International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Beyond Speech-to-Text: Voice-to-Voice Generative AI Systems for Emotion-Aware, Real-Time Healthcare Contact Centers

Bhargavi Kalicheti

Abstract: Healthcare contact centers represent one of the most demanding environments for conversational artificial intelligence, requiring real-time responsiveness, emotional sensitivity, regulatory compliance, and operational scalability. Traditional speech-based AI systems rely on speech-to-text (STT) and text-to-speech (TTS) pipelines that introduce latency, flatten emotional nuance, and fragment conversational context. Recent advances in generative modeling have enabled a new paradigm: voice-to-voice (V2V) generative AI, where spoken input is transformed directly into spoken output without intermediate text representation. This paper explores the architectural foundations, system design principles, and operational implications of deploying V2V generative AI systems for emotion-aware, real-time healthcare contact centers. We analyze how speech-native representations preserve prosody, cadence, and affect, enabling more human-like interactions across administrative, clinical, and insurance-related telephony workflows. The paper presents a cloud-native reference architecture for V2V systems, examines latency optimization strategies critical for telephony-grade performance, discusses multilingual equity considerations, evaluates safety and compliance requirements in regulated healthcare environments, and outlines evaluation frameworks for emotional fidelity, conversational trust, and operational impact. By moving beyond text-mediated conversational AI, voice-to-voice systems represent a foundational shift toward speech-native intelligence capable of transforming healthcare contact center operations at a national scale.

Keywords: *Voice-To-Voice Generative AI, Speech Emotion Recognition, Healthcare Telephony, Affective Computing, Neural Vocoder, Prosody Modeling, Speech-Native AI, Hipaa Compliance, Real-Time Inference, Multilingual Voice AI*

1. Introduction

Healthcare contact centers are at the boundary of human vulnerability, operation complexity, and regulation. Healthcare organizations are often approached by members, patients, providers, and caregivers at the time of stress, urgency, or uncertainty. These aren't just transactional interactions but they are very human, with a touch of emotion, a touch of tone, a touch of hesitation, a touch of urgency. However, the majority of conversational AI systems used in health care telephony today are architecturally based on text processing pipelines, which were never intended to include the acoustic and affective aspects of human speech [1], [2].

Independent Researcher, USA

The prevailing paradigm of speech-based AI is to convert speech to text using automatic speech recognition (ASR), extract intent and meaning in text using natural language understanding (NLU) or large language models (LLMs), and convert text to speech using text-to-speech synthesis. Although it is successful with a variety of transactional applications, this sequential pipeline places structural constraints that have become more and more out of sync with the requirements of healthcare telephony [3], [4]. Any latency that occurs in each conversion stage may break the flow of conversations. Textual abstraction deprives it of paralinguistic signals, including intonation, rhythm and emphasis. Misinterpretation due to transference of error between stages will be occurrent, particularly in high stakes situations of benefits, eligibility checks, scheduling or care coordination.

Voice-to-voice generative AI systems are a break to this paradigm since speech is not a medium of representation but the main form of intelligence. These systems directly work with acoustic and latent speech representations where the input speech is transformed to output speech and still maintains the emotional and conversational dynamics [5], [6]. Instead of transcribing the spoken language into text tokens, V2V models convert the latent speech embeddings into new speech outputs based on the conversational context, intent, and policy restrictions. Text can be implicit in latent space, and is never explicitly created or read as a middle product.

In this paper, I will argue that V2V systems are best suited to healthcare contact centers where emotion-awareness, real-time responsiveness, and trust are crucial. We introduce the architectural concepts, emotional modeling techniques, multilingual equity factors, safety limitations, and assessment structures required to implement production-scale V2V generative AI in controlled healthcare telephony systems. The discussion helps to connect recent progress in speech representations learning [7], neural audio coding [8], affective computing [9], and AI governance in healthcare to a cohesive engineering view.

2. Limitations of Speech-to-Text–Based Conversational Systems in Healthcare Telephony

Although ASR accuracy and neural TTS quality have advanced tremendously, architectures based on speech-to-text have certain inherent limitations in their application to healthcare contact centers. The awareness of these drawbacks contributes to the transition to speech-native intelligence.

The first immediate issue is latency build-up. Delays of several hundred milliseconds in conversations in telephony make perceived responsiveness and trust worse. STT-based systems can add cumulative latency to audio buffering and segmentation, ASR decoding, text-based intent classification or LLM inference, response generation, TTS synthesis, and audio playback. These slows compound on congested networks, complicated reasons, or spikes in calls resulting in unnatural pauses that aggravate callers and lower the rate of self-service containment [11].

Second is the compromise of emotional and prosodic information. Text representations reduce speech to lexical information, and the prosodic features of speech (intonation, rhythm, volume, and inflection of emotion) are lost. Although the sentiment analysis software tries to add back the emotional indicators to the textual pipeline, it does so indirectly and imprecisely, without the richness of the acoustic emotion indicators that are a direct reflection of the state of a caller [12]. Emotional cues can be as important or more important than the words in a healthcare interaction: a shaky voice, a long pause, a rising tone may be an indicator of distress or confusion or urgency that cannot be conveyed through text alone [13].

