International Journal of Science and Technology

International Journal of Science and Technology, 2025, 33–38

doi: 10.70728/tech.v2.i07.012
Volume 02, Issue 07
ISSN: 3030-3443
Paper

PAPER

MULTIMODAL SPEECH PROCESSING

Akramov Azizbek Akmal ugli [1], and Rakhimberdiev Sanjarbek Alisher ugli[2]

[1]TUIT, Department of Convergence of Digital Technologies, Master's degree and [2]TUIT, Department of Convergence of Digital Technologies, Master's degree

*akramovazizbek21@gmail.com, raximberdiyevsanjarbek@gmail.com

Abstract

The aim of this research is to evaluate the effectiveness of multimodal speech processing techniques in improving speech recognition accuracy, addressing the key issue of integrating audio, visual, and contextual cues in real-time applications; to solve this problem, data is required from diverse speech datasets that encompass various modalities, including video input, audio recordings, and contextual language information.
This dissertation examines the effectiveness of multimodal speech processing techniques in enhancing speech recognition accuracy, focusing particularly on the integration of audio, visual, and contextual cues for real-time applications. The research identifies significant challenges in the current speech recognition systems and emphasizes the necessity of utilizing diverse speech datasets that incorporate modalities such as video input, audio recordings, and contextual language information. Key findings reveal that incorporating visual cues along with auditory signals considerably improves recognition rates, particularly in environments with ambient noise, leading to a marked reduction in misunderstanding critical messages in communication. The implications of these findings are particularly significant within the healthcare sector, where effective communication between patients and providers is essential for accurate diagnosis and treatment. Improved speech recognition capabilities can enhance telemedicine services, assistive technologies for individuals with speech impairments, and overall patient-provider interactions, ultimately leading to better health outcomes. This study not only contributes to the existing body of knowledge in speech processing but also highlights the potential for transformative applications in healthcare, encouraging future research and development that leverages multimodal approaches to enhance communication efficacy in clinical settings.

Introduction

Significant advancements in speech processing technologies have been shaping the interface between humans and machines, ultimately transforming communication dynamics in various fields, including healthcare, education, and entertainment. The ability to interpret and process speech effectively has profound implications, particularly as interactions increasingly occur in noisy environments or through multiple channels, such as visual cues accompanying auditory signals.

Current speech recognition systems predominantly rely on audio inputs, often overlooking the complexity introduced by multimodal interactions, which integrate various forms of data—including visual stimuli and contextual information. This limitation presents an ongoing research problem:how to optimize multimodal speech processing techniques to enhance speech recognition accuracy in real-time applications, thereby improving user experiences and communication efficacy.

Consequently, the objectives of this research include developing innovative frameworks for multimodal speech processing, integrating audio, visual, and contextual cues to bolster recognition rates, and addressing masking effects encountered in environments with background noise. Furthermore, this analysis seeks to explore the synergies between different modalities to construct more robust algorithms that can learn from diverse datasets, thus expediting the transition from traditional speech processing to more holistic, multimodal systems [1][2][3]. The significance of this research extends beyond theoretical implications; it has actionable consequences in clinical settings where precise communication is vital, such as telemedicine, where misunderstandings can lead to adverse health outcomes.

By determining effective methods for integrating multimodal inputs, the potential to develop applications that enhance assistive technologies for individuals with speech impairments, improve accessibility in diverse communication environments, and foster empathetic human-computer interactions is substantial [4][5][6]. Moreover, this exploration lays the groundwork for future research that could lead to breakthroughs in artificial intelligence, enabling machines to engage in more naturalistic dialogues with humans, thereby addressing complex emotional

Compiled on: May 16, 2025.
Manuscript prepared by the author.

and contextual nuances within conversations [7][8][9][10].

Literature Review

The ability to process spoken language seamlessly integrates a variety of cognitive mechanisms, thereby facilitating effective communication. The advent of multimodal processing frameworks has expanded our understanding of how humans access, interpret, and respond to auditory stimuli in conjunction with other sensory modalities. This scholarly inquiry holds significant implications across various fields, including linguistics, psychology, and artificial intelligence. Existing literature has extensively examined the mechanisms underlying the integration of auditory information with visual and contextual cues, revealing that multimodal interactions greatly enhance language comprehension and processing efficiency [1].

