://doi.org/10.1016/j.eswa.2024.123300.
[364] C. Sur, "CRUR: coupled-recurrent unit for unification, conceptualization and context capture for language representation - a generalization of bi directional LSTM," *Multimed. Tools Appl.*, vol. 80, no. 7, pp. 9917–9959, 2021, https://doi.org/10.1007/s11042-020-09865-8.
[365] H. Yu, J. Bae, J. Choi, and H. S. Kim, "Lux: Smart mirror with sentiment analysis for mental comfort," *Sensors*, vol. 21, no. 9, 2021, https://doi.org/10.3390/s21093092.
[366] R. Rajan, J. Antony, R. A. Joseph, J. M. Thomas, C. D. H and A. C. V, "Audio-Mood Classification Using Acoustic-Textual Feature Fusion," *2021 Fourth International Conference on Microelectronics, Signals & Systems (ICMSS)*, pp. 1-6, 2021, https://doi.org/10.1109/ICMSS53060.2021.9673592.
[367] S. Choi and J. Nam, "A Melody-Unsupervision Model for Singing Voice Synthesis," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7242-7246, 2022, https://doi.org/10.1109/ICASSP43922.2022.9747422.
[368] S. Kusal, S. Patil, J. Choudrie, K. Kotecha, S. Mishra, and A. Ajith, "AI-Based Conversational Agents: A Scoping Review From Technologies to Future Directions," *IEEE Access*, vol. 10, pp. 92337–92356, 2022, https://doi.org/10.1109/ACCESS.2022.3201144.
[369] K. Kogila and M. Sadanandam, "Emotion Recognition from Speech Utterances with Hybrid Spectral Features Using Machine Learning Algorithms," *Trait. du Signal*, vol. 39, no. 2, pp. 603–609, 2022, https://doi.org/10.18280/ts.390222.
[370] I. Manco, E. Benetos, E. Quinton and G. Fazekas, "MusCaps: Generating Captions for Music Audio," *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8, 2021, https://doi.org/10.1109/IJCNN52387.2021.9533461.
[371] V. Knappich and T. P. Schrader, "Controllable Active-Passive Voice Generation using Prefix Tuning," in *International Conference Recent Advances in Natural Language Processing, RANLP*, pp. 23–32, 2023, https://doi.org/10.26615/issn.2603-2821.2023_003.
[372] G. Franceschelli and M. Musolesi, "On the creativity of large language models," *AI Soc.*, vol. 40, no. 5, pp. 3785–3795, 2025, https://doi.org/10.1007/s00146-024-02127-3.
[373] D. Niizumi, D. Takeuchi, M. Yasuda, B. Thien Nguyen, Y. Ohishi, and N. Harada, "M2D-CLAP: Exploring General-Purpose Audio-Language Representations Beyond CLAP," *IEEE Access*, vol. 13, pp. 163313–163330, 2025, https://doi.org/10.1109/ACCESS.2025.3611348.
[374] X. Wang, J. He, Z. Jin, M. Yang, Y. Wang, and H. Qu, "M2Lens: Visualizing and Explaining Multimodal Models for Sentiment Analysis," *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 1, pp. 802–812, 2022, https://doi.org/10.1109/TVCG.2021.3114794.
[375] Z. Huang, J. Epps, D. Joachim, and V. Sethu, "Natural Language Processing Methods for Acoustic and Landmark Event-Based Features in Speech-Based Depression Detection," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 2, pp. 435–448, 2020, https://doi.org/10.1109/JSTSP.2019.2949419.
[376] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal Intelligence: Representation Learning, Information Fusion, and Applications," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 3, pp. 478–493, 2020, https://doi.org/10.1109/JSTSP.2020.2987728.