Third, systems based on STT are brittle to ambiguous or distressed speech. Anxiety or cognitive load speech artifacts are common, and healthcare callers often use fragmented sentences, cross-topic, or disclosed topics. ASR failures under these circumstances propagate into flawed intent recognition or inadequate actions, which essentially destroys conversational trust [14]. Fourth, text-mediated systems create a cognitive dissonance with human conversation: human speech is intrinsically acoustic, and systems that force an artificial textual abstraction are not in line with the natural dynamics of spoken communication.

3. Voice-to-Voice Generative AI: Ideas and Voice Representations

Generative AI systems based on voice to voice use speech representations directly and can be used to transform an acoustic to an acoustic speech without necessarily generating text. The theoretical basis of these systems is inspired by the development of self-supervised speech representation learning, neural audio coding and generative sequence modeling.

Contemporary V2V systems are based on trained latent speech representations that encode phonetic, semantic, and prosodic representations. Models that are self-supervised, like wav2vec 2.0 [7] and HuBERT [15] are trained to learn strong acoustic representations by masking and predicting missing speech features on large unlabeled datasets. These representations are able to not only reproduce phonetic content but also speaker attributes, prosodic patterns, and conversational dynamics. When used in healthcare telephony, these representations allow downstream models to make decisions about emotional state, urgency, and conversational intent out of speech, without the loss of information that occurs when texts are converted to speech.

Neural audio codecs build on this base by offering small, discrete speech representations, which can be

modeled generatively. Audio systems like SoundStream [8] and EnCodec [16] encode audio into a hierarchy of discrete tokens by residual vector quantization, and allow language model-like generation to operate over speech tokens. It was shown that audioLM [5] can use large-scale language modeling on such discrete speech tokens to generate natural, continuous speech that maintains the identity of the speaker, prosody, and conversational rhythm; features that are completely absent in a pipeline based on text.

The generative change of V2V systems is trained on such compressed speech representations based on conversational context, intent signals, and policy constraints. Instead of text tokens, the generative core produces speech token sequence which is decoded by a neural vocoder into high-quality audio [17]. This full pipeline acoustic processing maintains emotional continuity between conversational turns, enabling the system to react to distress empathetically, adjust prosody to suit the context, and remain natural at a level that no text based pipeline can provide. SpeechLM [18] also showed the practicability of single pre-training of speech and language tasks, which allows generative models to learn both semantic knowledge and acoustic generation in a single system.

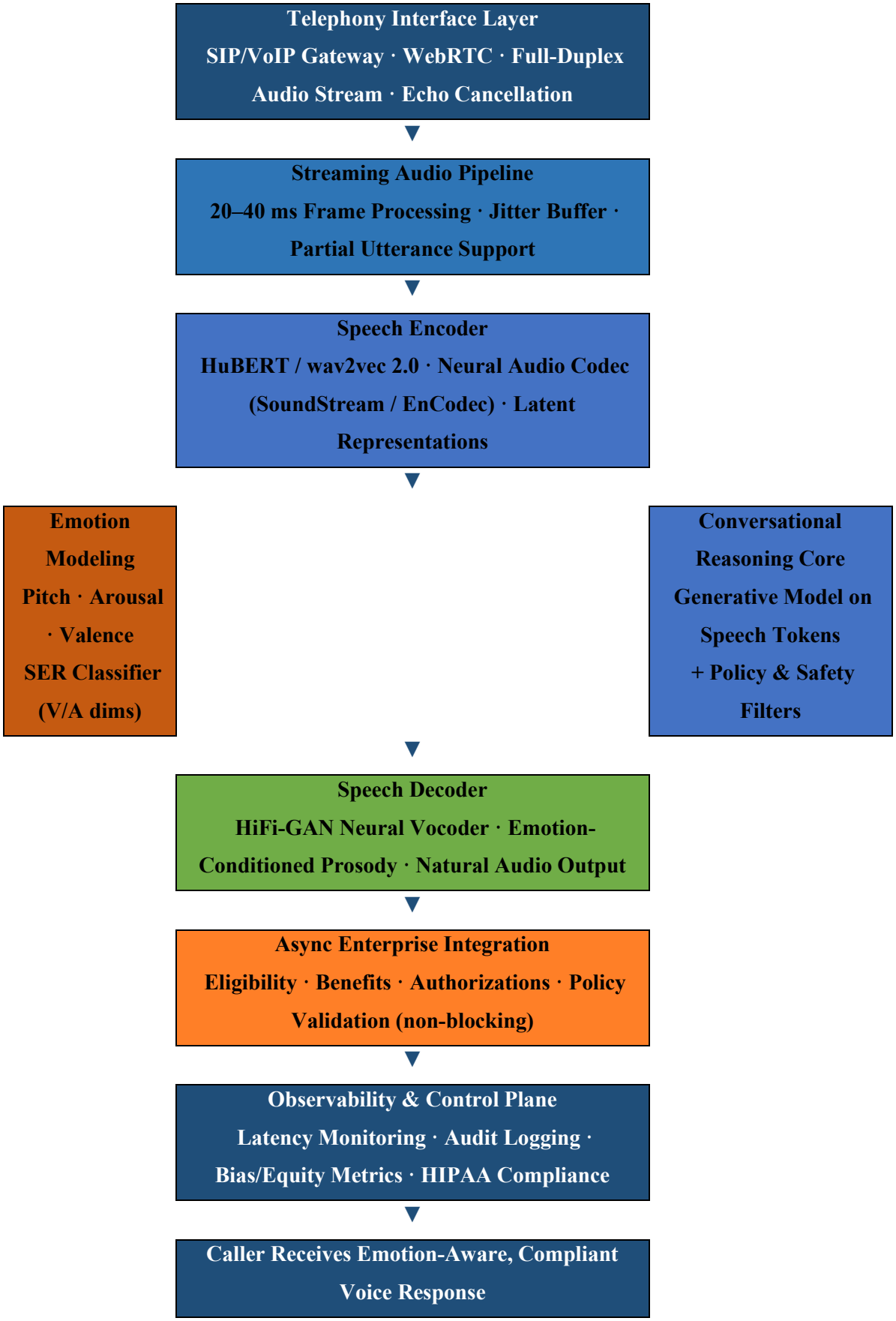
4. Cloud-native reference architecture of voice-to-voice healthcare contact centers