Several studies have highlighted the role of technological advancements in the study of speech processing, where researchers have utilized machine learning and neural networks to model human-like language understanding [2]. However, these frameworks often overlook the nuanced cognitive processes that govern individual differences in multimodal speech processing [3], indicating a critical gap in the literature that requires further examination.The themes emerging from current research underscore the interplay between cognitive load, sensory input, and context in shaping speech perception.

For instance, work by [4] identifies how environmental factors can modulate attentional resources, thereby impacting auditory processing. Additionally, the integration of non-verbal communication signals, such as facial expressions or gestures, has been shown to significantly enhance the clarity and meaning derived from spoken language [5]. This aligns with findings from [6], where the researchers argue that an individual's visual attention can redirect auditory processing pathways, suggesting a highly interactive model of communication. Despite these advances, significant gaps persist, particularly concerning the underlying neural mechanisms that facilitate these multimodal integrations [7].

Moreover, much of the existing research has been predominantly focused on specific demographic groups, inadequately representing the diversity of experiences and processing abilities among different populations [8].Furthermore, while prior studies have greatly contributed to our understanding of typical speech processing, less attention has been given to atypical populations, such as individuals with hearing impairments or cognitive disorders [9]. This highlights a poignant need to explore how multimodal processing frameworks can be adapted to meet the needs of varied communicative contexts, including educational environments and interpersonal interactions [10].

Another area requiring further exploration is the role of culture in shaping multimodal speech processing strategies; research has largely overlooked how cultural variations influence the interpretation of multimodal information [11]. In light of these observations, the present review seeks to synthesize existing findings into a coherent understanding of multimodal speech processing, outlining key theoretical frameworks and methodologies while highlighting the pressing gaps that remain.

By doing so, the literature review aims to pave the way for more targeted research that encompasses a wider spectrum of human communication, as well as fostering interdisciplinary collaboration that can address the complexities of multimodal interactions [12]. Ultimately, this review aspires to not only enrich the academic discourse surrounding speech processing mechanisms but also contribute to applied practices in educational and therapeutic settings [13]. As we delve into the existing literature, it becomes increasingly evident that multimodal processing is not merely an enhancement to speech comprehension; rather, it is fundamental to our understanding of human cognition and communication in an increasingly complex world [14][15][16][17][18][19][20].

The exploration of multimodal speech processing has evolved significantly over the decades, illustrating a dynamic interplay between technological advancements and theoretical frameworks. Early investigations into speech processing were largely focused on auditory mechanisms, with foundational works setting the stage for subsequent research by examining how humans perceive sound in isolation [1]. As computational methods developed, scholars began to incorporate visual elements, leading to a richer understanding of speech as a multifaceted experience. Notably, studies in the late 1990s highlighted the importance of audiovisual integration, demonstrating that visual cues could enhance speech recognition in noisy environments [2][3].Further, the 2000s witnessed a surge in empirical studies that underscored the significance of contextual information in speech processing. Research during this period emphasized how situational factors, such as the speakers facial expressions and gestures, contribute to the interpretation of spoken language [4][5].

This era marked a pivotal shift toward viewing speech recognition as an inherently multimodal process, prompting investigations that interlinked auditory and visual stimuli more closely [6].By the 2010s, advances in machine learning and neural networks catalyzed new paradigms in multimodal speech processing, allowing for more sophisticated modeling of how humans integrate various sensory inputs [7][8].

Contemporary research continues to explore this integration, with a growing emphasis on real-world applications, such as assistive technologies for individuals with hearing impairments [9]. Thus, the lineage of multimodal speech processing exemplifies an ongoing quest to understand and replicate the complexities of human communication through an ever-expanding lens. The exploration of multimodal speech processing reveals a rich tapestry of themes that underscore the complexity of effective communication. One prominent theme is the integration of auditory and visual modalities, which has been shown to enhance speech perception significantly.

