Complex Systems Informatics and Modeling Quarterly (CSIMQ)
eISSN: 2255-9922
Published online by RTU Press, https://csimq-journals.rtu.lv
Article 265, Issue 47, June/July 2026, Pages 105–127
https://doi.org/10.7250/csimq.2026-47.05

Toward Real-Time Medical IVR in Latvian: Evaluation Framework and Comparative Analysis of STT and TTS Models

Sintija Petrovica-Klavina[1*], Ingars Erins[1], Roberts Kirsteins[2], Ints Meijers[2], and Biruta Eliza Aunina-Kirmuska[2]

[1] Institute of Applied Computer Systems, Riga Technical University, 6A Kipsalas Street, Riga, LV-1048, Latvia
[2] Tet SIA, Dzirnavu Street 105, Riga, LV-1011, Latvia

sintija.petrovica-klavina@rtu.lv, ingars.erins@rtu.lv, roberts.kirsteins@tet.lv, ints.meijers@tet.lv, eliza.aunina@tet.lv

Abstract. This article investigates state-of-the-art speech-to-text (STT) and text-to-speech (TTS) technologies for the development of a conversational interactive voice response (IVR) solution in the Latvian medical domain. The study addresses challenges associated with low-resource languages, domain-specific medical terminology, real-time interaction, and sensitive data processing. A requirement-driven evaluation framework is proposed that integrates linguistic and domain accuracy, architectural and deployment suitability, real-time performance, scalability, and regulatory considerations relevant to healthcare environments. Based on evidence reported in scientific literature, model documentation, and publicly available repositories, candidate STT and TTS models with Latvian support or adaptation potential are comparatively assessed against these requirements. Rather than providing a standardized experimental benchmark, the analysis identifies architectures and models whose documented characteristics appear most closely aligned with the operational requirements of real-time Latvian medical IVR deployment. The study provides a structured basis for candidate model selection and defines the metrics and deployment conditions to be addressed in subsequent common-condition empirical evaluation and clinical validation.
Keywords: Speech-to-Text, Text-to-Speech, Interactive Voice Response, Medical Domain, Latvian Language, Low-Resource Language.

1 Introduction

The rapid advancement of artificial intelligence (AI) and speech technologies has significantly increased the adoption of voice-based interfaces across various industries, including healthcare [1]. Interactive voice response systems, powered by speech-to-text (STT), or automatic speech recognition (ASR), and text-to-speech (TTS) technologies, are increasingly used to improve

* Corresponding author

Reference: S. Petrovica-Klavina, I. Erins, R. Kirsteins, I. Meijers, and B. E. Aunina-Kirmuska, "Toward Real-Time Medical IVR in Latvian: Evaluation Framework and Comparative Analysis of STT and TTS Models," Complex Systems Informatics and Modeling Quarterly, CSIMQ, no. 47, pp. 105–127, 2026. Available: https://doi.org/10.7250/csimq.2026-47.05

Additional information. Author ORCID iD: S. Petrovica-Klavina – https://orcid.org/0000-0002-7670-638X, I. Erins – https://orcid.org/0000-0003-0138-2277, and I. Meijers – https://orcid.org/0000-0002-9256-5628. PII S225599222600265X. Received: 20 March 2026. Accepted: 28 July 2026. Available online: 31 July 2026.

accessibility, automate routine communication tasks, and enhance service efficiency [2]. In healthcare, speech-based solutions offer substantial potential for optimizing appointment scheduling, patient prioritization, information retrieval, and administrative support, thereby reducing staff workload and improving service availability [3].

Today, healthcare systems face growing pressures from increasing patient volumes, workforce shortages, and the need for continuous communication channels. Traditional interactive voice response (IVR) systems based on fixed menu navigation may be insufficient for complex medical communication, whereas modern AI-based speech systems enable more natural interaction through free-form speech [4]. However, reliable healthcare deployment requires effective integration of domain-adapted STT and TTS components. Despite substantial progress in general-purpose speech technologies, state-of-the-art models remain primarily optimized for high-resource languages and general domains [5], while medical communication introduces specialized terminology, abbreviations, variable acoustic conditions, and higher accuracy requirements due to safety considerations [2], [4]. Speech recognition or synthesis errors may lead to potentially serious misunderstandings [6], and these challenges are even more complicated for Latvian, a morphologically rich and low-resource language [7]. The availability of annotated Latvian speech corpora and domain-specific datasets remains limited, with comparatively less large-scale model optimization than for widely spoken languages [8]. Consequently, speech systems may demonstrate lower accuracy and robustness and insufficient handling of specialized terminology, particularly in healthcare applications.

Several research efforts have attempted to address these limitations. Early works focused on developing foundational ASR resources (e.g., [9]) and language models for Latvian, primarily targeting general-domain speech recognition. More recent studies have concentrated on domain adaptation and specialized corpora. For instance, Dargis et al. [8] and Auzina et al. [10] reported the development of medical-domain Latvian speech corpora and ASR systems adapted for clinical dictation, demonstrating measurable improvements in word error rates through domain-specific language modeling and semi-supervised learning. Znotins et al. [11] introduced LATE, an open-source toolkit for automatic speech recognition and formatted transcription that integrates Whisper-based models and provides state-of-the-art ASR solutions for Latvian and the first publicly available ASR model for Latgalian (a variation of the Latvian language). However, despite these advances, the majority of Latvian-focused research remains centered on offline transcription or domain-specific dictation workflows rather than real-time interactive dialogue systems. Furthermore, systematic deployment-oriented assessment of both STT and TTS technologies for integrated IVR applications in the Latvian medical domain remains limited. In particular, available evidence is fragmented across studies, datasets, model documentation, and implementation environments, making direct comparison of accuracy, latency, streaming capability, scalability, and integration characteristics difficult. This creates a research gap in establishing a common set of requirements for identifying candidate speech technologies suitable for subsequent empirical evaluation in conversational healthcare services. Therefore, the central research question addressed in this study is how existing STT and TTS technologies can be systematically assessed and shortlisted for real-time IVR deployment in the Latvian medical domain under linguistic, accuracy, real-time performance, deployment, and regulatory constraints.

The primary objective of this study is to review existing STT and TTS technologies, define technical and domain-specific requirements for their integration into a Latvian medical IVR solution, and comparatively assess candidate models against these requirements to identify those most suitable for planned empirical evaluation. To achieve this objective, a structured literature review was conducted to identify relevant research on STT and TTS technologies, with particular emphasis on Latvian-language and medical-domain applications. Scientific sources were retrieved from major academic databases, including IEEE Xplore, ScienceDirect, ACM Digital Library, SpringerLink, Web of Science, ProQuest, EBSCOhost, and Wiley Online Library.

The search strategy employed combinations of domain-specific keywords, including "*speech recognition models*," "*speech recognition Latvian*," "*speech recognition medicine*," "*text-to-*

speech models," "*speech synthesis*," "*STT models*," "*TTS models*," "*speech Latvian*," and "*medical speech technologies*." Boolean operators and database-specific filtering tools were used to refine results. The initial search yielded 136 publications. After abstract and full-text screening based on relevance to STT/TTS architectures, multilingual or low-resource language adaptation, medical-domain applicability, model evaluation methodologies, and system integration aspects, 55 peer-reviewed sources were selected for in-depth analysis. To ensure technological relevance and up-to-date model coverage, supplementary materials were also examined, including open-source model documentation, technical reports, official model cards, and publicly available repositories (e.g., GitHub, Hugging Face). These sources were used to complement academic findings with implementation-level insights, particularly for recently released models not yet extensively documented in peer-reviewed literature.

The contribution of this article is twofold. First, it provides a structured, deployment-oriented overview of current STT and TTS technologies relevant to Latvian as a low-resource language, with particular attention to medical-domain and real-time IVR requirements. Second, it combines linguistic, domain-specific, architectural, operational, integration, and data-protection considerations into a requirement-driven evaluation framework and applies these requirements to comparatively assess candidate models for planned empirical testing. The resulting analysis is intended to support transparent candidate selection rather than establish a definitive performance ranking, since the available models have not yet been evaluated under common datasets, hardware configurations, and measurement protocols. Further validation of the proposed framework involving speech-technology specialists and healthcare professionals is therefore required to assess and refine the relevance and relative importance of the selected evaluation criteria for real-world medical IVR deployment.

The rest of the article is organized as follows. Section 2 discusses the specific challenges of medical-domain speech technologies and additional constraints set by the Latvian language and reviews research in the Latvian healthcare context. Section 3 lists criteria for the STT and TTS models supporting IVR solution development in the Latvian medical domain. Section 4 comparatively assesses STT and TTS models against the defined criteria and identifies candidates for subsequent empirical evaluation. Conclusions and future work are given in Section 5 and Section 6 respectively.

2 Challenges of Medical Speech Technologies in Low-Resource Languages

The implementation of STT and TTS technologies in healthcare environments poses substantial linguistic, technical, and regulatory challenges, which become even more pronounced in low-resource languages. This section examines these challenges from two perspectives. First, it outlines the specifics of the medical domain, where high terminological precision, robustness, and data security are critical. Second, it discusses the additional complexities introduced by Latvian as a morphologically rich and data-scarce language, highlighting existing research initiatives and current limitations. Together, these aspects frame the technical and operational constraints relevant for deploying reliable IVR-oriented medical speech solutions.

