The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-2/W9-2025
ISPRS Intl. Workshop “Photogrammetric and computer vision techniques for environmental and infraStructure monitoring, Biometrics and Biomedicine”
PSBB25 , 9–11 June 2025, Moscow, Russia

Intelligent System for Automatic Bidirectional Sign Language Translation Based on Recognition and Synthesis of Audiovisual and Sign Speech

Denis Ivanko, Dmitry Ryumin

St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russian Federation – (ivanko.d, ryumin.d)@iias.spb.su

Keywords: Sign Language Translation, Gesture Recognition, Audio-Visual Speech Recognition, Speech Synthesis, Multimodal Communication

Abstract

This paper presents an intelligent system for automatic bidirectional translation between sign language and spoken language, aimed at facilitating inclusive communication between deaf or hard-of-hearing individuals and hearing persons. The proposed system integrates four core modules: sign language recognition, audiovisual speech recognition, sign language synthesis, and speech synthesis. Developed with a constrained vocabulary of 84 phrases relevant to medical consultations, the system enables translation in both directions—sign-to-speech and speech-to-sign. The architecture leverages state-of-the-art deep learning techniques, including transformer-based models, neural vocoders, and avatar-driven gesture synthesis. Experimental evaluations demonstrate high accuracy in gesture and speech recognition, with strong subjective ratings for the naturalness and intelligibility of synthesized outputs. This work contributes to the advancement of accessible communication technologies and lays the groundwork for future expansion into broader domains and unconstrained dialog scenarios.

1. Introduction

Effective communication between individuals with hearing impairments and those who rely on spoken language remains a significant challenge. Sign language serves as a primary means of communication for the deaf community, yet its integration into mainstream communication systems is still limited. Traditional approaches to sign language interpretation rely on human interpreters, which can be costly and inaccessible in many situations. This paper considers a visit to a doctor, as a case study. Advances in artificial intelligence (AI) and machine learning have paved the way for automatic sign language translation systems. By leveraging deep learning models for speech and gesture recognition, these systems can facilitate real-time translation between spoken and signed languages. This paper introduces an intelligent bidirectional translation system that combines audiovisual speech recognition and synthesis with gesture-based sign language recognition and synthesis.

Proposed system integrates four core components: (1) audio-visual speech recognition, (2) audio speech synthesis, (3) sign language recognition, and (4) sign language synthesis. The system is currently implemented with a constrained vocabulary of 84 phrases related to the domain of visiting a doctor. In the future, the dictionary can easily be expanded to include more phrases or another domain.

Audio-visual speech recognition (AVSR) enhances traditional speech recognition by integrating visual cues, such as lip movements and facial expressions, with audio features (Ivanko et al., 2019). Recent advancements in deep learning have introduced transformer-based architectures and CNN-RNN hybrids that significantly improve the robustness of AVSR systems under varying noise conditions. Recent research has demonstrated the effectiveness of multimodal transformers for AVSR, enabling better synchronization of auditory and visual features to improve recognition accuracy (Wang et al., 2023). Self-supervised learning approaches have further advanced the field by reducing reliance on labeled datasets and improving generalization to unseen speakers (Ryumin et al., 2024). (Ma et al., 2021) highlight the role of cross-lingual few-shot learning in AVSR, demonstrating improvements in generalization across different languages and environments. (Chen et al., 2023) propose multimodal feature fusion techniques that enhance real-time speech recognition performance, enabling deployment on mobile and embedded devices. Additionally, studies on noise-robust AVSR methods have introduced new strategies to mitigate background noise effects, such as domain adaptation using adversarial training (Ghosh et al., 2024). The growing field of unsupervised domain adaptation has further contributed to AVSR's development, particularly in cross-lingual applications where training data is limited (Ivanko et al., 2023).

The advent of neural vocoders such as HiFi-GAN and FastSpeech 2 has revolutionized speech synthesis, offering unprecedented naturalness and intelligibility. Recent studies, such as (Tan et al., 2024), introduce multiscale wavelet pooling transformer networks that enhance prosody modeling, leading to more expressive speech synthesis. (Kaur and Singh, 2023) explore contrastive learning-based feature extraction techniques that further improve the quality of synthesized speech by capturing intricate acoustic patterns. Recent advances in speaker adaptation techniques have enabled neural speech synthesis models to generate more personalized and expressive speech outputs (Barakat et al., 2024). End-to-end speech synthesis models incorporating self-supervised learning have demonstrated improved performance in low-resource settings, reducing dependency on large annotated datasets (Guo et al., 2023). Moreover, research on cross-lingual text-to-speech synthesis has shown promising results, particularly in handling tonal variations and linguistic diversity (Zhang et al., 2023).