Research indicates that when visual cues are available, individuals demonstrate improved understanding of speech in noisy environments, highlighting the importance of context in auditory processing [1][2]. This synchronicity between hearing and seeing not only facilitates comprehension but also reflects the brains inherent ability to merge sensory information, a point emphasized across various studies [3][4].Another essential theme is the role of technology in advancing multimodal speech processing. Advances in machine learning and artificial intelligence have opened new avenues for understanding speech through a multimodal lens, allowing for more sophisticated interpretations and applications in fields such as linguistics, psychology, and human-computer interaction [5][6].

This technological dimension is crucial, as it bridges the gap between theoretical research and practical applications, particularly in enhancing communication devices that assist individuals with hearing impairments [7][8].Furthermore, the cognitive mechanisms underlying multimodal integration are crucial to grasp. Several studies have demonstrated how the brain processes and prioritizes information from different modalities, suggesting a hierarchical organization that favors certain types of stimuli over others [9][10]. These findings underscore a deeper understanding of how empirical evidence can inform theories of perception and cognition, illustrating the intersection of sensory inputs and cognitive functions [11][12]. Overall, the literature emphasizes a multidisciplinary approach that weaves together technology, cognitive science, and sensory integration, promoting a comprehensive understanding of multimodal speech processing.

In exploring the multimodal processing of speech, a

variety of methodological approaches have emerged, each contributing distinct insights to the field. Notably, some researchers have emphasized the importance of integrating auditory and visual stimuli to understand speech perception more accurately. This integration has been supported by findings that demonstrate how visual cues enhance auditory processing, suggesting a synergistic relationship that is crucial for effective communication [1]. Moreover, recent studies leveraging neuroimaging techniques have provided robust evidence regarding the neural underpinnings of multimodal speech processing, highlighting areas of the brain that activate when both auditory and visual information is present [2][3].

In contrast, other methodological perspectives have prioritized the examination of individual components of speech processing in isolation, revealing nuanced details about how speech elements operate independently. This approach has illuminated the complexities inherent in factors such as intonation and prosody, which can dramatically alter meaning when processed without accompanying visual context [4][5]. Additionally, computational modeling has emerged as a vital method, simulating the cognitive processes involved in speech perception to predict how multimodal information is integrated or misinterpreted [6][7].

Ultimately, the interplay of these methodologies highlights the multidimensional nature of speech processing. By examining the contributions of various approaches—ranging from neural imaging to behavioral studies and computational models—researchers can construct a more comprehensive understanding of how multimodal factors influence communication dynamics [8][9][10]. Each methodological lens offers valuable perspectives, underscoring the complexity and richness of speech processing in real-world contexts. The exploration of multimodal speech processing engages various theoretical perspectives, each contributing unique insights into the neural mechanisms and cognitive processes involved. A significant body of literature emphasizes the integration of auditory and visual information, highlighting the role of cross-modal interactions in enhancing speech comprehension. For instance, studies have demonstrated that visual cues, such as lip movements, significantly improve speech recognition in noisy environments, underscoring the effectiveness of multimodal processing in real-world communication settings [1][2].

Moreover, theoretical frameworks such as the dual-coding theory support the notion that the simultaneous processing of auditory and visual signals can create richer memory traces, thus facilitating better recall and understanding [3][4]. This aligns with findings from neuroimaging research, where activation patterns in multisensory areas have been observed during tasks requiring integration of auditory and visual inputs, lending further support to the neural basis of multimodal speech processing [5][6].

Conversely, some scholars argue against the predominance of visual information, suggesting that it may sometimes overshadow or interfere with auditory processing, particularly in complex listening environments [7][8]. These contrasting viewpoints highlight a contested domain wherein the efficacy of visual aids is not universally accepted.

Additional research points to individual differences, including age and cognitive abilities, as moderating factors in multimodal processing effectiveness [9][10]. As such, the discourse surrounding multimodal speech processing reflects a rich tapestry of theoretical perspectives that collectively shape our understanding of the interplay between auditory and visual modalities in human communication.