2.1 The Medical Domain and the Role of STT/TTS Solutions

The healthcare sector is characterized by complex communication processes, extensive documentation requirements, and strict quality and safety standards. In this context, natural language processing (NLP) technologies have become increasingly important, particularly for the analysis of clinical notes, electronic health records, and radiology reports [1]. By extracting structured information from unstructured textual data, NLP systems support disease diagnosis, patient management, and medical research, thereby facilitating more efficient and data-driven clinical decision-making [2].

One of the major operational challenges in healthcare is the time-consuming nature of clinical documentation. Activities such as note-taking, information retrieval, and manual data entry

substantially reduce the time medical professionals can dedicate to direct patient care [12]. Paper-based or only partially digitized workflows further hinder timely access to critical patient information. Speech technologies represent a promising approach to mitigating these burdens by enabling real-time dictation, automated transcription, and voice-driven interaction within clinical systems [2].

Nevertheless, the deployment of speech technologies in the medical domain presents significant technical and methodological challenges. Medical language is highly specialized and characterized by complex terminology, abbreviations, drug names, institutional jargon, and context-dependent expressions [4]. Ensuring accurate recognition and semantic interpretation of such terminology is critically important. Misrecognition or semantic distortion of medical terms during transcription may directly affect clinical decision-making and potentially endanger patient safety [6].

In addition to linguistic complexity, medical audio recordings frequently contain background noise, overlapping speech, device-generated sounds, and variability in recording quality (e.g., 8 kHz telephone recordings versus 16 kHz clinical dictation). Consequently, STT systems must demonstrate robustness not only at the linguistic level but also in handling acoustic variability [2]. Prior research indicates that the scarcity of domain-specific training data represents a primary source of transcription errors in medical speech recognition systems [12]. The availability of large-scale, structured, and clinically representative datasets is therefore essential to ensure model robustness, generalizability, and reliability in real-world clinical settings [13].

Modern medical speech solutions increasingly incorporate large language models (LLMs) as post-processing components [14]. While STT systems primarily capture phonetic information, LLMs can refine transcriptions by correcting terminology, resolving ambiguities, and structuring clinical information [6]. This layered approach with speech recognition followed by contextual validation and semantic structuring has become common in digital medical assistants and telemedicine platforms. Following validation, TTS systems may serve as an output layer to generate clinically accurate and comprehensible spoken responses, such as treatment explanations or medication instructions [4].

Reliability and security considerations are also particularly critical. Medical audio recordings contain sensitive personal data and must comply with strict data protection regulations [6]. In addition, speech models may demonstrate uneven performance across demographic groups, dialects, and accents, increasing the risk of systematic biases. The potential for hallucinations, misinterpretations, or clinically inaccurate outputs is particularly problematic in healthcare contexts. Such risks necessitate robust human oversight, transparent validation mechanisms, and comprehensive quality assurance protocols to ensure safe and trustworthy system deployment [4].

2.2 Latvian as a Low-Resource Language and Its Use in the Medical Domain

The deployment of speech technologies in the Latvian context introduces additional linguistic and resource-related challenges. Latvian is a morphologically rich language characterized by extensive inflectional variation [15]. Word forms vary according to case, number, gender, and other grammatical categories, resulting in high lexical variability. This morphological complexity increases the demands placed on both acoustic and language modeling components, particularly within specialized or domain-specific applications [10], [12].

In comparison with high-resource languages such as English, Latvian is constrained by a limited availability of publicly accessible annotated corpora, fewer domain-specific medical datasets, and comparatively weaker commercial model support [8], [16]. Contemporary transformer-based speech architectures typically require large-scale labeled datasets and substantial computational resources, which are often difficult to secure in low-resource settings. Consequently, several Latvian automatic speech recognition initiatives (e.g., [12], [17], [18]) have adopted hybrid or n-gram-based language modeling approaches, prioritizing computational efficiency and stable performance under constrained data conditions. These linguistic and resource limitations become particularly critical in specialized domains such as healthcare, where the amount of domain-specific terminology is high and tolerance for recognition errors is minimal [12], [19].

To date, only a limited number of research initiatives in Latvia have specifically addressed speech technologies in the medical context. One significant project, "Latvian Speech Recognition and Synthesis for Medical Applications," was conducted by the Institute of Mathematics and Computer Science (University of Latvia), in collaboration with Riga East University Hospital [16]. The project focused on developing specialized Latvian-language resources, including a medical speech corpus, a pronunciation lexicon covering medical terminology and abbreviations, and adapted ASR/TTS models for radiology and related domains [19]. The researchers developed anonymized corpora of historical medical reports (i.e., radiology examination and epicrisis reports), constructed pronunciation dictionaries for medical terms and drug names, and implemented advanced pre- and post-processing rules using finite-state transducers (OpenFST/Thrax framework) [17], [19]. Acoustic models were built using the Kaldi toolkit with time-delay neural networks, while language models were developed using KenLM and Kaldi-RNNLM tools. Training data included a combination of general Latvian speech corpora (100-hour Latvian Speech Recognition Corpus), dictation corpora, and parliamentary speech data [12], [17].

A key methodological contribution was semi-supervised training, where pseudo-transcriptions generated by a general-domain ASR model were used to adapt acoustic and language models to medical subdomains, including radiology, epicrisis, psychiatry, and pediatrics. This approach significantly reduced dependence on manually transcribed data, which is particularly costly in medical contexts [12]. Reported results demonstrated substantial Word Error Rate (WER) improvements across domains, with relative improvements of up to 39% for epicrisis data and 27–29% for psychiatry data. A WER below 25%, considered sufficient for practical deployment in that study, was achieved in high-quality psychiatric recordings, although the radiology and epicrisis domains required further improvement [12].

Another significant initiative in the Latvian medical domain, "Automated Voice Communication Solutions for the Healthcare Industry," was implemented between 2021 and 2022 in cooperation between the Latvian Children's Clinical University Hospital and Tilde SIA [20]. The project explored NLP and speech recognition methods adapted to the medical domain and focused on developing two prototypes: (1) a post-processing environment for automatically generated speech transcripts, and (2) the integration of speech-recognition solutions into clinical workflows [21]. This collaboration continued within the COMPRISE[†] project, resulting in two conversational AI demonstrators – Hospital Concierge and Doctor's Assistant [22]. Hospital Concierge is a voice-enabled virtual assistant deployed at hospital reception areas, supporting visitors through spoken and text-based interaction by answering frequently asked questions, providing navigation guidance, and assisting users with special needs. Doctor's Assistant is designed to support clinical workflows by enabling medical professionals to record and retrieve patient-related information through voice commands. The system integrates custom speech recognition and NLP components tailored to the medical domain and supports structured data integration into hospital information systems [22].

More recently, commercial solutions specifically targeting Latvian medical speech have also emerged. One example is Viola [23], developed by the Latvian company SoftMedico as an AI-based medical transcription and documentation tool adapted to Latvian and medical terminology. The system converts dictated or recorded speech into structured medical text, supports punctuation and formatting, and is designed to recognize specialized medical terminology [23], [24]. According to publicly available information, the solution has been used by more than 300 medical professionals and has processed more than 60,000 minutes of medical audio [24]. Viola can be accessed as a web-based solution and also offers on-premise deployment for healthcare organizations requiring greater control over data processing [23]. However, its primary use case is medical dictation and documentation rather than bidirectional real-time interaction with a patient through an IVR system. Thus, while Viola demonstrates the practical applicability of Latvian medical speech recognition, real-time conversational IVR additionally requires streaming recognition, low turn-taking latency, concurrent call processing, and TTS integration.

[†] COMPRISE project - https://www.compriseh2020.eu/

2.3 Summary and Research Gap

The analysis of Latvian medical language technologies demonstrates substantial foundational work, including medical speech corpora, pronunciation lexicons, adapted language models, and domain-specific ASR systems. Semi-supervised learning has also been applied to address limited medical speech data, demonstrating that domain adaptation can improve recognition performance. However, existing Latvian solutions primarily target clinical dictation and transcription rather than real-time conversational interaction with patients. Consequently, evidence on integrated STT-TTS pipelines under IVR-specific conditions remains limited, particularly regarding latency, concurrent use, telephone-quality audio, and real-time dialogue management.

Furthermore, prior studies predominantly assess speech recognition performance using conventional metrics such as WER, without systematically addressing operational constraints essential for IVR deployment, including streaming capability, system integration feasibility, and GDPR-compliant data processing. These challenges are particularly notable for Latvian as a low-resource, morphologically rich language, where extensive inflectional variation, specialized medical terminology, and accent diversity increase modeling complexity. Also, on the TTS side, comparatively limited attention has been paid to the pronunciation accuracy of domain-specific terminology and low-latency streaming performance in real-time conversational settings.

Taken together, these findings highlight a clear research gap. Beyond the technological dimension, this gap is particularly significant in the context of the Latvian healthcare system. Latvia represents a relatively small healthcare market characterized by workforce shortages, high administrative burden, and increasing demand for accessible patient communication channels. National healthcare policy documents (e.g., *Health Workforce Development Strategy*, 2025-2029[‡]) emphasize the need to reduce administrative workload and promote digital and AI-based solutions to optimize human resource allocation.

Despite these strategic objectives, empirically validated guidance for selecting, integrating, and deploying real-time Latvian-language STT and TTS technologies within healthcare IVR workflows remains limited. In particular, there is insufficient evidence from common-condition evaluations addressing medical-domain accuracy, real-time performance, scalability, system integration, and secure processing of sensitive medical data. Consequently, healthcare providers currently have limited evidence-based guidance for determining which technological configurations are appropriate for such applications and under which operational conditions they can be reliably deployed. Addressing this gap is therefore relevant not only from a technological perspective but also to the broader digitalization of healthcare services in Latvia and other low-resource language contexts.