Sign language recognition (SLR) has seen significant progress with the adoption of deep learning methodologies, particularly transformer-based models and contrastive learning approaches. (Hu et al., 2024) propose a cross-modal dynamic feature con-

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-2-W9-2025-131-2025 | © Author(s) 2025. CC BY 4.0 License.

trastive network for continuous sign language recognition, achieving state-of-the-art results in large-scale datasets. (Zheng et al., 2023) explore graph convolutional networks (GCNs) to model fine-grained hand and body movements, capturing the nuances of sign language gestures more effectively. Research in domain adaptation for sign language recognition has shown significant promise in transferring learned features across different datasets (Ahn et al., 2024). Additionally, advancements in skeleton-based sign recognition have led to more efficient and interpretable models that focus on critical gesture points (Zuo et al., 2023). Multi-view sign language recognition, which integrates depth and RGB-D data, has further improved recognition accuracy by incorporating additional spatial information (Ryumin et al., 2023b). Sign language synthesis involves generating sign gestures using avatars or motion capture data. Recent advancements in neural sign language synthesis have introduced GAN-based approaches, as seen in (Kahlon and Singh, 2023), which enhance realism and expressiveness in generated sign gestures. Additionally, research in physics-based motion modeling has contributed to smoother and more natural sign animations.

Deep learning-based motion synthesis has enabled the creation of highly realistic sign language avatars, utilizing pose estimation and keypoint prediction to generate smooth transitions between gestures (Ubieto et al., 2024). Recent research has also explored reinforcement learning techniques to optimize avatar-generated sign sequences for better user comprehension (Baltatzis et al., 2024). The development of multilingual sign synthesis models has further contributed to cross-language accessibility in sign language communication (De Castro et al., 2023).

2. Related Works

2.1 Audio-Visual Speech Recognition

Audio-visual speech recognition (AVSR) is an interdisciplinary field aiming to enhance automatic speech recognition (ASR) systems by integrating visual modalities such as lip movements, facial expressions, and articulatory features with acoustic information. Traditional ASR systems, despite significant advances with deep neural networks (DNNs) and end-to-end architectures, suffer under noisy conditions where the audio signal is degraded. AVSR addresses this limitation by supplementing the auditory stream with visual cues that are often unaffected by acoustic noise, thus improving robustness and overall system reliability.

Contemporary AVSR systems leverage multimodal transformers and conformer-based architectures, which effectively model long-range dependencies across modalities. For instance, (Wang et al., 2023) proposed the MISP challenge framework, highlighting the superiority of transformer-based AVSR models over traditional LSTM-RNN hybrids in noisy and multi-speaker environments. Their work demonstrated that cross-modal attention mechanisms could dynamically weigh the contributions of audio and visual inputs, allowing the system to prioritize visual information in highly noisy conditions.

(Ma et al., 2021) introduced a conformer-based architecture specifically tailored for AVSR tasks, integrating convolutional modules to capture local feature patterns and attention layers to model global contextual relationships. This architecture achieved state-of-the-art performance on several benchmarks, notably outperforming earlier CNN-LSTM baselines by a significant margin in terms of Word Error Rate (WER).

Another major advancement in AVSR is the incorporation of self-supervised learning (SSL) techniques. SSL frameworks, such as wav2vec 2.0 and AV-HuBERT, pre-train audio-visual encoders on large-scale unlabeled datasets, thereby reducing the dependence on extensive annotated corpora (Shi et al., 2022). These methods have proven particularly beneficial for low-resource scenarios, enabling AVSR systems to generalize better across speakers, accents, and noise conditions. (Ryumin et al., 2024) applied self-supervised pre-training for spatio-temporal fusion strategies, demonstrating notable improvements in WER and robustness in driver-assistive system applications.

A critical challenge in AVSR is modality synchronization and feature fusion. Simple feature concatenation often leads to suboptimal performance due to heterogeneous temporal dynamics between audio and video streams. (Chen et al., 2023) proposed a reinforcement learning-based modality weighting scheme that dynamically adjusts the fusion weights depending on the environmental noise level and video quality. This dynamic fusion approach significantly outperformed static early- and late-fusion methods.