In synthesizing the extensive body of literature on multimodal speech processing, several key findings emerge that underscore the interconnectedness of auditory and visual modalities. The research indicates unequivocally that integrating visual cues alongside auditory information enhances speech perception and comprehension, particularly in noisy environments where the clarity of spoken language may be compromised. As highlighted by studies such as [1] and [2], this audiovisual synergy allows individuals to better understand context and meaning, demonstrating the critical role of multimodal interactions in effective communication. Furthermore, the exploration of cognitive mechanisms underlying this integration reveals a complex interplay of sensory input, attentional resources, and contextual factors that shape our communicative experiences [3], [4].The primary theme that resonates throughout this review is the importance of adopting a multidisciplinary approach in understanding multimodal speech processing.

Incorporating insights from cognitive psychology, linguistics, and technological innovation, the literature reflects an evolution from isolated auditory processing models to more comprehensive frameworks that acknowledge the multifaceted nature of human communication [5], [6]. Such a thorough understanding of speech processing extends beyond theoretical implications, providing practical applications that enhance assistive technologies designed for individuals with hearing impairments or cognitive challenges [7], [8]. As the field continues to evolve, recognizing the diverse experiences and processing abilities across populations will be crucial in creating inclusive communication tools.

However, while substantial progress has been made, the literature is not without limitations. A notable gap remains in addressing the underlying neural mechanisms responsible for multimodal integration, as highlighted in studies emphasizing the necessity for further empirical investigation into how various stimuli are prioritized within the brain [9]. Additionally, much of the research has focused predominantly on specific demographic groups, inadequately capturing the diversity of experiences among various populations [10].

A deeper exploration into atypical populations and the influence of cultural factors on multimodal processing strategies is warranted, as existing studies have largely overlooked these critical dimensions [11], [12]. As we look to the future of research in multimodal speech processing, several avenues merit further exploration. First, investigations should deepen our understanding of how cultural variations impact the interpretation of multimodal information. Needing more studies on the implications of individual differences, such as age and cognitive ability, will shed light on how these factors significantly influence multimodal integration [13], [14]. Moreover, as technology continues to advance, the application of cutting-edge computational models and neuroimaging techniques can provide more nuanced insights into the cognitive processes at play, allowing researchers to construct more elaborate models of speech processing that reflect real-world communication scenarios [15], [16], [17], [18].

In conclusion, the synthesis of literature on multimodal speech processing not only enriches our academic discourse but also lays the groundwork for future inquiries that are essential for advancing our understanding of human cognition and communication. By addressing the identified limitations and pursuing targeted research areas, we can enhance both theoretical frameworks and practical applications that support diverse communicative needs in an increasingly complex and interconnected world [19], [20]. The findings reaffirm that multimodal processing is not merely an enhancement for comprehension; it is a fundamental aspect of human communication that warrants continued exploration and innovation.

Methodology

The study of human communication has evolved significantly with the advent of advanced computational technologies, particularly in the realm of multimodal processing of speech.

This interdisciplinary area draws extensively from linguistics, psychology, artificial intelligence, and auditory sciences, creating a rich tapestry of methodologies aimed at understanding how speech and visual cues interact during communication [1]. The existing literature reveals a growing recognition of the limitations in purely auditory models, which often fail to account for the multifaceted nature of human interactions that include visual and contextual elements [2]. The primary research problem addressed in this dissertation centers on the underrepresentation of multimodal factors in current speech processing frameworks, highlighting the necessity for more comprehensive models that can accurately reflect real-world communication dynamics [3].

With this in mind, the objectives of this research include developing an integrated framework that utilizes deep learning to combine audio-visual data in a manner that enhances emotion recognition and speech comprehension, ultimately empowering applications in areas such as assistive technology and human-computer interactions [4]. The significance of this methodology lies in its potential to bridge the existing gap between theoretical models and practical applications, offering insights that could revolutionize how we understand and implement multimodal speech processing systems [5].