3 Evaluation Framework for STT/TTS Models for Latvian Medical IVR

The development of a conversational IVR solution for the Latvian medical domain requires a structured and multidimensional evaluation framework reflecting linguistic complexity, technical feasibility, operational constraints, and regulatory requirements. Unlike general-purpose speech applications, medical IVR systems operate in a safety-critical and privacy-sensitive environment where recognition accuracy, terminological precision, response latency, conversational robustness, and secure data processing directly influence communication quality and system reliability. Model selection must therefore consider not only STT and TTS quality but also system-level requirements, including real-time interaction, streaming, concurrent call processing, infrastructure integration, and GDPR-compliant data processing.

Accordingly, this section presents a requirement-driven framework for assessing candidate speech technologies and their integration into a medical IVR system. Section 3.1 defines criteria for STT models, focusing on linguistic and domain accuracy, architectural and deployment suitability, real-time performance, and compliance-related considerations. Section 3.2 defines corresponding criteria

[‡] Cabinet of Ministers. Health Workforce Development Strategy Plan 2025–2029 - https://likumi.lv/ta/id/357507

for TTS models, emphasizing Latvian-language synthesis, medical-term pronunciation and pronunciation control, speech quality, intelligibility, real-time performance, and deployment characteristics. Section 3.3 extends the assessment to the integrated IVR pipeline, covering end-to-end responsiveness, telephony robustness, concurrency, information capture, and error handling.

The framework is intended to support systematic candidate selection, identify trade-offs among the evaluation criteria, and define the conditions for subsequent common-condition empirical evaluation rather than to provide a validated weighting or ranking model. Numerical weights are not assigned at this stage because the relative importance of individual criteria depends on the specific healthcare workflow and deployment environment, while the currently available evidence is not sufficiently comparable to justify weighted scoring. Where established benchmarks are unavailable, the specified thresholds are treated as initial engineering targets rather than validated acceptance criteria. Both criterion weighting and acceptance thresholds should be refined and validated through empirical IVR testing and consultation with speech-technology specialists and healthcare professionals.

3.1 Evaluation Criteria for STT Models

3.1.1 Linguistic and Domain Accuracy

Given that Latvian is a low-resource and morphologically rich language, in the context of large-scale speech technologies, robust Latvian-language support is a primary requirement. In medical customer-service and IVR scenarios, STT systems must handle Latvian inflectional complexity, pronunciation and accent variability, spontaneous telephone speech, medical and administrative vocabulary, and background noise typical of call environments.

Domain-specific robustness is equally essential. Models must accurately recognize medical terminology, drug names, medical abbreviations, dates, numeric expressions, and personal names. In medical contexts, not all transcription errors are equivalent; misrecognition of a clinically relevant entity can significantly affect downstream interpretation and response generation.

General recognition accuracy should therefore be evaluated using WER and Character Error Rate (CER) on a common Latvian test set representative of the intended IVR scenario. These measures should be complemented by an Entity Error Rate (EER) for predefined clinically and operationally important entities, such as medical terms, medication names, dates, numbers, and personal names. EER can be calculated analogously to WER by considering substitutions, deletions, and insertions of reference entities relative to the total number of reference entities. Entity-level recall and exact-match accuracy may additionally be reported where appropriate.

In general, no universal WER threshold defines safe or acceptable medical speech recognition across clinical applications. Rather than using a fixed accuracy threshold to determine clinical suitability, the framework considers WER, CER, and entity-level accuracy, with particular attention to errors involving safety-relevant information. Based on prior research suggesting that a WER below approximately 10% indicates acceptable recognition under controlled conditions [10], this value is adopted as an initial target for the intended prototype. However, its adequacy must be assessed together with entity-level performance and validated under representative medical IVR conditions.

3.1.2 Model Architecture and Deployment Suitability

Both open-source and commercial models with explicit Latvian support or demonstrated adaptation potential may be considered if they meet the linguistic, operational, and integration requirements of the intended application. Open-source architectures can be advantageous in medical environments because they permit controlled local deployment, customization, and domain adaptation without requiring speech data to be transferred to an external model provider. Commercial services may also be considered if they provide adequate Latvian support and meet the required performance, data-processing, and integration requirements.

Additional criteria include support for domain adaptation or fine-tuning, vocabulary or contextual biasing where available, streaming inference, and practical service integration through APIs or equivalent interfaces. These capabilities are evaluated as technical characteristics rather than as direct indicators of recognition quality.

Deployment architecture is also considered from a data-governance perspective. Local or controlled-cloud deployment can facilitate GDPR-compliant processing by providing greater control over sensitive data, but does not ensure compliance by itself. GDPR compliance depends on the complete processing workflow, including data minimization, retention, access control, security measures, and organizational governance.

3.1.3 Real-Time Performance and Operational Scalability

Conversational IVR systems require sufficiently low processing latency to avoid disruptive pauses between user input and system response. STT latency should therefore be measured from detected end-of-utterance to the availability of the final transcript, while streaming systems should additionally report time to the first usable partial hypothesis. For the intended prototype, approximately one second from end-of-utterance to final recognition is adopted as an initial engineering target for responsive interaction, subject to validation through subsequent user testing rather than as an established clinical threshold.

Scalability should be evaluated by measuring latency, accuracy, and throughput under increasing numbers of concurrent speech streams. The initial prototype will use at least three concurrent sessions, as the target clinic currently has two employees handling incoming calls; supporting three parallel interactions would therefore already increase simultaneous call-handling capacity. This remains a prototype-specific goal rather than a general healthcare capacity requirement, and subsequent operational capacity should be determined based on expected call volume, service-level requirements, and available computational resources. For reproducibility, empirical latency and concurrency testing should report the hardware configuration, inference runtime, model precision or quantization, decoding parameters, batch size, and other runtime settings that may affect performance.

3.2 Evaluation Criteria for TTS Models

3.2.1 Linguistic and Terminological Accuracy

In medical IVR scenarios, synthesized speech must prioritize clarity, intelligibility, and consistent pronunciation. The TTS system must generate fluent Latvian speech and provide sufficient control over speech rate and prosody for different prompt types, including confirmations, procedural instructions, and informational messages. Accurate pronunciation of medical terms, medication names, personal names, dates, abbreviations, and administrative information is essential. Candidate systems should therefore support pronunciation control through phoneme input, custom lexicons, grapheme-to-phoneme (G2P) overrides, or text-normalization rules. Medical-term pronunciation accuracy should be evaluated using representative IVR prompts and assessed by native Latvian listeners, including healthcare professionals for specialized terminology. Results are reported as the proportion of terms judged correctly pronounced, with error analysis distinguishing terminology, names, abbreviations, dates, and numerical expressions.

3.2.2 Model Architecture and Customization Capabilities

Candidate TTS systems are assessed with respect to Latvian-language availability, generation architecture, pronunciation and text-normalization control, deployment model, and integration capabilities. Both open-source and commercial systems may be considered where sufficient technical information is available for assessment.

Relevant architectural characteristics include autoregressive or non-autoregressive synthesis, integrated or separate waveform generation, and support for streaming. Pronunciation and domain customization may involve phoneme input, pronunciation dictionaries, G2P adaptation, and text normalization. Service interfaces such as REST APIs or WebSockets are considered implementation requirements rather than speech-quality criteria.

Deployment architecture is also considered from a data-governance perspective. Local or controlled-cloud deployment can provide greater control over sensitive data and facilitate GDPR-compliant processing, but does not ensure compliance by itself. As with STT, compliance depends on a complete technical and organizational processing workflow.

3.2.3 Speech Quality, Intelligibility, and Real-Time Performance

TTS evaluation should combine perceptual, pronunciation-focused, and operational measures. Perceived naturalness and overall speech quality should be evaluated through native-listener Mean Opinion Score (MOS) ratings using a consistent protocol and common Latvian medical-domain prompts across all candidate voices. Because MOS results depend on the listener population, prompt set, playback conditions, and evaluation protocol, values obtained across different datasets or languages should not be considered directly comparable. Short-Time Objective Intelligibility (STOI) is not used as a general TTS quality measure but may be applied in subsequent experiments to assess degradation caused by telephony transmission, codecs, or other acoustic processing.

Intelligibility can be evaluated through transcription-based listening tests or ASR-based WER/CER calculated for synthesized utterances, while domain-specific pronunciation accuracy should be assessed separately for medical terminology and other critical entities. In the later IVR pilot, these measures should be complemented by comprehension testing to determine whether users correctly understand instructions and clinically relevant information.

Real-time performance should be evaluated using time to first audio (TTFA), defined as the interval between submission of the text synthesis request and availability of the first playable audio segment, together with total synthesis time or real-time factor (RTF). For the intended prototype, a TTFA of approximately one second or less is adopted as an initial engineering target for responsive interaction rather than as a universally established clinical threshold.

Output must also be compatible with the target IVR infrastructure. PCM/WAV at 16 kHz or higher is used as a baseline synthesis format, while conversion to the codec and sampling rate required by the telephony channel should be evaluated separately as part of system-level testing.