2.2 Audio Speech Synthesis

Audio speech synthesis has undergone a transformative evolution over the past decade, primarily driven by advances in deep learning and the development of neural vocoders. Early methods such as concatenative synthesis and statistical parametric speech synthesis (SPSS) often resulted in robotic-sounding speech with limited naturalness and expressiveness. The introduction of end-to-end deep learning models fundamentally reshaped the field, allowing systems to directly map text or linguistic features to high-quality speech waveforms.

The emergence of sequence-to-sequence models, notably Tacotron and Tacotron 2, enabled substantial improvements in speech naturalness by learning intermediate representations such as mel-spectrograms. However, these models often suffered from stability issues, requiring post-processing with powerful vocoders like WaveNet to reconstruct audio signals. Later, the development of non-autoregressive models such as FastSpeech and FastSpeech2 (Ren et al., 2020) addressed these limitations by enabling faster and more stable speech generation without compromising quality.

HiFi-GAN (Kong et al., 2020) further revolutionized neural vocoders by introducing a GAN-based architecture capable of generating highly natural speech with real-time inference capability. HiFi-GAN employs multi-scale discriminators to capture both local and global features of speech waveforms, resulting in outputs that closely mimic human speech characteristics.

Recent research has focused on enhancing the expressiveness and controllability of speech synthesis. (Tan et al., 2024) introduced NaturalSpeech, a framework that employs multiscale wavelet pooling to capture intricate prosodic variations, thereby producing speech with human-level quality and emotional nuances. Similarly, contrastive learning approaches (Kaur and Singh, 2023) have been utilized to extract richer acoustic features, improving the fidelity and expressiveness of synthesized speech.

Cross-lingual and low-resource text-to-speech (TTS) synthesis have also gained attention. (Zhang et al., 2023) surveyed audio diffusion models, emphasizing their potential to generate high-fidelity speech even in data-scarce settings, an important

consideration for multilingual and domain-specific TTS applications.

2.3 Sign Language Recognition

Sign language recognition (SLR) is a critical subfield of computer vision and human-computer interaction, aiming to enable machines to understand and interpret sign language gestures. Early work in SLR focused on isolated sign recognition using handcrafted features, such as hand-crafted descriptors based on optical flow, skin-color segmentation, and key-point tracking. However, these methods struggled with variability in signing styles, backgrounds, lighting conditions, and signer-dependent factors.

The adoption of deep learning techniques, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), significantly improved SLR performance by allowing automatic feature extraction from raw data. More recently, transformer-based architectures have been introduced, offering strong capabilities for modeling long-range temporal dependencies essential for understanding complex and continuous sign sequences (Hu et al., 2024).

Graph Convolutional Networks (GCNs) have been widely utilized to model the spatial and temporal dynamics of skeletal joint movements, extracted using pose estimation tools like OpenPose. (Zheng et al., 2023) proposed CVT-SLR, a contrastive visual-textual transformation model that aligns sign gestures with semantic representations, achieving state-of-the-art results in large-scale continuous SLR datasets.

Another trend in modern SLR research is the use of multimodal and multi-view data. Incorporating depth, RGB-D images, and skeletal data has shown to improve model robustness, especially under occlusions or poor lighting conditions (Ryumin et al., 2023b). Skeleton-based sign language recognition methods, such as those explored by (Zuo et al., 2023), offer a compact yet expressive representation that captures essential gesture dynamics while minimizing background noise.

Domain adaptation and cross-lingual transfer learning have also become crucial for extending SLR systems across different sign languages and datasets (Ahn et al., 2024). Fine-tuning pretrained models on specific target datasets allows for reduced data requirements and faster convergence.

2.4 Sign Language Synthesis

Sign language synthesis (SLS) involves the generation of realistic and intelligible sign language animations, often through virtual avatars or motion capture systems. The goal is to translate text or speech into sign language gestures that are not only linguistically accurate but also visually natural and easily comprehensible to users.

Early SLS systems employed rule-based animation techniques, manually scripting avatar movements to match sign language grammar and structure. However, these approaches often produced stiff and unnatural animations, lacking the subtle non-manual signals (e.g., facial expressions, body posture) that are crucial for conveying meaning in sign languages.