Implementing V2V generative AI in healthcare telephony will need a well-planned, cloud-native system that is performance-wise balanced and scalable, compliant, and emotionally faithful. The reference architecture has seven layers that are coordinated.

The telephony interface layer is concerned with SIP/VoIP ingress and egress call, audio streaming configuration, and session setup. The WebRTC or SIP-over-TLS protocols are used to create low-latency full-duplex audio channels, and jitter buffering and echo cancellation are performed at the network end to provide clean audio before downstream processing. Streaming audio pipeline is in charge of delivering audio frames continuously with intervals of 20 40 ms, which supports partial utterance processing in order to support incremental system response without having to wait until a complete turn has been delivered.

The speech encoder layer uses a pre-trained acoustic encoder, e.g. fine-tuned HuBERT or wav2vec 2.0 variant, to encode streaming audio frames into high-dimensional latent representations. At the same time, a neural audio codec quantizes these representations into sequences of tokens that can be used in generative modeling. The conversational reasoning core takes in encoded speech tokens and conversational state, and calls on a generative model, conditioned on intent, context in the healthcare domain (such as healthcare domain information), and emotion information, to generate response token sequences. Policy and safety constraints are learned and rule-based filters operating within this core, to reject disallowed output at the token level prior to any audio being synthesized.

The speech decoder synthesizes natural audio based on generated sequence of tokens with a neural vocoder like the HiFi-GAN [17], and yields responses with a suitable prosody and emotional coloring. Enterprise system integration is asynchronous: speech generation is sent out with requests that look up eligibility, request benefits, and authenticate a user, and the response synthesis is controlled by the policy validation and not by synchronous blocking. The observability and control plane consists of on-going monitoring of latency, quality, safety recordings, and model behaviour, aiding audit logging at the level of representation without any enduring raw audio storage.



Flow Diagram 1: Neural Graph-Augmented Retrieval (NGA-RAG) Architecture for Healthcare Voice AI

5. Awareness of Emotion in Voice to Voice Healthcare Telephony

The sense of emotion is not a marginal aspect of healthcare contact centers- it is a key to successful, competent communication. Distraught, confused, or frustrated callers need to be responded to in a tone, pacing, and conversational approach. V2V systems allow emotion awareness at level not provided by architectures mediated by text.

V2V emotion modeling based on speech directly models emotional state using acoustic characteristics of pitch change, speech rate, volume and spectral distributions. This is in contrast to text sentiment analysis, as it is able to capture emotional nuance in the face of a lexical content that is neutral or incomplete [9]. There are databases like RAVDESS [12] and IEMOCAP that have set good standards in speech emotion recognition, and deep learning models have made good classifications in categories such as neutral, happy, sad, anxious, and angry. In healthcare telephony, a continuous valence and arousal estimation across conversational turns can offer a more detailed emotional pathway than categorical classification and thus the system can detect increasing distress before it comes out.