3.3 System-Level IVR Integration Criteria

Component-level STT and TTS performance do not alone determine the usability of an integrated medical IVR system. System-level evaluation should therefore consider the complete interaction pipeline, including speech capture, endpoint detection, STT, dialogue processing, response generation, TTS, and telephony transmission. The following criteria define the measurements proposed for subsequent prototype evaluation (where established clinical thresholds are unavailable, the specified values are treated as initial engineering targets to be validated through empirical testing and stakeholder assessment):

- *Responsiveness.* End-to-end response latency should be measured from the end of the caller's utterance to the first playable system audio. Component-level measurements should include STT final-transcript and partial recognition results, and TTS time to first audio and real-time factors. Approximately one second for final STT output and TTFA can be used as an initial prototype target, while acceptable end-to-end latency should be established through pilot testing.
- *Domain-specific accuracy and speech quality.* STT should be evaluated using WER, CER, and entity-level errors on a common medical-domain speech set, with a WER below approximately 10% treated as an initial engineering target. TTS evaluation should include

native-listener MOS and correct pronunciation of predefined medical terminology, with healthcare professionals involved in assessing specialized terms.

- *Telephony robustness.* Recognition accuracy and speech intelligibility should be compared between source or 16 kHz audio and representative 8 kHz telephony conditions, including background noise where applicable. Changes in WER, CER, entity-level accuracy, and user comprehension should be reported rather than assuming a fixed degradation threshold.
- *Concurrency and scalability*. Latency, recognition accuracy, and TTS synthesis performance should be measured under increasing numbers of concurrent sessions. The initial prototype will compare single-session operation with at least three concurrent sessions. This represents a prototype testing target rather than a general healthcare capacity requirement.
- *Critical-information handling and error recovery.* The integrated system should evaluate successful confirmation of safety-relevant information, such as names, dates, medications, and other critical entities. Error-recovery tests should introduce recognition errors, missing input, and ambiguous responses and measure the proportion of interactions successfully recovered.
- *Data protection.* Data protection should be assessed at the workflow level through review of data flows, storage and retention, access controls, security measures, and other relevant safeguards. Models should not be classified individually as "GDPR compliant," since compliance depends on the complete processing and governance environment.

4 Comparative Assessment and Candidate Selection for Latvian Medical IVR

This section applies the requirement-driven framework defined in Section 3 to assess candidate STT and TTS models for subsequent empirical evaluation and development of a Latvian medical IVR prototype. The assessment draws on scientific publications, model documentation, public repositories, and Latvian-language studies. The results therefore represent a structured comparative assessment rather than a standardized benchmark or direct comparison of model performance under identical conditions.

Reported accuracy and speech-quality values should be interpreted with caution because they originate from different datasets, languages, model variants, and evaluation protocols. Real-time characteristics are assessed primarily from documented architectural and implementation properties rather than directly comparable latency measurements. Where Latvian or medical-domain evidence is unavailable, this is explicitly indicated.

The purpose of the comparison is to identify models whose documented characteristics align with the linguistic, architectural, operational, and deployment requirements defined in Section 3, to identify relevant trade-offs, and to select a limited set of candidates for subsequent testing. A definitive ranking requires common-condition evaluation using the same Latvian medical-domain test data, hardware configurations, runtime settings, and measurement procedures.

4.1 Assessment of STT Models for a Real-Time Latvian Medical IVR Solution

4.1.1 Comparative Assessment of STT Models

The comparative assessment of STT models follows the criteria defined in Section 3.1, focusing on Latvian and medical-domain recognition, architecture and deployment, and real-time performance and scalability. IVR-specific aspects from Section 3.3, including telephony robustness, endpoint detection, output characteristics, and integration feasibility, are also considered.

Accuracy alone does not determine suitability for conversational IVR and must be considered together with responsiveness, computational requirements, and reliable utterance boundary detection. Although some models provide internal silence or segmentation mechanisms, IVR endpoint detection may still require a dedicated VAD or endpointing component and should not be considered an inherent advantage of a particular STT architecture.

Within the scope of this study, a broad spectrum of STT models with documented Latvian support or multilingual capabilities including Latvian is considered. The reviewed models include *OpenAI Whisper (large-v3)* [25], [26], *Whisper-large-v3-lv-late-cv19* (fine-tuned Latvian version) [10], [11], [27], *Whisper-small-lv* (fine-tuned Latvian version) [28], *Meta Wav2Vec 2.0 (base)* [29], [30], *Wav2Vec2-Large-XLSR-Latvian* [31], Meta MMS ASR (*MMS-1B-all*) [32], [33], Meta Omnilingual ASR (*OmniASR_LLM_1B*) [34], [35], *Canary-1b-v2* [36], [37], *Parakeet-tdt-0.6b-v3* [36], [38], and *mHuBERT-147* [39], [40]. However, to avoid redundancy in Table 1, generic base models such as *Whisper-large-v3* and *Wav2Vec 2.0* are not included separately. Instead, only their Latvian-specific fine-tuned versions (*Whisper-large-v3-lv-late-cv19* and *Wav2Vec2-Large-XLSR-Latvian*) are shown, since fine-tuning does not fundamentally change the underlying architecture or computational characteristics of the base models. Therefore, a separate analysis would largely duplicate the technical considerations. Moreover, language-adapted versions provide a more realistic representation of expected performance in the Latvian medical IVR context.

Table 1 consolidates the model characteristics most relevant to candidate selection. Input file formats are not treated as a primary criterion because audio can be converted and resampled as required; instead, robustness to 8 kHz telephony audio is evaluated at the system level. Output assessment focuses on native recognition characteristics, including punctuation, capitalization, and timestamps where available. Medical-specific normalization of numbers, dates, abbreviations, and terminology may require additional post-processing.

Reported WER and CER values are drawn from the cited sources and reflect different datasets and evaluation conditions. They are therefore used as indicative evidence of Latvian-language performance rather than as directly comparable benchmark results, particularly in the medical domain. For instance, [10] reports substantial variation across medical subdomains and adaptation conditions, including a WER of 45.9% for visual imaging and 12.1% for physical rehabilitation after domain fine-tuning. Consequently, a single WER value should not be generalized as representative of Latvian medical speech recognition.

Real-time characteristics in Table 1 are based primarily on documented architectural and implementation properties rather than standardized latency measurements. Actual latency and throughput depend on the hardware, runtime, model configuration, and streaming setup. These factors will therefore be controlled and reported in the subsequent common-condition evaluation.

Licensing and deployment restrictions are also considered. Most models permit local deployment, while *MMS-1B-all* and *mHuBERT-147* are distributed under licenses that restrict commercial use. Local deployment may facilitate greater control over sensitive speech data, but should not be interpreted as evidence of GDPR compliance, which depends on the complete technical and organizational processing workflow discussed in Section 3.1.2.

Overall, the candidates represent different trade-offs in accuracy, computational requirements, and real-time capability; the available evidence supports shortlisting for common-condition testing rather than a definitive performance ranking. Section 4.1.2 further examines these trade-offs and defines the candidates selected for empirical evaluation.

4.1.2 Candidate Selection for Empirical Evaluation

Our comparative assessment reveals a central trade-off between recognition accuracy, streaming capability, and computational efficiency. No candidate can currently be identified as superior for Latvian medical IVR because the available accuracy evidence originates from different datasets and evaluation conditions, while real-time characteristics have not been evaluated under common runtime conditions.

Whisper-large-v3-lv is considered as an accuracy-oriented reference candidate for evaluating medical-domain WER, CER, and entity-level accuracy. Published results demonstrate substantial variation across medical subdomains, reinforcing the need for evaluation under the intended IVR conditions. The main limitation is the lack of native low-latency streaming, which may require segmentation, external endpoint detection, optimization, and appropriate computational resources.

Table 1. Comparison of STT models for real-time Latvian medical IVR

Model (architecture; parameters; references)	Latvian and medical-domain evidence	Adaptation capability	Streaming and endpointing	Output characteristics	Indicative computational and latency profile	Key IVR considerations
Whisper-large-v3-lv-late-cv19 Seq2Seq model (~2B) **[10], [11], [27]**	Latvian-adapted; WER ~3.2% on Common Voice (CV'19) and ~12.8% on LATE-Media. Medical-domain WER varies by subdomain: 45.9% for visual imaging and 12.1% for physical rehabilitation after fine-tuning [10].	Fine-tuning and domain adaptation are possible; medical adaptation has been reported.	Primarily segment-based, without native low-latency streaming; IVR use may require chunking and external endpoint detection.	Punctuation and capitalization supported; timestamps available. Medical numbers, abbreviations, and terminology normalization may require post-processing.	Relatively high computational demand; GPU preferred; quantization can reduce requirements.	Available Latvian and medical-domain accuracy evidence, with performance varying across medical subdomains; real-time IVR performance requires validation.
Whisper-small-lv Seq2Seq (~0.2B) **[28]**	Latvian fine-tuned; reported WER ~9.6% on FLEURS [28]. No identified published Latvian medical-domain benchmarks.	Fine-tuning/ domain adaptation possible.	Chunk-based partial recognition possible; external endpoint detection may be required for reliable IVR operation.	Punctuation/ capitalization and timestamps are available; domain-specific normalization may require post-processing.	Lower computational requirements than large Whisper; CPU/GPU deployment possible; inference speed depends on implementation.	Resource-efficient Latvian candidate; medical-domain accuracy and streaming behavior require further validation.
Wav2Vec2-Large-XLSR-Latvian Connectionist Temporal Classification model (~0.3B) **[31], [41]**	Latvian fine-tuned; reported WER ~29.95% on CV LV [31]. No identified medical-domain evaluation.	Fine-tuning is supported; external language models and domain lexicons can be integrated.	CTC architecture is compatible with low-latency processing, but reliable IVR endpointing typically requires an external component.	Primarily raw/ unformatted transcript; punctuation, capitalization, and domain-specific normalization generally require post-processing.	Relatively low inference cost; CPU deployment possible. Language-model decoding can increase latency.	Lower computational requirements, but weaker reported Latvian accuracy; medical domain adaptation likely required.
MMS-1B-all CTC (Wav2Vec2 extension) (~1B) **[10], [32], [33]**	Latvian supported; WER ~16.0% on CV'19 and 29.3% on LATE-Media. Medical-domain WER varies: 55.9% for visual imaging and 17.7% for physical rehabilitation [10].	Fine-tuning possible; multilingual adaptation support.	CTC-based processing can support low-latency implementations; external endpointing/VAD generally required.	Basic transcript output; punctuation, capitalization, and medical normalization require post-processing.	Moderate-to-high resource requirements; GPU preferred, CPU possible with optimization/ quantization. Latency depends on implementation.	Medical-domain evidence is available, with substantial variation across subdomains; non-commercial license limits deployment.