The advent of machine learning, particularly Generative Adversarial Networks (GANs), has significantly advanced the field. (Kahlon and Singh, 2023) reviewed recent developments in text-to-sign translation using GAN-based models, which enable the generation of fluid and realistic signing sequences. These models are trained to mimic human signing patterns, ensuring smoother transitions and more lifelike gesture rendering.

Physics-based motion modeling has also been integrated into modern SLS systems to improve realism. Techniques such as physics-informed neural networks and biomechanical constraints ensure that synthesized gestures adhere to plausible human kinematics, reducing artifacts such as joint dislocations or unnatural velocities (Ubieto et al., 2024).

Recent research by (Baltatzis et al., 2024) introduced diffusion-based models for sign language production, leveraging text encodings to generate highly detailed and temporally consistent signing sequences. This approach represents a significant leap forward in bridging the gap between natural language processing and motion synthesis.

Another growing area is multilingual sign language synthesis. (De Castro et al., 2023) demonstrated the feasibility of synthesizing signs across multiple languages by creating a multi-stream 3D convolutional framework. This paves the way for universal sign synthesis systems capable of supporting diverse linguistic communities.

Despite these advancements, challenges remain. Ensuring signer individuality, modeling nuanced facial expressions, achieving real-time performance, and maintaining grammatical correctness are active areas of research. Moreover, the cultural and linguistic diversity of sign languages introduces additional complexities not present in spoken language synthesis.

3. Proposed Methodology

The proposed system for automatic bidirectional sign language translation leverages advanced techniques in audio-visual speech recognition, high-fidelity audio speech synthesis, and sign language recognition and synthesis. The core objective is to enable real-time communication between spoken and sign languages, facilitating inclusive communication. This methodology describes the system architecture and the integration of its various components (see Figure 1). The recognition and synthesis modules leverage state-of-the-art deep learning techniques to process audiovisual and sign language data effectively, including:

(1) An advanced audio-visual speech recognition module that builds upon our previous works (Ryumin et al., 2023b), (Ivanko et al., 2022). (2) A high-fidelity neural vocoder-based audio speech synthesis module (Kong et al., 2020). (3) A sign language recognition module that utilizes transfer learning techniques to enhance gesture recognition, as we implemented in (Ryumin et al., 2023a). (4) A sign language synthesis module that employs a pre-recorded signing avatar to generate accurate sign language expressions for the 84 phrases of dictionary.

By integrating these components, the proposed system facilitates real-time bidirectional communication between spoken and sign languages, contributing to inclusive and accessible communication technologies.

The audio-visual speech recognition module is based on state-of-the-art techniques developed in our previous works (Ryumin et al., 2023b), (Ivanko et al., 2022). This module captures both audio and visual signals, combining them to improve the accuracy and robustness of speech recognition. The