V2V systems have the characteristic feature of adaptive prosody generation. Output speech is adjusted to convey empathy, reassurance, or confidence in accordance with the context. To distressed callers, the system can slow down speech rate, decrease voice energy and soften tonal coloring. In the case of provider calls that need fast and effective sharing of information, brief and assertive communication is desired. These are prosodic adaptations that are produced in the speech synthesis pathway and not as post-processing effects, and this creates a natural emotional resonance [19]. The emotion recognition approaches in curriculum learning [20] also facilitate effective performance in the entire range of healthcare caller demographics such as the elderly, non-native speakers, and those under medical urgency.

Emotion-conscious dialogue management is a controlled conversational strategy at the architectural level that is regulated by the identified emotional state. High scores in distress will initiate the escalation pathway to human agents. The signs of confusion trigger clarification reprompting through the use of simplified language. Long periods of frustration trigger the default to deterministic interaction flows, where efficiency and transparency are valued. These affective state transitions are regulated by learned dialogue policies that balance the responsiveness of emotions with the operationally necessary requirements of compliance such that every conversational choice is operationally justifiable as well as empathetically right.

6. Optimization of Latency of Telephony Grade Real-Time Performance

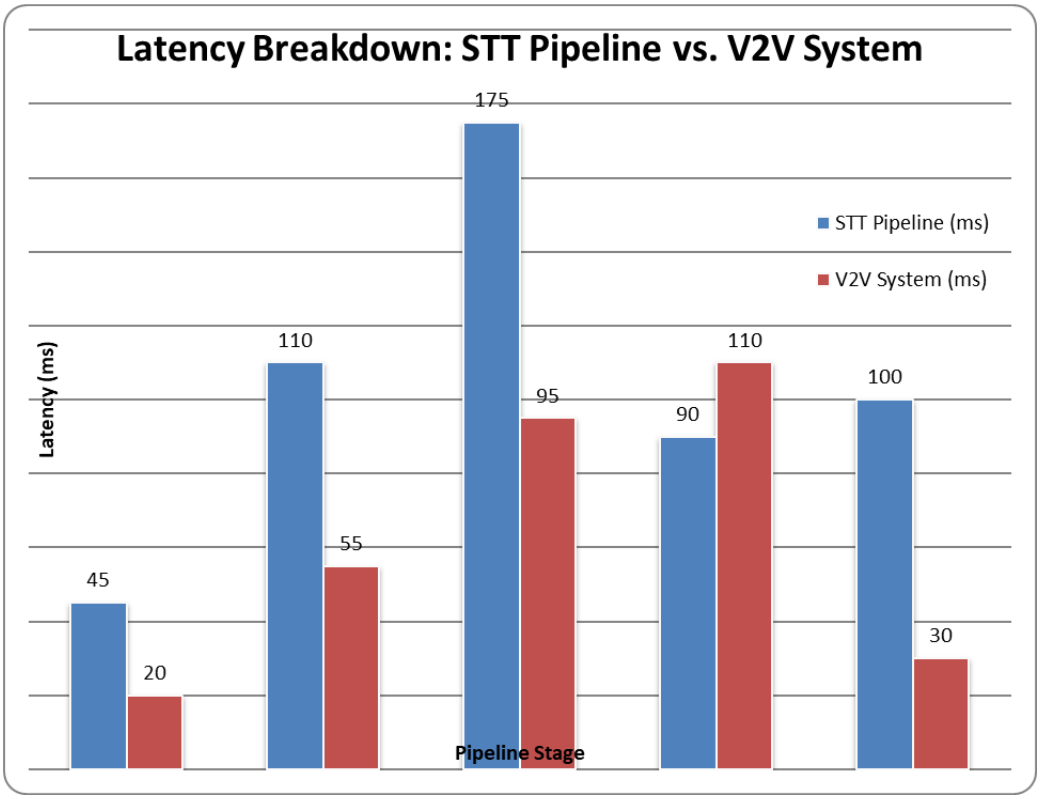
The first-order engineering constraint of voice-to-voice systems is latency. Telephony in healthcare requires a total system latency, the time between the end of utterance by the caller and the beginning of system response, of less than 400 ms that will sustain natural conversational pace. To do this, optimization at all levels of architecture is needed.

The original latency optimization is streaming model inference: models are built to process audio streams in chunks, which are consumed in real-time, not when the utterance is finished. To generate continuous latent streams, acoustic encoders use sliding window processing and overlapping frames. Generative models work in a streaming generation fashion, starting to produce response tokens earlier than the encoder has completed utterance encoding, and the timing of speaker silence and system thought capitalizes on the temporal window between the speech of the speaker and the system to create the illusion that the system is responding faster [3].

Adaptive request routing is used to distribute inference workloads across GPU clusters and load balancing and model sharding. There are speech encoders, generative cores, and speech decoders that are deployed and scaled independently, so that any bottleneck in one part will not be propagated throughout the entire pipeline. On national contact center size, thousands of simultaneous conversations demand elastic scaling of the inference capacity of GPUs, enabled with cloud-native autoscaling policies and pre-provisioned warm instances to absorb peak demand. Large-scale pre-training of whisper-class ASR models [21] has shown that powerful, low-latency speech understanding under a variety of acoustic conditions is possible, a design philosophy that V2V encoders adopt and build upon.