Table 1. Continued

Model (architecture; parameters; references)	Latvian and medical-domain evidence	Adaptation capability	Streaming and endpointing	Output characteristics	Indicative computational and latency profile	Key IVR considerations
OmniASR_LLM_1B Wav2Vec + Llama decoder (~2.2B) **[34], [35]**	Multilingual model including Latvian; average CER ~7.2% on FLEURS-102 and ~6.5% on CV'22 [34]; no identified Latvian-specific or medical-domain evaluations.	Multilingual architecture offers adaptation potential; model-specific customization should be verified.	Not primarily designed as an IVR-native streaming recognizer; endpointing and streaming behavior depend on implementation.	LLM-assisted output processing; medical normalization and factual accuracy require validation.	GPU-oriented architecture; higher computational requirements than smaller CTC models; latency depends strongly on runtime and decoding configuration.	Potentially strong multilingual modeling but limited Latvian medical evidence and uncertain real-time IVR performance.
Canary-1b-v2 FastConformer (~1B) **[36], [37]**	Latvian supported; average WER ~8.3% on FLEURS-24; Latvian-specific WER ~9.66% on FLEURS-LV [36]; no identified medical-domain evaluation.	Supports multilingual adaptation within the NVIDIA NeMo ecosystem.	Streaming depends on the implementation; explicit endpoint detection may be required for IVR.	Supports punctuation/ capitalization and structured transcription features depending on implementation; medical normalization may require post-processing.	GPU-oriented; potentially lower processing latency with optimized NVIDIA runtime, but performance depends on hardware and configuration.	Promising multilingual candidate, but Latvian medical accuracy and real-time behavior require empirical validation.
Parakeet-tdt-0.6b-v3 FastConformer with Token-and-Duration Transducer (TDT) (~0.6B) **[36], [38]**	Latvian supported; average WER ~11.52% on FLEURS-24 and ~22.8% on FLEURS-LV [36]; no identified Latvian medical-domain evaluation.	Supports fine-tuning and domain adaptation.	Supports streaming recognition; external VAD/endpoint detection may still be required for reliable IVR turn detection.	Punctuation/ capitalization depends on implementation; timestamps require verification; medical normalization may require post-processing.	Optimized for GPU inference and streaming; performance depends on runtime and configuration.	Suitable for low-latency evaluation; Latvian medical accuracy requires empirical validation.
mHuBERT-147 HuBERT (95M) **[39], [40]**	Latvian supported; average CER ~21.1% on FLEURS-102 and Latvian-specific CER ~7.6% on FLEURS-LV [39]; no identified Latvian medical-domain evaluation.	Fine-tuning required for task/domain-specific ASR use.	Streaming depends on the ASR implementation; external endpoint detection may be required.	Output formatting depends on the downstream decoder; additional punctuation and normalization components may be required.	Computational requirements depend on fine-tuned implementation and decoder; no standardized IVR latency evidence identified.	Multilingual adaptation candidate; limited evidence for Latvian medical IVR; non-commercial license.

Parakeet-tdt-0.6b-v3 represents a complementary streaming-oriented candidate due to its support for real-time inference and lower computational requirements. However, the available evidence does not establish Latvian medical-domain accuracy comparable to the Latvian-adapted Whisper models. It is therefore included as a candidate for assessing the trade-off between accuracy and latency rather than as a preferred overall STT model.

Whisper-small-lv provides a resource-oriented candidate with lower computational requirements, allowing evaluation of whether sufficient Latvian medical-domain accuracy can be maintained while increasing concurrency or reducing hardware requirements.

Accordingly, *Parakeet-tdt-0.6b-v3*, *Whisper-large-v3-lv*, and *Whisper-small-lv* are shortlisted for common-condition prototype evaluation, representing streaming-oriented, accuracy-oriented, and resource-oriented profiles, respectively. Their relative suitability will be evaluated using the same Latvian medical-domain speech data and runtime conditions, including WER, CER, clinically relevant entity errors, recognition latency, robustness under 8 kHz telephony conditions, and concurrent-call performance.

4.2 Assessment of TTS Models for a Real-Time Latvian Medical IVR Solution

4.2.1 Comparative Assessment of TTS Models

The comparative assessment of TTS models follows the criteria defined in Section 3.2, focusing on Latvian and medical-domain pronunciation support, architecture and customization aspects, speech-quality evidence, real-time performance, and computational requirements. Licensing and deployment characteristics relevant to healthcare IVR integration are also considered.

The analysis includes internationally recognized TTS architectures and Latvian-specific implementations. The architecture-level models include *Tacotron 2* [42], [43], *FastSpeech 2* [44], [45], *VITS* [46], [47], *Deep Voice 3* [48], [49], *VALL-E X* [50], [51], and *Llama-OuteTTS* [52]. Latvian-trained or Latvian-adapted implementations include the *Latvian Female TTS Model* [53], [54], *Latvian Male VITS TTS "Ilvars"* [55], *Latvian Piper TTS "Aivars"* [56], and *MMS TTS (lav)* [57]. Together, these models represent different synthesis architectures and generation pipelines with varying computational and deployment characteristics.

Architecture-level results are distinguished from the evidence for specific Latvian implementations. Results for generic architectures are typically obtained on non-Latvian datasets and cannot be interpreted as evidence of Latvian speech quality. Latvian-specific implementations demonstrate language availability, but often lack standardized evaluations of naturalness, pronunciation, real-time performance, or medical-domain suitability. Table 2 therefore presents these evidence types separately and avoids generalizing results across languages or implementations.

Speech-quality results are interpreted cautiously. MOS values are reported only where a documented evaluation of the corresponding architecture or implementation is available and are not considered directly comparable across different datasets, languages, listener populations, or evaluation protocols. Architecture-level MOS results from non-Latvian datasets are not used to infer the quality of Latvian implementations.

For medical IVR, pronunciation control is particularly important for medical terminology, drug names, abbreviations, personal names, dates, and numerical expressions. The comparison therefore considers support for phoneme-level input and other pronunciation-control mechanisms, as well as whether waveform generation is integrated or requires a separate vocoder.

Real-time characteristics in Table 2 are assessed primarily from documented architectural and implementation properties rather than standardized latency or throughput measurements. TTFA, RTF, and throughput will therefore be measured under common hardware and runtime conditions during the subsequent empirical evaluation.

Table 2. Comparative analysis of TTS models for Latvian medical IVR

Model (architecture; parameters; references)	Latvian-specific evidence	Pronunciation and domain-control capabilities	Generation pipeline	Real-time and computational characteristics	Speech-quality evidence	License and deployment	Key IVR considerations
Tacotron 2 Autoregressive Seq2Seq (attention-based CNN + LSTM) (~25-30M) **[42], [43], [58]**	Latvian Tacotron 2 implementation evaluated using a 60-h corpus derived from Latvian broadcast speech [58]	Grapheme/phoneme input depending on implementation; medical pronunciation requires G2P, lexicon, or preprocessing.	Generates mel-spectrograms; separate neural vocoder required (WaveGlow in the Latvian implementation).	Autoregressive synthesis limits real-time efficiency; GPU preferable. Standardized Latvian IVR latency not reported.	MOS 4.53 reported for the original Tacotron 2 [42]; Latvian Tacotron 2 + WaveGlow: MOS 3.7/5 (10 evaluators) [58]	Apache 2.0 for commonly used implementations; local deployment possible.	Latvian speech-quality evidence is available, but medical pronunciation and real-time IVR performance require evaluation.
FastSpeech 2 Non-autoregressive transformer (~30-35M) **[44], [45]**	No Latvian-specific FastSpeech 2 evaluation identified.	Duration, pitch, and energy control; phoneme-based pronunciation adaptation requires Latvian G2P/lexicon support.	Generates mel-spectrograms; separate neural vocoder required.	Non-autoregressive architecture supports fast synthesis; actual latency depends on the model, vocoder, hardware, and runtime.	MOS 3.83 on LJSpeech dataset (English, ~24 h) [44]; no Latvian MOS identified.	Common implementations use permissive licenses; license depends on the implementation and vocoder. Local deployment is possible.	Fast synthesis and prosody control, but Latvian medical pronunciation and speech quality require empirical evaluation.
VITS (general) End-to-end transformer (acoustic model, vocoder) (~28-35M) **[46], [47]**	Generic architecture underlying several Latvian implementations is considered below.	Text/phoneme front ends depending on implementation; pronunciation control depends on linguistic preprocessing.	End-to-end architecture integrating acoustic modeling and waveform generation; no separate vocoder required at inference.	Relatively efficient end-to-end synthesis; CPU/GPU inference depends on implementation and optimization.	MOS 4.43 on LJSpeech and 4.38 on VCTK dataset [46]; no Latvian MOS identified.	Implementation-dependent license; open implementations support local deployment.	Integrated waveform generation simplifies deployment; Latvian medical pronunciation and real-time performance require evaluation.
Deep Voice 3 CNN-based Seq2Seq (~20-25M) **[48], [49]**	Generic architecture; no Latvian medical-domain evaluation identified.	Supports phoneme-based preprocessing; medical terminology requires pronunciation rules or adapted lexicon.	Generates acoustic features; separate waveform-generation component/vocoder required.	Autoregressive architecture; less efficient than newer non-autoregressive or end-to-end approaches.	MOS 3.78 with WaveNet on single-speaker data [48]; no Latvian MOS identified.	MIT for commonly cited implementation; local deployment possible.	Separate vocoder and autoregressive synthesis increase complexity; Latvian medical performance requires evaluation.