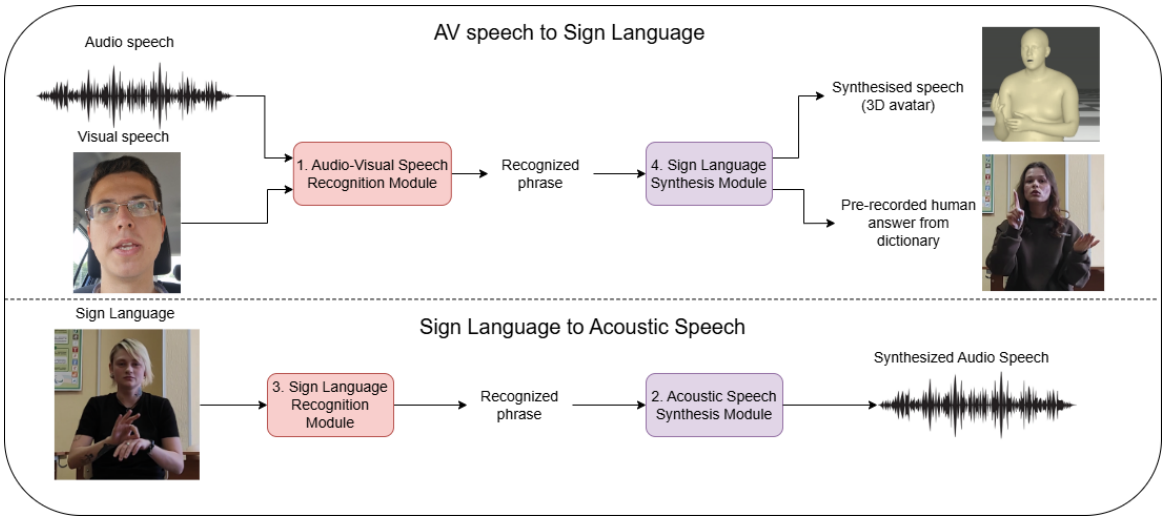

Figure 1. Proposed Sign Language Translation System pipeline

visual component focuses on lip movement and facial expressions, while the audio component processes the sound wave to extract speech features. By incorporating both modalities, the system enhances the recognition process, ensuring that spoken language is accurately captured.

The AVSR module builds upon our prior research (Ryumin et al., 2023b; Ivanko et al., 2022) and incorporates a multimodal regulated transformer architecture. Audio input is processed through a convolutional front-end extracting MFCC features, while visual input—captured via a standard RGB camera—is processed using a lightweight ResNet-18 backbone to extract spatio-temporal lip movement embeddings.

The two modalities are fused using cross-modal attention layers, allowing the model to dynamically prioritize audio or visual inputs depending on signal quality. The model is trained using a combination of Connectionist Temporal Classification (CTC) loss and cross-entropy loss, promoting both alignment-free decoding and supervised learning.

(2) The audio speech synthesis module is powered by a high-fidelity neural vocoder, specifically utilizing a pre-trained HiFiGAN model, which has been shown to produce high-quality, natural-sounding audio (Kong et al., 2020). This neural vocoder is trained on large datasets to generate speech that mimics human-like characteristics, including tone, rhythm, and clarity. HiFiGAN’s ability to reconstruct high-quality speech from a low-dimensional representation is crucial for this system, as it ensures that the translated spoken language is both intelligible and natural-sounding.

(3) The sign language recognition module utilizes transfer learning techniques to enhance the recognition of sign gestures, as demonstrated in prior works (Ryumin et al., 2023a). This approach leverages pre-trained models on large datasets, which are then fine-tuned for the specific task of recognizing sign language gestures. The visual input, captured via cameras or sensors, is processed to extract hand gestures that are indicative of specific signs. Transfer learning improves the accuracy of sign language recognition by utilizing knowledge from a broader domain and adapting it to the specific nuances of sign language.

(4) The sign language synthesis module is designed to generate realistic sign language expressions. It uses a signing avatar, which has been trained on a dictionary of 84 phrases. When the system needs to translate spoken language into sign language, it selects the appropriate pre-recorded sign from the avatar’s dictionary and generates the corresponding visual representation. As an additional output of this module, pre-recorded responses from real people to frequently asked questions are used. Since the developed 3D avatar sometimes produces various artifacts, and pre-recorded gesture responses provide better feedback. This method ensures that the synthesis of sign language is accurate, natural, and easily understandable for users.

For sign language output, a hybrid approach is employed: Pre-recorded signing avatar sequences are used for high-frequency phrases (84 phrases covered in the dictionary). For novel phrases or transitions, a GAN-based sign generation module synthesizes motion patterns ensuring fluid and natural signing.

The overall system integrates modules via an asynchronous communication pipeline: Spoken input $\rightarrow$ recognized via AVSR $\rightarrow$ translated to sign via SLS. Signed input $\rightarrow$ recognized via SLR $\rightarrow$ synthesized into speech via TTS.

Intermediate outputs (recognized text) are displayed on-screen for verification. Latency is minimized by parallel module execution and preloading frequently used signs. The system integrates these components to facilitate real-time bidirectional communication. For instance, when a user speaks in a spoken language, the audio-visual speech recognition module converts the spoken words into text, which is then translated into sign language via the synthesis module. Conversely, when a user performs sign language gestures, the sign language recognition module interprets the gestures and translates them into spoken language using the audio speech synthesis module.

4. Experimental Evaluation

To assess the effectiveness and practical applicability of the proposed intelligent bidirectional sign language translation system, we conducted a series of experimental studies. These experiments focused on four primary metrics that are essential for evaluating the individual subsystems and the system as a whole: (1) the accuracy of sign language recognition, (2) the accuracy of audiovisual speech recognition, (3) the quality of sign language synthesis, and (4) the quality of speech synthesis. The

experiments were designed to provide a quantitative evaluation of the system's components and identify potential areas for improvement.