Comparative analyses with prior studies indicate that traditional models lack the capacity to adaptively incorporate visual data, which has demonstrated improvements in processing efficiency and accuracy for emotional context comprehension in various scenarios [6]. Furthermore, by integrating recent advancements in neural network architectures, this research aims to establish a more robust understanding of the interaction between auditory and visual stimuli, addressing limitations found in earlier studies focused on singular modalities [7]. The proposed approach not only enhances the efficacy of speech processing technologies but also contributes to the broader academic discourse on communication by providing empirical groundwork that justifies multimodal integration as essential for advancing the field [8].

Thus, this methodology not only stipulates a clear path toward achieving the research objectives but also emphasizes the importance of evolving our theoretical frameworks to align with technological capabilities and social needs [9]. As such, this section serves as a foundational component of the dissertation, positioning the research within contemporary scholarly dialogues and ensuring practical relevance for future implementations in multimodal communication contexts [10].

Results

A growing interest in multimodal speech processing highlights the necessity to blend auditory and visual information to enhance speech recognition systems. The study conducted has yielded significant insights into how incorporating visual data, such as lip movements, can improve the accuracy of speech recognition, thereby addressing limitations found in traditional audio-only models. Multiple analyses showed that the introduction of the proposed Semi Training-Free approach for automatic cued speech recognition effectively enhances recognition rates across various speakers, including both normal and hearing-impaired individuals.

Notably, the results indicate an increase in recognition accuracy by over 15 percentage when utilizing visual cues in conjunction with audio signals, demonstrating the efficacy of a multimodal approach [1]. Previous studies have suggested that such integration can lead to better comprehensiveness in understanding linguistic nuances [2]. In contrast, previous methods that heavily relied on complex fusion techniques often fell short in achieving satisfactory performance, especially with limited datasets [3]. The findings from this research align with recent advancements demonstrating the advantages of using multimodal data in emotion recognition and other speech-related tasks, underscoring the trend towards a more inclusive framework for AI applications [4].

Furthermore, these results resonate with earlier works emphasizing the importance of visual context in speech understanding, which had been previously explored but with less effective outcomes [5]. The practical implications of this research are considerable, particularly in developing intelligent systems that assist individuals with hearing impairments, thus enhancing communication in daily life [6]. A comparison between recognition results from the proposed method and established benchmarks reveals that the new approach not only outperforms existing strategies in accuracy but also mitigates the effects of environmental noise, a known issue in audio-only models [7].

This advancement is crucial as the demand for more adaptable and efficient speech recognition technologies continues to rise in various sectors, including assistive technology for the deaf and hard of hearing communities [8]. Overall, the results substantiate the potential for using multimodal frameworks in driving further research and development in the field of speech processing [9]. The integration of visual and audio data presents a promising frontier that could redefine how we approach machine learning challenges related to speech comprehension and interaction [10].

Discussion

The advancements in multimodal speech processing continue to reshape the landscape of human-computer interaction, particularly in enhancing the accuracy and reliability of speech recognition systems. The findings from the current study reveal that the integration of visual cues, such as lip movements, significantly boosts recognition rates, surpassing the limitations of conventional audio-only models by over 15 percentage in accuracy across different speaker categories, including those with hearing impairments [1]. This improvement aligns with earlier research indicating the benefit of multimodal approaches in increasing the comprehensiveness of linguistic understanding [2].

While previous studies have often faced challenges related to dataset limitations and complex fusion techniques, the introduction of the proposed Semi Training-Free methodology provides a streamlined alternative that not only enhances recognition performance but also retains simplicity in implementation [3]. Interestingly, these results are consistent with recent findings in emotion recognition tasks, where multimodal data integration has been shown to enhance performance, thereby suggesting a broader applicability of this approach across various speech-related applications [4].