Table 2. Continued

Model (architecture; parameters; references)	Latvian-specific evidence	Pronunciation and domain-control capabilities	Generation pipeline	Real-time and computational characteristics	Speech-quality evidence	License and deployment	Key IVR considerations
VALL-E X, Neural codec language model (~0.2-0.8B) **[50], [51]**	Multilingual architecture; no Latvian-specific or medical-domain evaluation identified.	Multilingual and voice-cloning capabilities; Latvian medical pronunciation control requires validation.	Generates neural codec tokens subsequently decoded to waveform audio.	Computationally demanding; real-time performance depends on implementation, hardware, and optimization.	Naturalness MOS 3.54 on EMIME dataset [50]; no Latvian MOS identified.	Research implementation; licensing conditions require verification for deployment.	Voice cloning and multilingual capabilities are available, but Latvian medical pronunciation, generation reliability, and real-time performance require validation.
Llama-OuteTTS-1.0-1B LLM-based autoregressive TTS (~1B) **[52]**	No Latvian-specific or medical-domain evaluations identified.	Prompt-based control; Latvian pronunciation and medical terminology require validation and potential adaptation.	LLM/token-based autoregressive speech generation.	High computational requirements; quantization may reduce resource demand. Standardized IVR latency unavailable.	No MOS evaluation identified; no Latvian-specific results available.	CC-BY-NC-SA-4.0; non-commercial license restricts commercial deployment.	Limited Latvian evidence, high computational requirements, and non-commercial licensing restrict commercial IVR deployment.
Latvian Female VITS-based (parameters not reported) **[53], [54]**	Latvian-specific VITS model; no medical-domain listening evaluation identified.	Phoneme-based front end; medical terminology may require lexicon/G2P adaptation and text normalization.	End-to-end VITS waveform generation; no separate vocoder required.	Lightweight local synthesis; CPU/GPU testing feasible. Standardized TTFA/RTF not reported.	No MOS evaluation was identified for this Latvian implementation.	BSD-3-Clause; local and commercial deployment permitted.	Native Latvian candidate with permissive licensing; medical pronunciation, naturalness, and telephony intelligibility require evaluation.

Table 2. Further continued

Model (architecture; parameters; references)	Latvian-specific evidence	Pronunciation and domain-control capabilities	Generation pipeline	Real-time and computational characteristics	Speech-quality evidence	License and deployment	Key IVR considerations
Latvian Male "Ilvars" VITS-based (parameters not reported) **[55]**	Native Latvian model trained on audiobook speech; no identified medical-domain TTS evaluation; development acknowledges support from the Latvian medical speech project.	Latvian phonetic preprocessing; medical terminology may require additional pronunciation control and text normalization.	End-to-end VITS waveform generation.	Lightweight local inference; standardized TTFA/RTF under IVR conditions not reported.	No MOS evaluation was identified for this Latvian implementation.	Non-commercial use; commercial deployment requires permission from AiLab.	Native Latvian candidate; licensing restricts commercial deployment, while medical pronunciation and speech quality require evaluation.
Latvian Piper "Aivars" VITS, ONNX optimized (parameters not reported) **[56]**	Latvian-specific Piper voice for local synthesis; no medical-domain listening evaluation identified.	Phoneme-based preprocessing provides a basis for lexicon/G2P adaptation and medical text normalization.	VITS-based ONNX model with direct waveform generation; no separate vocoder required.	CPU-oriented local inference; standardized TTFA, RTF, and concurrent-call performance not reported.	No MOS evaluation was identified for this Latvian implementation.	CC0 1.0; permissive for commercial use and local deployment.	CPU-oriented local deployment and permissive licensing; medical pronunciation and speech quality require empirical evaluation.
MMS TTS (lav) VITS-based (~36M) **[57]**	Latvian-specific model within Meta MMS; no medical-domain TTS evaluation identified.	Primarily text-based synthesis; medical terminology may require additional pronunciation preprocessing.	VITS-based end-to-end waveform generation.	Relatively lightweight CPU/GPU inference; standardized Latvian IVR latency not reported.	No MOS evaluation was identified for this Latvian implementation.	CC-BY-NC-4.0; non-commercial license restricts commercial deployment.	Useful Latvian reference model for research comparison, but licensing restrictions and absence of medical-domain quality evidence limit deployment.

Overall, the candidates represent different relative strengths and limitations in speech quality, computational requirements, pronunciation control, and deployment characteristics. The available evidence supports candidate shortlisting for common-condition testing rather than a definitive performance ranking. Section 4.2.2 further examines these trade-offs and defines the candidates selected for empirical evaluation.

4.2.2 Candidate Selection for Empirical Evaluation

The available evidence highlights a trade-off between speech quality, pronunciation control, computational efficiency, and deployment requirements. Although published MOS results are available for several architectures and a Latvian *Tacotron 2* implementation, they were obtained under different datasets and evaluation protocols and do not support a definitive ranking for Latvian medical IVR.

FastSpeech 2 represents a controllability-oriented architecture because of its explicit duration, pitch, and energy prediction. Such control may be useful for medical prompts requiring clear articulation and consistent pacing. However, its practical suitability depends on the availability or development of an adequately trained Latvian implementation, and its non-Latvian MOS results cannot be generalized to Latvian speech.

VITS-based models are particularly relevant because acoustic modeling and waveform generation are integrated, reducing deployment complexity compared with systems requiring a separate vocoder. Several Latvian implementations are available for empirical comparison. *Piper "Aivars"* provides CPU-oriented ONNX inference and permissive licensing, while the *Latvian Female VITS model* provides another locally deployable candidate with permissive licensing. *"Ilvars"* and *MMS-TTS (lav)* provide additional Latvian implementations but have licensing restrictions relevant to commercial deployment. These characteristics do not establish superior speech quality for any of the Latvian VITS-based voices. No MOS evaluations were identified for the *Latvian Female VITS model*, *"Ilvars," Piper "Aivars,"* or *MMS-TTS (lav)*, and their medical-term pronunciation has not been evaluated under a common protocol. Their relative suitability therefore requires direct empirical comparison.

Accordingly, *Piper "Aivars"* and the *Latvian Female VITS model* are prioritized for the initial common-condition TTS evaluation because they combine Latvian-language availability, local deployment, and permissive licensing. *"Ilvars"* and *MMS-TTS (lav)* may be retained as research references subject to licensing conditions, while *FastSpeech 2* may be included as a controllability-oriented architecture if an appropriate Latvian implementation is available. This shortlist represents candidates for empirical testing rather than a final ranking.

The evaluation should use the same Latvian medical-domain prompt set for all candidate voices and include native-listener MOS, medical-term pronunciation accuracy, intelligibility assessed through listener transcription or ASR-based WER/CER, TTFA, RTF, and performance under the target telephony conditions. The subsequent integrated IVR pilot should additionally assess comprehension of medical and administrative instructions. These common-condition measurements will provide the basis for selecting the final TTS configuration.

5 Conclusions

This study addressed the selection of speech technologies for a real-time Latvian medical IVR system by developing a requirement-driven evaluation framework and applying it to candidate STT and TTS models. The framework considers language and medical-domain performance together with real-time capabilities, computational requirements, deployment characteristics, and system-level IVR requirements. Particular attention was given to the challenges of Latvian as a low-resource and morphologically rich language and to medical terminology and privacy-sensitive healthcare communication.

The comparative assessment does not support identifying a single STT or TTS model as superior for Latvian medical IVR. Available results originate from different datasets, domains, hardware configurations, implementations, and evaluation protocols, while some real-time characteristics are derived from model documentation or architectural properties rather than standardized measurements. The results should therefore be interpreted as a structured candidate-selection and shortlisting rather than as a comparative experimental benchmark.

For STT, Latvian-adapted *Whisper* models provide documented Latvian recognition results, including medical-domain evaluations, but their segment-based processing presents challenges for low-latency conversational use. *Parakeet-tdt-0.6b-v3* provides a complementary streaming-oriented candidate, although its Latvian medical-domain accuracy remains to be established. *Whisper-small-lv* provides a resource-oriented alternative for evaluating compromises between accuracy, computational requirements, and concurrent-call performance. These candidates represent complementary configurations for subsequent common-condition evaluation rather than a performance ranking.

For TTS, Latvian-specific implementations demonstrate the availability of locally deployable speech synthesis, while architectures such as *FastSpeech 2* provide additional options where suitable Latvian adaptation is available. Published MOS evidence was identified for several general architectures and for a Latvian *Tacotron 2* implementation, but these results were obtained under different datasets and evaluation protocols and are not directly comparable. *Piper "Aivars"* and the *Latvian Female VITS model* are prioritized for initial testing based on Latvian-language availability, local deployment, and permissive licensing rather than demonstrated superiority in speech quality. Comparable MOS, medical-term pronunciation, and standardized real-time measurements for these voices remain to be established experimentally.