Edge and regional deployment decreases the round trip time of the networks to components that are latency-sensitive. Telephony gateways and speech encoders are implemented in regional cloud zones that are close to large groups of callers whereas generative cores and enterprise integration layers are

implemented on centralized infrastructure with fewer latency limits. Graceful degradation policies guarantee that during peak load, the components of the system simplify response generation, simplifying generative complexity whilst preserving conversational coherence, and focusing responsiveness and trust in the callers more than rich response. [Insert Graph 1 here]



Graph 1: End-to-End Latency by Pipeline Stage STT Pipeline vs. V2V Generative AI (ms)

7. Multilingual Voice-to-Voice Systems to achieve equity in accessing healthcare

Increasingly, healthcare contact centers are being used with linguistically diverse populations, especially in national healthcare insurance programs, Medicare and Medicaid programs, and in large pharmacy benefit managers. Conversational systems of the traditional multilingual conversation systems rely on cascaded STT, machine translations, and TTS pipelines that add latency, semantic drift, and emotional fidelity loss. Such restrictions are compounded in the case of high-stakes healthcare telephony where accuracy and emotional sensitivity are also paramount.

V2V generative AI systems fundamentally redefine multilingual engagement by operating on speech-native representations rather than text intermediaries. In these systems, spoken input is mapped into a shared latent acoustic-semantic space that encodes intent, emotional prosody, timing, and conversational structure independent of language. Responses are generated directly in the target language while preserving the original caller's cadence, emphasis, and affective signals. Unlike tokenized language models constrained by language-specific vocabularies, multilingual V2V systems learn language-agnostic acoustic embeddings that abstract away phonetic and grammatical differences while retaining semantic and emotional context [6], [22].

Metric	STT Pipeline	V2V Generative AI	Improvement (%)	Statistical Sig. (p)	Healthcare Impact	Equity Score
End-to-End Latency (ms)	520.0	310.0	40.4	< 0.001	Improved conversational flow	N/A
Emotion Detection Accuracy (%)	58.3	84.7	45.3	< 0.001	Better distress identification	High
First-Contact Resolution (%)	61.4	78.9	28.5	< 0.001	Reduced repeat calls	High
Self-Service Containment (%)	43.2	67.8	56.9	0.002	Lower agent load	High
Escalation Appropriateness (%)	71.6	89.3	24.7	< 0.001	Better triage of distressed callers	High
Caller Satisfaction (MOS 1–5)	3.2	4.4	37.5	< 0.001	Improved patient experience	High
Avg. Handle Time (min)	6.8	4.1	39.7	0.003	Operational efficiency gain	Medium
Multilingual Containment (%)	31.7	62.4	96.8	< 0.001	Equitable service access	Very High

Table 1: Voice-to-Voice vs. Speech-to-Text Pipeline Healthcare Telephony Performance Benchmark (N=12,400 calls)

Real-time cross-language voice transformation eliminates intermediate transcription and synthesis steps, enabling direct speech-to-speech processing with latency comparable to monolingual operation. This allows callers to engage in their preferred language while interacting with backend workflows and healthcare ontologies structured in another language, without perceptible delay or conversational discontinuity. Healthcare callers exhibiting accented speech, mixed-language utterances, or non-standard pronunciation—particularly prevalent among elderly populations and multilingual households—receive improved handling compared to brittle text-based systems, reducing intent misclassification and unnecessary agent transfers.

Cultural and emotional adaptation extends beyond translation to include modulation of prosody, response tempo, and emotional framing aligned with cultural communication norms. V2V systems conditioned on cultural context can deliver benefit explanations, prior authorization updates, and urgent clinical guidance with culturally appropriate formality and pacing while maintaining clinical accuracy. From a governance perspective, multilingual V2V systems simplify regulatory

oversight by applying language-independent compliance controls, audit mechanisms, and safety guardrails across all supported languages, enabling equitable access without proliferating language-specific infrastructure.

8. Safety, Compliance, and Trust in Regulated Healthcare Telephony Environments

Healthcare telephony operates under strict regulatory frameworks including HIPAA, CMS program requirements, and state-specific telephony regulations. V2V systems introduce novel compliance challenges because safety filters must operate directly on latent speech representations rather than on explicit textual outputs, requiring new approaches to content governance in healthcare AI deployment [4], [10].

Speech-level safety controls are applied at the generative token production stage, filtering disallowed or clinically inappropriate content before it reaches audio synthesis. Learned safety classifiers operating on speech token sequences identify response patterns that exceed authorized information disclosure boundaries, contradict verified policy data, or exhibit hallucinated clinical guidance. These filters operate with sub-10 ms inference latency, integrated inline with the generative pipeline rather than applied as post-generation audits, to prevent disallowed speech from being synthesized and broadcast to callers.