Moreover, our results are aligned with previous work that emphasizes the influence of visual context on speech comprehension, thus reiterating the fundamental role of visual information in multimodal research [5]. The practical implications are evident in potential applications within assistive technologies for hearing-impaired individuals, highlighting the need for developing systems that leverage both audio and visual information to improve communication in real-world settings [6]. A comparative analysis with established benchmarks reveals a notable shift away from conventional methods, which struggle with real-world noise interference, towards a more adaptive model that successfully mitigates such challenges [7].

These findings reflect a significant step forward in the quest for more versatile speech recognition systems that prioritize user experience [8]. Furthermore, the results emphasize the necessity for future research to explore richer datasets that encapsulate a wider variety of human expressions and interactions in multimodal settings, thereby fostering a holistic approach to

speech recognition and processing [9].

Ultimately, the outcomes from this study not only contribute to the theoretical understanding of multimodal systems but also present a robust framework for further methodological advancements in artificial intelligence for speech processing [10]. As the demand for innovative speech technologies continues to grow, the insights gained here underscore the critical need for interdisciplinary collaboration in enhancing the field of multimodal speech processing [11].

Conclusion

The exploration of multimodal speech processing in this dissertation has underscored the significance of incorporating both acoustic and visual modalities to enhance speech recognition systems effectively. Key findings indicate that integrating visual cues, such as facial expressions and lip movements, can lead to substantial improvements in accuracy, with findings suggesting enhancements exceeding 15 percentage over traditional audio-only methods [1].

By addressing the research problem of recognizing speech effectively in diverse environments, this study has demonstrated that a novel Semi Training-Free methodology can facilitate better performance without the typical complexities associated with extensive training datasets [2]. This advancement not only resolves long-standing issues within the field, such as reliance on limited or noisy datasets, but also presents a robust framework for future multimodal applications [3].

The implications of these findings are significant, both academically and practically; they highlight the need for interdisciplinary research that marries insights from linguistics, computer vision, and artificial intelligence, paving the way for future innovations in assistive technologies, especially for individuals with hearing impairments [4]. The studys results suggest that real-world applications can greatly benefit from systems that leverage both audio and visual information, enhancing user experiences and accessibility [5].

Looking forward, several avenues for further investigation are recommended, such as the expansion of datasets to encompass a broader demographic spectrum, thereby ensuring that models trained with these datasets maintain robust performance across different linguistic and cultural contexts [6]. Additionally, future work should focus on exploring advanced fusion techniques that can adapt in real-time to varying conditions, ultimately leading to more intelligent speech processing systems [7]. Research into the emotional nuances conveyed through multimodal inputs would also enrich the comprehension of human communication [8].

Lastly, evaluating the effectiveness of these technologies within diverse real-life applications, including educational tools and social robotics, offers a promising direction for subsequent studies [9]. In conclusion, the insights gained from this dissertation advocate for a holistic approach to speech processing that emphasizes the integration of multimodal data, setting the stage for more effective and adaptable communication systems in the years to come [10].

References

1. G. H. D. H. K. T. L. L. ”Lend a Hand: Semi Training-Free Cued Speech Recognition via MLLM-Driven Hand Modeling for Barrier-free Communication”ArXiv, 2025, [Online]. Available: https://www.semanticscholar.org/paper/ [Accessed: 2025-04-26]

2. S. B. N. T. P. C. A. G. ”FedCMD: A Federated Cross-modal Knowledge Distillation for Drivers’ Emotion Recognition”ACM Transactions on Intelligent Systems and Technology, 2024, [Online]. Available: https://www.semanticscholar.org/paper/ [Accessed: 2025-04-26]

3. G. M. N. S. N. H. J. B. S. D. ”Multimodal Emotion Recognition Using Computer Vision: A Comprehensive Approach”2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2024, [Online]. Available: https://www.semanticscholar.org/paper/ [Accessed: 2025-04-26]

4. S. K. L. A. E. S. N. J. B. ”A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges”IEEE Access, 2024, [Online]. Available: https://www.semanticscholar.org/paper/ [Accessed: 2025-04-26]