The analysis also indicates that model-level characteristics alone are insufficient for determining suitability for medical IVR. STT evaluation must consider clinically relevant entity errors, endpoint detection, telephony robustness, latency, and concurrent-call performance, while TTS requires common-condition evaluation of naturalness, intelligibility, medical-term pronunciation, and real-time synthesis. Local or controlled deployment may facilitate greater control over sensitive healthcare data but does not itself establish GDPR compliance, which depends on the complete technical and organizational processing workflow.

Overall, the study provides a structured basis for identifying suitable STT and TTS candidates and defining the measurements required for subsequent prototype evaluation. The resulting shortlist narrows the range of candidate models for subsequent empirical testing but does not constitute a final model ranking. The proposed framework and identified trade-offs may also apply to speech-enabled healthcare services in other low-resource language settings where domain-specific data and standardized evaluations remain limited.

6 Future Work

The next stage of the research will focus on empirical validation of the selected STT and TTS candidates under common and reproducible conditions reflecting the intended Latvian medical IVR environment. This will complement the literature- and documentation-based assessments with directly comparable experimental evidence.

For STT, the shortlisted models will be evaluated using common Latvian speech data containing medical and administrative vocabulary corresponding to the intended IVR workflow. Evaluation will include WER, CER, and entity-level analysis of clinically and operationally important information, including medical terminology, drug names, personal names, dates, and numerical expressions. A further experiment will investigate whether an LLM-based post-processing layer can reduce medical terminology and entity errors without introducing semantic substitutions or unsupported information. Robustness will be evaluated under telephone-quality conditions, including 8 kHz speech and realistic background noise. Real-time measurements will include final-transcript latency, partial-hypothesis latency where streaming is supported, and throughput

under concurrent calls. Hardware, inference runtime, numerical precision or quantization, decoding parameters, and other relevant settings will be reported to ensure reproducibility.

For TTS, the selected Latvian voices will synthesize a common set of medical-domain prompts and be evaluated using a consistent listening protocol. Native Latvian listeners will assess naturalness and overall quality using MOS, while medical-term pronunciation will be evaluated separately, involving healthcare professionals where domain expertise is required. Intelligibility will be assessed using listener transcription and/or ASR-based WER/CER. TTFA, RTF, and synthesis throughput will be measured under standardized hardware and runtime conditions. As an additional assessment, the effect of conversion to the target telephony format may be evaluated for potential degradation in intelligibility and pronunciation clarity.

Following component-level testing, the selected STT and TTS configurations will be integrated into an end-to-end medical IVR prototype. System-level evaluation will examine interaction latency, turn-taking, information-capture accuracy, error recovery, and concurrent-call performance. Comprehension testing will assess whether users correctly understand medical and administrative information delivered through synthesized speech. The validation stage will involve speech-technology specialists, healthcare professionals, and representative end users. Their input will support refinement of the evaluation criteria and, where appropriate, establishment of criterion weights and acceptance thresholds for the intended healthcare workflow.

Finally, security and regulatory requirements will be assessed at the level of the complete IVR processing workflow rather than individual speech models. The assessment will consider data protection, security, and governance requirements across the complete IVR workflow and its integration within healthcare IT infrastructure. The combined technical, user-oriented, and governance evaluation will provide the evidence required for final model selection and assessment of operational deployment feasibility.

Acknowledgment

The research leading to these results has received funding from the project “Competence Centre of Information and Communication Technologies for Digitalization” of the Recovery and Resilience Facility, contract No. 2.2.1.3.i.0/1/24/A/CFLA/006 signed between the IT Competence Centre and the Central Finance and Contracts Agency, Research No. 2.3 “Development of an Artificial Intelligence Voice Solution Adapted to Customer Service in the Medical Field for Process Digitisation and Its Quality Improvement”.

AI Assistance Disclosure: Portions of the manuscript were edited for language clarity and stylistic refinement using ChatGPT (OpenAI). The authors reviewed and validated all content and remain fully responsible for the accuracy, interpretation, and conclusions of the work.

References

[1] Y. Chen, C. Zhang, R. Bai, T. Sun, W. Ding, and R. Wang, “A Review of Medical Text Analysis: Theory and Practice,” *Information Fusion*, vol. 119, article 103024, 2025. Available: https://doi.org/10.1016/j.inffus.2025.103024

[2] N. Kondai, A. Devapangu, S. K. Hosur, P. Raj, N. Tiwari, and K. S. Nataraj, “Enhancing Medical ASR Accuracy through the Integration of T5-Based Small Language Models and Error Correction Mechanisms,” in *Proceedings of the 2024 IEEE Conference on Engineering Informatics (ICEI)*, pp. 1–6, 2024. Available: https://doi.org/10.1109/ICEI64305.2024.10912105

[3] S. J. Adams, J. N. Acosta, and P. Rajpurkar, “How Generative AI Voice Agents Will Transform Medicine,” *npj Digital Medicine*, vol. 8, article 353, 2025. Available: https://doi.org/10.1038/s41746-025-01776-y

[4] L. Liu et al., “A Survey on Medical Large Language Models: Technology, Application, Trustworthiness, and Future Directions,” *arXiv preprint arXiv:2406.03712*, 2024. Available: https://doi.org/10.48550/arXiv.2406.03712

[5] H. Ahlawat, N. Aggarwal, and D. Gupta, “Automatic Speech Recognition: A Survey of Deep Learning Techniques and Approaches,” *International Journal of Cognitive Computing in Engineering*, vol. 6, pp. 201–237, 2025. Available: https://doi.org/10.1016/j.ijcce.2024.12.007

[6] A. M. Alkalbani et al., "A Systematic Review of Large Language Models in Medical Specialties: Applications, Challenges and Future Directions," *Information*, vol. 16, no. 6, article 489, 2025. Available: https://doi.org/10.3390/info16060489

[7] R. Safarik and L. Mateju, "Automatic Development of ASR System for an Under-Resourced Language," in *Proceedings of the 41st International Conference Telecommunications and Signal Processing (TSP)*, pp. 100–103, 2018. Available: https://doi.org/10.1109/TSP.2018.8441243

[8] R. *Dargis* et al., "BalsuTalka.lv – Boosting the Common Voice Corpus for Low-Resource Languages," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2080–2085, 2024. Available: https://doi.org/10.63317/3and6sceax9m

[9] M. Pinnis, A. Salimbajevs, and I. Auzina, "Designing a Speech Corpus for the Development and Evaluation of Dictation Systems in Latvian," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 775–780, 2016. Available: https://doi.org/10.63317/3vzmvy6qw7va

[10] I. Auzina et al., "Recent Latvian Speech Corpora for Linguistic Research and Technology Development," *Baltic Journal of Modern Computing*, vol. 12, no. 4, pp. 646–658, 2024. Available: https://doi.org/10.22364/bjmc.2024.12.4.24

[11] A. Znotins, D. Gosko, and N. Gruzitis, "LATE: Open Source Toolkit for Latvian and Latgalian Speech Transcription," in *Proceedings INTERSPEECH 2025*, pp. 306–307, 2025. Available: https://www.isca-archive.org/interspeech_2025/znotins25_interspeech.pdf

[12] A. Salimbajevs and J. Kapočiūtė-Dzikienė, "Automatic Speech Recognition Model Adaptation to Medical Domain Using Untranscribed Audio," in *Digital Business and Intelligent Systems. Baltic DB&IS 2022. Communications in Computer and Information Science*, vol. 1598, Springer, pp. 65–79, 2022. Available: https://doi.org/10.1007/978-3-031-09850-5_5

[13] J. Huh, S. Park, J. E. Lee, and J. C. Ye, "Improving Medical Speech-to-Text Accuracy using Vision-Language Pre-training Models," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 3, pp. 1692–1703, 2024. Available: https://doi.org/10.1109/JBHI.2023.3345897

[14] A. Znotins, N. Gruzitis, and R. Dargis, "From Conversational Speech to Readable Text: Post-Processing Noisy Transcripts in a Low-Resource Setting," in *Proceedings of the 10th Workshop on Noisy and User-generated Text (WNUT)*, pp. 143–148, 2025. Available: https://doi.org/10.18653/v1/2025.wnut-1.15

[15] R. Dargis and I. Auzina, "Towards a Modern Text-to-Speech System for Latvian," in *Human Language Technologies – The Baltic Perspective*, vol. 307, IOS Press, pp. 26–29, 2018. Available: https://doi.org/10.3233/978-1-61499-912-6-26

[16] I. Auzina, et al., "Development of a Specialized Latvian Speech Corpus and Pronunciation Dictionary for the Linguistic Analysis and Systematic Transcription of Visual Diagnostic Examinations," *Letonica*, vol. 47, pp. 244–262, 2022 (in Latvian). Available: https://doi.org/10.35539/LTNC.2022.0047.I.A.R.D.B.S.N.G.M.G.A.S.K.S.244.263

[17] R. Dargis, N. Gruzitis, I. Auzina, and K. Stepanovs, "Creation of Language Resources for the Development of a Medical Speech Recognition System for Latvian," in *Human Language Technologies – The Baltic Perspective*, vol. 328, IOS Press, pp. 135–141, 2020. Available: https://doi.org/10.3233/FAIA200615

[18] M. Kronis, A. Salimbajevs, and M. Pinnis, "Code-Mixed Text Augmentation for Latvian ASR," in *Proceedings of the 2024 Joint International Conference Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 3469–3479, 2024. Available: https://doi.org/10.63317/33xr6bnbtmun

[19] N. Gruzitis, R. Dargis, V. J. Lasmanis, G. Garkaje, and D. Gosko, "Adapting Automatic Speech Recognition to the Radiology Domain for a Less-Resourced Language: The Case of Latvian," in *Intelligent Sustainable Systems*, *Lecture Notes in Networks and Systems*, vol. 333, Springer, pp. 267–276, 2022. Available: https://doi.org/10.1007/978-981-16-6309-3_27

[20] I. Skadina, et al., "Latvian Language in the Digital Age: The Main Achievements in the Last Decade," *Baltic Journal of Modern Computing*, vol. 10, no. 3, pp. 490–503, 2022. Available: https://doi.org/10.22364/bjmc.2022.10.3.21

[21] Children's Clinical University Hospital, "Automated Voice Communication Solutions for the Healthcare Industry," 2024 (in Latvian). Available: https://www.bkus.lv/lv/automatizeti-balss-komunikacijas-risinajumi-veselibas-nozarei. Accessed on Feb. 16, 2026.