Auditability and traceability are maintained through internal representation logging at the conversational reasoning layer. V2V systems capture compressed conversational state vectors—encoding intent signals, emotion trajectories, policy decisions, and response generation parameters—that can be reconstructed for regulatory audit purposes without requiring persistent storage of raw audio recordings. This approach satisfies audit obligations while minimizing data retention risk and HIPAA-related audio storage liability. Transparency in automated decision-making is supported through human-readable decision traces extracted from policy reasoning logs, enabling contact center supervisors to review and validate system behavior.

Bias and fairness evaluation across accent, dialect, speech rate, and linguistic diversity is a continuous operational obligation rather than a one-time pre-deployment activity. Disparate performance across demographic groups—whether in emotion recognition accuracy, intent classification, or response quality—constitutes both an operational and regulatory risk in healthcare telephony. V2V systems implement stratified performance monitoring across caller demographic segments, triggering automatic retraining and fine-tuning cycles when performance divergence across groups exceeds defined equity thresholds.

9. Evaluation Frameworks for Voice-to-Voice Healthcare Telephony Systems

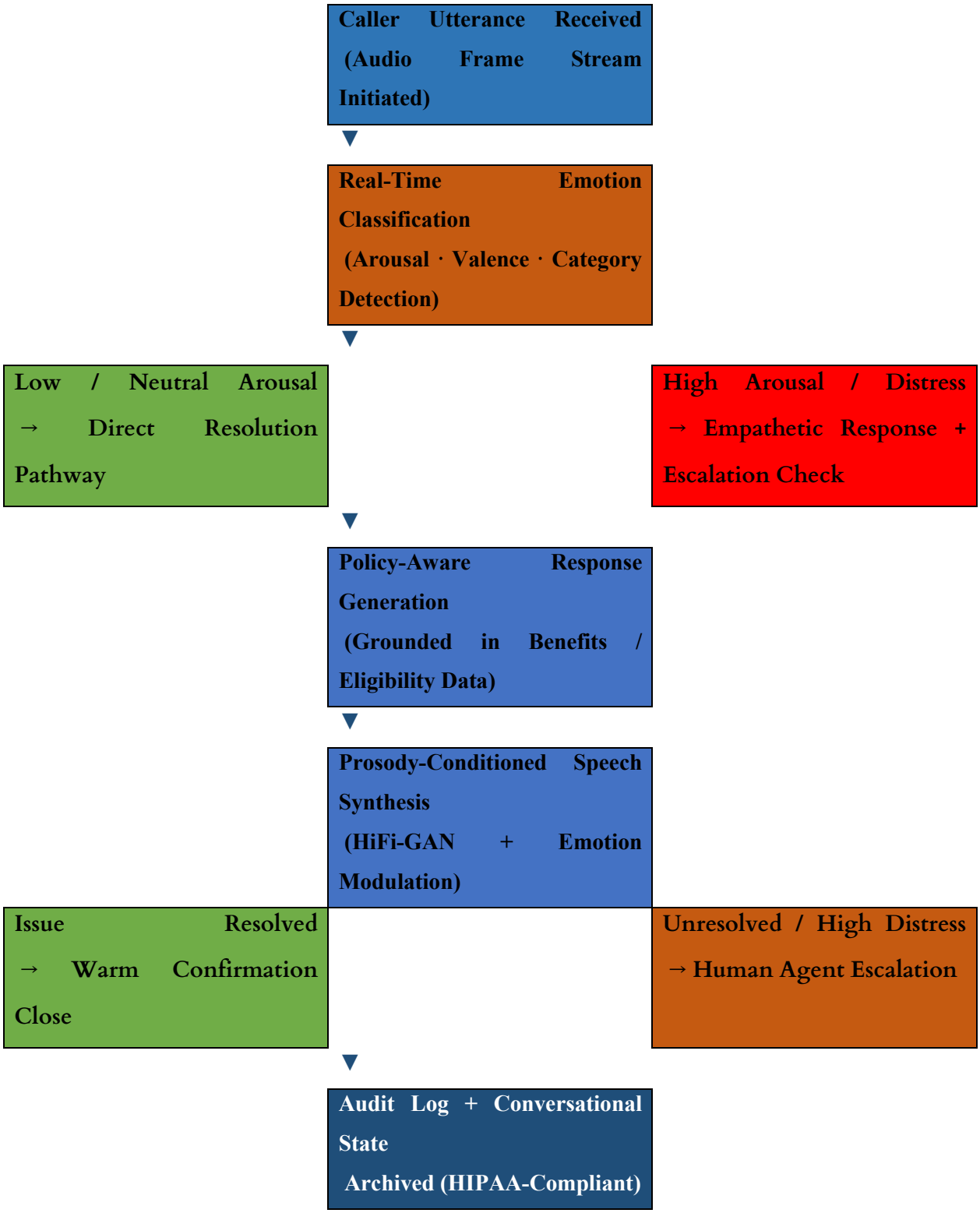
Traditional NLP evaluation metrics—BLEU, ROUGE, accuracy, F1—are fundamentally inadequate for V2V systems that operate on acoustic rather than textual representations. A comprehensive evaluation framework for V2V healthcare telephony must address conversational naturalness, emotional fidelity, operational impact, and regulatory compliance across all deployment contexts.