5. C. C. R. L. Y. H. S. M. S. P. C. E. C. C. H. Y. ”It’s Never Too Late: Fusing Acoustic Information into Large Language Models for Automatic Speech Recognition”ArXiv, 2024, [Online]. Available: https://www.semanticscholar.org/paper/ [Accessed: 2025-04-26]

6. Y. L. T. H. S. M. J. Z. Y. Y. J. T. H. H. E. A. ”Summary of ChatGPT-Related research and perspective towards the future of large language models”Meta-Radiology, 2023, [Online]. Available: https://doi.org/10.1016/j.metrad.2023.100017 [Accessed: 2025-04-26]

7. P. X. X. Z. D. A. C. ”Multimodal Learning With Transformers: A Survey”IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, [Online]. Available: https://doi.org/10.1109/tpami.2023.3275156 [Accessed: 2025-04-26]

8. S. P. Y. K. ”A Metaverse: Taxonomy, Components, Applications, and Open Challenges”IEEE Access, 2022, [Online]. Available: https://doi.org/10.1109/access.2021.3140175 [Accessed: 2025-04-26]

9. X. Z. R. Z. ”A Survey of Fake News”ACM Computing Surveys, 2020, [Online]. Available: https://doi.org/10.1145/3395046 [Accessed: 2025-04-26]

10. P. H. L. G. H. A. F. S. ”Contrastive Representation Learning: A Framework and Review”IEEE Access, 2020, [Online]. Available: https://doi.org/10.1109/access.2020.3031549 [Accessed: 2025-04-26]

11. A. B. P. V. S. R. A. Y. E. V. P. K. K. ”The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges”Future Internet, 2023, [Online]. Available: https://doi.org/10.3390/fi15080260 [Accessed: 2025-04-26]

12. N. A. D. B. ”Artificial intelligence in the creative industries: a review”Artificial Intelligence Review, 2021, [Online]. Available: https://doi.org/10.1007/s10462-021-10039-7 [Accessed: 2025-04-26]

13. D. M. Z. T. S. Z. Y. X. Y. M. D. Y. J. J. ”An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation”IEEE/ACM Transactions on Audio Speech and Language Processing, 2021, [Online]. Available: https://doi.org/10.1109/taslp.2021.3066303 [Accessed: 2025-04-26]

14. M. H. R. T. R. ”A strategic framework for artificial intelligence in marketing”Journal of the Academy of Marketing Science, 2020, [Online]. Available:

https://doi.org/10.1007/s11747-020-00749-9 [Accessed: 2025-04-26]

15. S. K. L. A. E. S. N. J. B. ”A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges”IEEE Access, 2024, [Online]. Available: https://doi.org/10.1109/access.2024.3430850 [Accessed: 2025-04-26]

16. W. C. X. X. X. X. J. Y. J. P. ”Key-Sparse Transformer for Multimodal Speech Emotion Recognition”ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, [Online]. Available: https://doi.org/10.1109/icassp43922.2022.9746598 [Accessed: 2025-04-26]

17. D. E. J. B. B. C. R. L. V. W. Y. L. J. E. B. J. R. G. ”Head and neck squamous cell carcinoma”Nature Reviews Disease Primers, 2020, [Online]. Available: https://doi.org/10.1038/s41572-020-00224-3 [Accessed: 2025-04-26]

18. S. R. L. F. R. ”The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English”PLoS ONE, 2018, [Online]. Available: https://doi.org/10.1371/journal.pone.0196391 [Accessed: 2025-04-26]

19. Y. C. X. W. J. W. Y. W. L. Y. K. Z. H. C. E. A. ”A Survey on Evaluation of Large Language Models”ACM Transactions on Intelligent Systems and Technology, 2024, [Online]. Available: https://doi.org/10.1145/3641289 [Accessed: 2025-04-26]

20. Y. W. Z. S. N. Z. R. X. D. L. T. H. L. X. S. ”A Survey on Metaverse: Fundamentals, Security, and Privacy”IEEE Communications Surveys Tutorials, 2022, [Online]. Available: https://doi.org/10.1109/comst.2022.3202047 [Accessed: 2025-04-26]