[22] Tilde, "TILDE Develops New Voice-Enabled AI Solution to Support Hospital Visitors and Doctors," *Tilde News*, 2024. Available: https://tilde.ai/news/tilde-develops-new-voice-enabled-ai-solution-to-support-hospital-visitors-and-doctors/. Accessed on Feb. 16, 2026.

[23] SoftMedico, "Viola," n.d. Available: https://viola.softmedico.eu/. Accessed on July 22, 2026.

[24] Labs of Latvia, “An AI-Powered Transcription Tool for the Medical Field in the Latvian Language Has Been Developed,” 2025 (in Latvian). Available: https://labsoflatvia.com/aktuali/izstradats-maksliga-intelekta-transkripcijas-riks-medicinai-latviesu-valoda. Accessed on July 22, 2026.

[25] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, vol. 202, article 1182, pp. 28492–28518, 2023. Available: https://dl.acm.org/doi/10.5555/3618408.3619590

[26] Hugging Face, “Whisper,” OpenAI Models, 2024. Available: https://huggingface.co/openai/whisper-large-v3. Accessed on Jan. 15, 2026.

[27] Hugging Face, “General-purpose Latvian ASR model,” AI Lab at IMCS (University of Latvia) Models, 2024. Available: https://huggingface.co/AiLab-IMCS-UL/whisper-large-v3-lv-late-cv19. Accessed on Jan. 15, 2026.

[28] Hugging Face, “Latvian Whisper Small Speech Recognition Model,” Raivis Dejus Models, 2024. Available: https://huggingface.co/RaivisDejus/whisper-small-lv/blob/main/README.md. Accessed on Jan. 15, 2026.

[29] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “Wav2Vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, article 1044, pp. 12449–12460, 2020. Available: https://dl.acm.org/doi/10.5555/3495724.3496768

[30] Hugging Face, “Wav2Vec2,” Transformers Documentation, 2025. Available: https://huggingface.co/docs/transformers/main/model_doc/wav2vec2. Accessed on Jan. 17, 2026.

[31] Hugging Face, “Wav2Vec2-Large-XLSR-Latvian,” Jim O'Regan Models, 2025. Available: https://huggingface.co/jimregan/wav2vec2-large-xlsr-latvian-cv. Accessed on Jan. 13, 2026.

[32] V. Pratap et al., “Scaling Speech Technology to 1,000+ Languages,” *The Journal of Machine Learning Research*, vol. 25, no. 1, article 97, pp. 4798–4849, 2024. Available: https://dl.acm.org/doi/10.5555/3722577.3722674

[33] Hugging Face, “Massively Multilingual Speech (MMS) – Finetuned ASR - ALL,” AI at Meta Models, 2023. Available: https://huggingface.co/facebook/mms-1b-all. Accessed on Jan. 17, 2026.

[34] G. Keren et al., “Omnilingual ASR: Open-Source Multilingual Speech Recognition for 1600+ Languages,” *arXiv preprint arXiv:2511.09690*, 2025. Available: https://doi.org/10.48550/arXiv.2511.09690

[35] Hugging Face, “Omnilingual ASR: Open-Source Multilingual Speech Recognition for 1600+ Languages,” AI at Meta Models, 2025. Available: https://huggingface.co/facebook/omniASR-LLM-1B. Accessed on Jan. 16, 2026.

[36] M. Sekoyan et al., “Canary-1B-v2 & Parakeet-TDT-0.6B-v3: Efficient and High-Performance Models for Multilingual ASR and AST,” *arXiv preprint arXiv:2509.14128*, 2025. Available: https://doi.org/10.48550/arXiv.2509.14128

[37] Hugging Face, “Canary 1B v2: Multitask Speech Transcription and Translation Model,” NVIDIA Models, 2025. Available: https://huggingface.co/nvidia/canary-1b-v2. Accessed on Jan. 17, 2026.

[38] Hugging Face, “parakeet-tdt-0.6b-v3: Multilingual Speech-to-Text Model,” NVIDIA Models, 2025. Available: https://huggingface.co/nvidia/parakeet-tdt-0.6b-v3. Accessed on Jan. 17, 2026.

[39] M. Z. Boito, V. Iyer, N. Lagos, L. Besacier, and I. Calapodescu, “mHuBERT-147: A Compact Multilingual HuBERT Model,” *arXiv preprint arXiv:2406.06371*, 2024. Available: https://doi.org/10.48550/arXiv.2406.06371

[40] GitHub, “mHuBERT-147: A Compact Multilingual HuBERT Model,” UTTER Project Repository, 2024. Available: https://github.com/utter-project/fairseq/tree/main/examples/mHuBERT-147. Accessed on Jan. 18, 2026.

[41] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, “Unsupervised Cross-Lingual Representation Learning for Speech Recognition,” in *Proceedings INTERSPEECH 2021*, pp. 2426–2430, 2021. Available: https://doi.org/10.21437/Interspeech.2021-329

[42] J. Shen et al., “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions,” in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783, 2018. Available: https://doi.org/10.1109/ICASSP.2018.8461368

[43] GitHub, “Tacotron 2,” NVIDIA Corporation Repository, 2024. Available: https://github.com/NVIDIA/tacotron2. Accessed on Jan. 13, 2026.

[44] Y. Ren et al., “FastSpeech 2: Fast and High-Quality End-to-End Text to Speech,” *arXiv preprint arXiv:2006.04558*, 2020. Available: https://doi.org/10.48550/arXiv.2006.04558

[45] GitHub, “FastSpeech 2,” Chung-Ming Chien Repository, 2023. Available: https://github.com/ming024/FastSpeech2. Accessed on Jan. 14, 2026.

[46] J. Kim, J. Kong, and J. Son, “Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech,” in *Proceedings of Machine Learning Research*, vol. 139, pp. 5530–5540, 2021. Available: https://proceedings.mlr.press/v139/kim21f.html

[47] Hugging Face, “VITS,” Transformers Documentation, 2025. Available: https://huggingface.co/docs/transformers/model_doc/vits. Accessed on Jan. 15, 2026.

[48] W. Ping et al., “Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning,” in *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, pp. 1-16, 2018. Available: https://openreview.net/pdf?id=HJtEm4p6Z

[49] GitHub, “Deep Voice 3 (PyTorch),” Ryuichi Yamamoto Repository. Available: https://github.com/r9y9/deepvoice3_pytorch. Accessed on Jan. 14, 2026.

[50] Z. Zhang et al., “Speak Foreign Languages with Your Own Voice: Cross-Lingual Neural Codec Language Modeling,” *arXiv preprint arXiv:2303.03926*, 2023. Available: https://doi.org/10.48550/arXiv.2303.03926

[51] GitHub, “VALL-E-X,” Songting Repository, 2024. Available: https://github.com/Plachtaa/VALL-E-X. Accessed on Jan. 14, 2026.

[52] Hugging Face, “OuteTTS Version 1.0,” OuteAI Models, 2025. Available: https://huggingface.co/OuteAI/Llama-OuteTTS-1.0-1B. Accessed on Jan. 13, 2026.

[53] AI Models, “Latvian Female TTS Model VITS Encoding Trained on CV Dataset at 22050Hz,” AI Models for Text-to-Speech Synthesis, 2022. Available: https://aimodels.org/ai-models/text-to-speech-synthesis/latvian-female-tts-model-vits-encoding-trained-on-cv-dataset-at-22050hz/. Accessed on Jan. 16, 2026.

[54] GitHub, “tts_models--lv--cv--vits.zip,” Coqui Repository, 2022. Available: https://github.com/coqui-ai/TTS/releases/tag/v0.8.0_models. Accessed on Jan. 20, 2026.

[55] R. Dargis and I. Auzina, “Ilvars – Latvian Male VITS Text-to-Speech Model (vers. 2023),” CLARIN-LV Digital Library at IMCS, University of Latvia, 2023. Available: http://hdl.handle.net/20.500.12574/89

[56] Hugging Face, “Latvian Piper TTS Voice ‘Aivars’,” Raivis Dejus Models, 2025. Available: https://huggingface.co/RaivisDejus/Piper-lv_LV-Aivars-medium. Accessed on Jan. 14, 2026.

[57] Hugging Face, “Massively Multilingual Speech (MMS): Latvian Text-to-Speech,” AI at Meta Models, 2023. Available: https://huggingface.co/facebook/mms-tts-lav. Accessed on Jan. 19, 2026.

[58] R. Dargis, P. Paikens, N. Gruzitis, I. Auzina, and A. Akmane, “Development and Evaluation of Speech Synthesis Corpora for Latvian,” in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC)*, pp. 6633–6637, 2020. Available: https://aclanthology.org/2020.lrec-1.818.pdf