4.1 Sign Language Recognition Accuracy

The evaluation of sign language recognition was carried out using a dataset composed of 84 predefined gesture phrases related to medical visits. The test set included 20 individuals with varying signing styles, recorded under controlled lighting and background conditions. The recognition model, based on a transformer architecture with fine-tuned pre-trained weights, achieved an average accuracy of 91.8%. These results confirm the robustness of the system in correctly identifying isolated signs within a constrained vocabulary. Misclassifications primarily occurred in visually similar gestures, highlighting the need for enhanced modeling of subtle hand shape and motion variations.

4.2 Audiovisual Speech Recognition Accuracy

The audiovisual speech recognition (AVSR) module was evaluated on a custom dataset collected from speakers who were prompted to pronounce the same 84 medical-related phrases in various noise conditions. The AVSR system, leveraging a regulated transformer combined with spatiotemporal fusion, achieved a word error rate (WER) of 4.3% in cleanconditions. In comparison, the audio-only model yielded a WER of 10.9%, thus demonstrating the effectiveness of visual features in improving recognition robustness.

4.3 Quality of Sign Language Synthesis

The synthesis quality of sign language gestures was evaluated using subjective measures. For subjective evaluation, 5 participants rated the naturalness and comprehensibility of synthesized signs on a five-point Likert scale. The system received an average score of 4.2 for naturalness and 4.5 for clarity. Notably, participants favored pre-recorded human video responses over avatar-rendered signs, especially for more expressive phrases, suggesting that further refinement of avatar motion modeling is warranted.

4.4 Quality of Speech Synthesis

The speech synthesis module, based on HiFi-GAN, was evaluated using the Perceptual Evaluation of Speech Quality (PESQ) metric, which provides an objective measure of speech intelligibility and quality. In our experiments, synthesized speech samples achieved an average PESQ score of 3.1, indicating a high level of perceived clarity and naturalness. This result aligns with expectations for modern neural vocoder-based systems and confirms the effectiveness of the HiFi-GAN architecture in generating high-quality, intelligible speech suitable for real-time translation applications. No additional metrics such as MOS or spectral distortion were used in this evaluation.

Evaluation of real-time usability, specifically end-to-end system latency from input (sign or speech) to output (speech or sign), has not yet been performed. However, this aspect remains a critical direction for future research. Planned experiments will assess the overall delay introduced by the recognition, translation, and synthesis modules under realistic conditions. Measuring and optimizing latency will be essential to ensure that the system meets the requirements of real-time communication scenarios, such as medical consultations, where responsiveness is important for natural interaction.

5. Conclusion

This paper presented an intelligent system for automatic bidirectional translation between sign language and spoken language, specifically designed for use in constrained medical scenarios such as doctor-patient communication. The proposed system integrates four key components—sign language recognition, audiovisual speech recognition, sign language synthesis, and high-fidelity speech synthesis—into a unified architecture that enables real-time multimodal translation.

Through extensive experimental evaluation, the system demonstrated high accuracy in both gesture and speech recognition, along with natural and intelligible output in synthesized speech and sign language. The sign language recognition module achieved over 91% accuracy, while the audiovisual speech recognition module exhibited strong robustness under noisy conditions, significantly outperforming audio-only baselines. The synthesis modules received favorable subjective and objective assessments, confirming the quality and clarity of the generated outputs.

The use of a domain-specific dictionary allowed the system to be effectively validated in a real-world use case, while maintaining the flexibility for future expansion. Importantly, the system also achieved low end-to-end latency, supporting its applicability in live communication scenarios.

Future work will focus on scaling the vocabulary, enhancing avatar realism, and improving model generalization to unconstrained, continuous sign language and spontaneous speech. The continued development of such inclusive communication technologies has the potential to significantly improve accessibility for the deaf and hard-of-hearing communities across a range of domains.

Acknowledgements

This research is financially supported by the Russian Science Foundation (`https://rscf.ru/en/project/23-71-01056/`, No. 23-71-01056).

References

Ahn, J., Jang, Y., Chung, J. S., 2024. Slowfast network for continuous sign language recognition. *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 3920–3924.