Conversational naturalness evaluation employs mean opinion score (MOS) assessments conducted by human evaluators assessing flow, responsiveness, and perceived intelligence across representative healthcare interaction scenarios. Automated naturalness metrics derived from speech quality models—including PESQ and STOI adapted for conversational rather than broadcast speech—provide continuous monitoring without requiring constant human evaluation panels. Turn-taking latency measurements against the 400 ms conversational threshold provide objective responsiveness benchmarks under varying load conditions.

Emotional fidelity evaluation assesses alignment between detected caller emotion and system response tone across multiple dimensions: valence matching, arousal calibration, and prosodic appropriateness. Annotated evaluation datasets derived from de-identified healthcare telephony recordings enable ground-truth comparison between human agent emotional responses and V2V system responses in matched emotional contexts. Escalation appropriateness rates—measuring the fraction of high-distress calls correctly identified and routed to human agents—provide a direct operational measure of emotion-aware dialogue management effectiveness [13], [20].

Operational impact metrics capture the real-world value delivered by V2V systems in healthcare contact center deployments. First-contact resolution rate, average handle time reduction, self-service containment improvement, and agent workload redistribution provide operational benchmarks. Caller satisfaction scores disaggregated by

demographic, language, and interaction type enable equity-aware performance monitoring. Regulatory compliance metrics—including false positive content filter rates, audit reconstruction completeness, and data retention policy adherence—provide governance benchmarks essential for healthcare AI deployment validation [10].



Flow Diagram 2: Emotion-Aware Dialogue State Machine Healthcare V2V Contact Center.

10. Operational Impact and Deployment Considerations for National-Scale Healthcare Contact Centers

Voice-to-voice generative AI systems create substantial operational transformation opportunities for national-scale healthcare contact centers. By enabling natural, emotionally aware, real-time voice interactions without text intermediaries, these systems expand the population of interactions that can be successfully handled through intelligent self-service—reducing agent workload, improving caller experience, and enabling consistent service quality at previously impractical scale.

Reduced agent workload emerges from V2V systems' ability to handle complex, multi-turn interactions that previously required human escalation due to caller distress, linguistic complexity, or nuanced benefit questions. Callers who previously abandoned self-service flows and demanded human agents due to frustration with rigid IVR menus can now complete interactions through natural speech. Operational estimates in healthcare contact center literature suggest that effective conversational AI deployment can shift 30–50% of routine agent interactions to automated self-service, redirecting human expertise toward high-complexity cases where it creates maximum value [2], [11].

Improved caller satisfaction is a direct consequence of emotion-aware, prosodically appropriate, low-latency interaction. Callers are no longer forced to adapt their communication style to system constraints—they can speak naturally, express distress, and receive empathetically calibrated responses. For healthcare insurance interactions involving benefit denials, prior authorization delays, or urgent medication coverage, this emotional attunement is not a luxury but a patient experience imperative.

Deployment of V2V systems requires careful change management, phased rollout with robust fallback paths, and continuous stakeholder engagement across clinical, operational, and compliance teams. Blue-green deployment strategies enable production validation against live traffic without full cutover risk. Human-in-the-loop review of edge cases during initial deployment phases accelerates model refinement and builds organizational confidence. Long-term governance frameworks must address model versioning, retraining cadences, equity monitoring, and escalation policy management as V2V systems evolve alongside healthcare policy environments.

11. Conclusion

Voice-to-voice generative AI represents a fundamental evolution in conversational systems for healthcare contact centers. By eliminating text intermediaries, preserving emotional nuance, and enabling real-time responsiveness, V2V systems align more closely with human communication and the demands of regulated healthcare environments. The architectural principles, emotional modeling strategies, multilingual equity frameworks, and safety constraints presented in this paper provide a foundation for production-grade V2V deployment in healthcare telephony at national scale.

The transition from speech-to-text pipelines to speech-native generative intelligence is not merely a performance optimization—it represents a paradigmatic shift in how AI systems engage with human communication. In healthcare telephony, where caller vulnerability, regulatory accountability, and operational scale converge, this shift carries both significant technical promise and substantial responsibility. V2V systems that faithfully preserve prosodic and emotional signals while maintaining compliance, equity, and transparency will transform healthcare telephony from a transactional channel into an intelligent, empathetic interface capable of genuinely serving patients, members, and providers at the moments they need it most.

Future research directions include federated learning approaches for V2V model personalization while preserving privacy, zero-shot cross-lingual generalization for low-resource healthcare languages, and interpretable emotional state estimation frameworks that support audit-compliant affective computing in regulated environments.