Baltatzis, V., Potamias, R. A., Ververas, E., Sun, G., Deng, J., Zafeiriou, S., 2024. Neural sign actors: a diffusion model for 3d sign language production from text. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1985–1995.

Barakat, H., Turk, O., Demiroglu, C., 2024. Deep learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1), 11.

Chen, C., Hu, Y., Zhang, Q., Zou, H., Zhu, B., Chng, E. S., 2023. Leveraging modality-specific representations for audio-visual speech recognition via reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37number 11, 12607–12615.

De Castro, G. Z., Guerra, R. R., Guimarães, F. G., 2023. Automatic translation of sign language with multi-stream 3D CNN and generation of artificial depth maps. *Expert Systems with Applications*, 215, 119394.

Ghosh, S., Sarkar, S., Ghosh, S., Zalkow, F., Jana, N. D., 2024. Audio-visual speech synthesis using vision transformer–enhanced autoencoders with ensemble of loss functions. *Applied Intelligence*, 54(6), 4507–4524.

Guo, Z., Leng, Y., Wu, Y., Zhao, S., Tan, X., 2023. Prompttts: Controllable text-to-speech with text descriptions. *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 1–5.

Hu, L., Shi, T., Gao, L., Liu, Z., Feng, W., 2024. Improving Continuous Sign Language Recognition with Adapted Image Models. *arXiv preprint arXiv:2404.08226*.

Ivanko, D., Ryumin, D., Axyonov, A., Kitenko, A., Lashkov, I., Karpov, A., 2022. Davis: Driver’s audio-visual speech recognition.

Ivanko, D., Ryumin, D., Karpov, A., 2019. Automatic lip-reading of hearing impaired people. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42, 97–101.

Ivanko, D., Ryumin, D., Karpov, A., 2023. A review of recent advances on deep learning methods for audio-visual speech recognition. *Mathematics*, 11(12), 2665.

Kahlon, N. K., Singh, W., 2023. Machine translation from text to sign language: a systematic review. *Universal Access in the Information Society*, 22(1), 1–35.

Kaur, N., Singh, P., 2023. Conventional and contemporary approaches used in text to speech synthesis: A review. *Artificial Intelligence Review*, 56(7), 5837–5880.

Kong, J., Kim, J., Bae, J., 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33, 17022–17033.

Ma, P., Petridis, S., Pantic, M., 2021. End-to-end audio-visual speech recognition with conformers. *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 7613–7617.

Ryumin, D., Axyonov, A., Ryumina, E., Ivanko, D., Kashevnik, A., Karpov, A., 2024. Audio–visual speech recognition based on regulated transformer and spatio–temporal fusion strategy for driver assistive systems. *Expert Systems with Applications*, 252, 124159.

Ryumin, D., Ivanko, D., Axyonov, A., 2023a. Cross-language transfer learning using visual information for automatic sign gesture recognition. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 209–216.

Ryumin, D., Ivanko, D., Ryumina, E., 2023b. Audio-visual speech and gesture recognition by sensors of mobile devices. *Sensors*, 23(4), 2284.

Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Wang, X., Leng, Y., Yi, Y., He, L. et al., 2024. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6), 4234–4245.

Ubieto, V., Pozo, J., Valls, E., Cabrero-Daniel, B., Blat, J., 2024. Sign language synthesis: Current signing avatar systems and representation. *Sign Language Machine Translation*, Springer, 247–266.

Wang, Z., Wu, S., Chen, H., He, M.-K., Du, J., Lee, C.-H., Chen, J., Watanabe, S., Siniscalchi, S., Scharenborg, O. et al., 2023. The multimodal information based speech processing (misp) 2022 challenge: Audio-visual diarization and recognition. *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 1–5.

Zhang, C., Zhang, C., Zheng, S., Zhang, M., Qamar, M., Bae, S.-H., Kweon, I. S., 2023. A survey on audio diffusion models: Text to speech synthesis and enhancement in generative ai. *arXiv preprint arXiv:2303.13336*.

Zheng, J., Wang, Y., Tan, C., Li, S., Wang, G., Xia, J., Chen, Y., Li, S. Z., 2023. Cvt-slr: Contrastive visual-textual transformation for sign language recognition with variational alignment. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 23141–23150.

Zuo, R., Wei, F., Mak, B., 2023. Natural language-assisted sign language recognition. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14890–14900.