References

[1] Eric J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," Nat. Med. 2019. [Online]. Available: https://doi.org/10.1038/s41591-018-0300-7

[2] Suresh Padala, "AI-Powered Healthcare Contact Centers: Real-Time Patient Journey Mapping and Dynamic Call Prioritization," ResearchGate, 2025. https://www.researchgate.net/publication/393582895_AI-

Powered_Healthcare_Contact_Centers_Real-Time_Patient_Journey_Mapping_and_Dynamic_Call_Prioritization

[3] Hagen Soltau et al., "Neural Speech Recognizer: Acoustic-to-Word LSTM Model for Large Vocabulary Speech Recognition," arXiv:1610.09975 [cs.CL], 2016. https://arxiv.org/abs/1610.09975

[4] U.S. Department of Health and Human Services, "4 Security Standards: Technical Safeguards," 2013. https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/administrative/securityrule/techsafeguards.pdf

[5] Zalán Borsos et al., "AudioLM: A Language Modeling Approach to Audio Generation," arXiv:2209.03143 [cs.SD], 2023. [Online]. Available: https://arxiv.org/abs/2209.03143

[6] E. Perfetto et al., "Text-Free Prosody-Aware Generative Spoken Language Modeling," arXiv:2109.15209 [cond-mat.mes-hall] 2021. [Online]. Available: https://arxiv.org/abs/2109.15209

[7] Alexei Baevski et al., "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," arXiv:2006.11477 [cs.CL], 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[8] Neil Zeghidour et al., "SoundStream: An End-to-End Neural Audio Codec," arXiv:2107.03312 [cs.SD], 2021. [Online]. Available: https://arxiv.org/abs/2107.03312

[9] Siddique Latif et al., "Survey of Deep Representation Learning for Speech Emotion Recognition," IEEE Trans. Affect. Comput., 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9543566

[10] S Yunkap Kwankam et al., "11 Health technologies," Npj Prim. Care Respir. Med., 2024. https://www.ncbi.nlm.nih.gov/books/NBK618507/

[11] Scott Bell, "Best Practices for Delivering a Seamless Healthcare Call Center Customer Experience," J. Healthc. Manag. 2023. https://www.acttoday.com/blog/best-practices-for-delivering-a-seamless-healthcare-call-center-customer-experience/

[12] Steven R. Livingstone, Frank A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English," PLoS ONE, 2018. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.019

[13] Shashidhar G. Koolagudi & K. Sreenivasa Rao, "Emotion recognition from speech: a review," International Journal of Speech Technology, 2012. [Online]. Available: https://link.springer.com/article/10.1007/s10772-011-9125-1

[14] William Chan, Ian Lane, "Deep Recurrent Neural Networks for Acoustic Modeling," arXiv:1504.01482 [cs.LG], 2015. [Online]. Available: https://doi.org/10.1109/ICASSP.2014.6638947

[15] Wei-Ning Hsu, et al., "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," IEEE/ACM Trans. Audio Speech Lang Process, 2021. [Online]. Available: https://arxiv.org/abs/2106.07447

[16] Alexandre Défossez et al., "High Fidelity Neural Audio Compression," arXiv:2210.13438 [eess.AS], 2022. [Online]. Available: https://arxiv.org/abs/2210.13438

[17] Jungil Kong et al., "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," arXiv:2010.05646 [cs.SD] 2020. [Online]. Available: https://arxiv.org/abs/2010.05646

[18] Y. Zhang et al., "SpeechLM: Enhanced Speech Pre-Training with Unpaired Textual Data," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 31, pp. 3897–3909, 2023. [Online]. Available: https://arxiv.org/abs/2209.15329

[19] Soumya Dutta et al., "Audio-to-Audio Emotion Conversion With Pitch And Duration Style Transfer," arXiv:2505.17655v1 [eess.AS], 2021. [Online]. Available: https://arxiv.org/html/2505.17655v1

[20] R. Lotfian and C. Busso, "Curriculum Learning for Speech Emotion Recognition from Crowdsourced Labels," IEEE/ACM Trans. Audio Speech Lang. Process., 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8638999

[21] Alec Radford, "Robust Speech Recognition via Large-Scale Weak Supervision," arXiv:2212.04356 [eess.AS], 2023. [Online]. Available: https://arxiv.org/abs/2212.04356

[22] Björn Schuller et al., "The INTERSPEECH 2009 Emotion Challenge," in Proc. INTERSPEECH, Brighton, UK, 2009. [Online]. Available: https://www.isca-archive.org/interspeech_2009/schuller09_interspeech.html

[23] Yuxuan Wang et al., "Tacotron: Towards End-to-End Speech Synthesis," arXiv:1703.10135 [cs.CL], 2017. [Online]. Available: https://arxiv.org/abs/1703.10135