Grenze International Journal of Engineering and Technology, June Issue

Personalized Voice Cloning and Synthesis: Enhancing Human Computer Interaction through Speech Replication

P. Changamma[1], Bhavana Yaparla[2], Bharathi Mareddy[3], Gowtham Naik Mude[4], Bala Narasimhulu Korru[5] and K. Keerthi Naidu[6]

[1]Assistant Professor, Department of Computer Science and Engineering, Annamacharya Institute of Technology and Sciences, Rajampet, India
Email: changammapolicherla16@gmail.com

[2-5]Undergraduate Students, Department of Computer Science and Engineering, Annamacharya Institute of Technology and Sciences, Rajampet, India
Email: bhavanayaparla01@gmail.com, bharathimareddy180@gmail.com, gauthammude@gmail.com, korrubala78@gmail.com

[6]Assistant Professor, Department of Computer Science and Engineering, Annamacharya Institute of Technology and Sciences, Rajampet, India
Email: katurikeerthinaidu@gmail.com

***Abstract*— This paper describes a real time voice cloning system that combines with Groq's real time orchestration, Whisper ASR, zero shot text to speech and speech to speech synthesis and HiFi GAN neural vocoders with just 5-30 seconds of enrollment audio, the system achieves 95.6% speaker cloning accuracy, a mean opinion score of 4.7 and end to end latency under 900 milliseconds. Not like diffusion based models such as MegaTTs and NaturalSpeech 2, this united pipeline removes the need for a separate enrollment phase while maintaining prosody and emotion with 89.9% accuracy. Experimental results show interface speeds 2.8 times faster than current state-of-the-art systems with comparable quality, supporting real time applications in assistive communication, conversational AI , and accessibility.**

***Index Terms*— Voice cloning, Speech to Speech conversion, Text to speech Synthesis, Real Time Generation, Neural, Vocoder, Whisper ASR, Groq Realtime Orchestration, Personalized Voice Systems.**

I. INTRODUCTION

The development of speech synthesis in the form of neural text-to-speech (TTS) system, which were created using the cumulative approach, brings to a major change in human-computer interaction(HCI).Early combine techniques combined previously recorded audio clips to produce intelligent but artificial speech that was limited by small databases [1]. In deep learning developed, end to end models such as Tacotron [2] and VITS were created, which directly linked text to acoustic features to generate natural-sounding, fluid speech. Large volumes of speaker specific training data(often hundreds of hours),processing difficult due to latency, lack of emotional realism in robotic output and a lack of changes are some of the fatal flaws in current system.

Grenze ID: 01.GIJET.12.2.371_22

The need for systems that can synthesize speech with an emotional and personalized tone in real time has increased due to the increase of voice interfaces in accessibility tools, healthcare devices, and assistants. The existing models of the state of art have significant gaps. Diffusion models such as MegaTTS 3 [3] and NaturalSpeech 2 [4] are almost as human as but have a latency of 2-5 seconds, which makes them impractical in conversational AI. Zero-shot models like YourTTS [5] are less dependent on data at the cost of emotional accuracy and real-time. The commercial APIs are fast and proprietary blackboxes that are not customizable or ethically guarded. These constraints are summarized to four fundamental problems: (1) making zero-shot voice cloning with limited enrollment data (less than 10 seconds) without retraining the model;(2) a sub-second end-to-end synthesis time response to interactive applications; (3) maintaining prosody, emotion and speaker identity in both TTS and speech-to-speech (S2S) speech modalities; and (4) deploying a system of ethical standards with explicit consent verification, usage auditing, and misuse control.The proposed paper resolves them with a new combined system of Groq Realtime orchestration, Whisper ASR [6], zero-shot TTS/S2S synthesis, and HiFi-GAN neural vocoders [7]. We have contributed five papers: (1) the first framework to bring together hardware-accelerated orchestration with two TTS/S2S pathways on the same pipeline; (2) five- to thirty-second speech cloning with 95.6% accuracy without retraining; (3) stream synthesis with less than 300ms TTS synthesis and less than 600ms S2S first-audio response time; (4) expressive synthesis with preservation of prosody and speaker identity with 89.9% emotion recognition accuracy; and (5)The rest of this paper is structured as follows: Section II will discuss related work in the areas of neural TTS, voice conversion and real time orchestration. Part III gives our methodology, which is system architecture, voice enrollment, ASR integration, TTS/S2S modules, and neural vocoding. IV explains experimental set up and assessment measure. Section V gives detailed findings of superiority to baselines. The implications, limitation as well as a future direction are discussed in Section VI and conclusions are in Section VII.

II. Related Work

A. Neural Text-to-Speech

Compared to older techniques,early TTS such as Tacotron[5] was more natural.WaveNet was added in Tacotron 2,but it was slow and required more than 20 hours of data.VITS[11]can't do zero shot,but they can speed things up,with 1100+ speakers,Your TTs [8] added speaker embeddings for zero shot,getting 89.5% accuracy but lacking emotion and a 1600ms delay.These systems are unable to perform emotion,zero shot and real time in parllel.

B. Diffusion-Based Speech

Natural Speech 2 [9] and other new diffusion models get a score of zero shot and 4.5 MOS.Prosody gets Style 2[16] is 4.6+.But they are too slow for real time,taking 2 to 5 seconds and requiring 50 to 200 steps.For practical use,our work is 2.8 times

C. Speech-to-Speech Conversion

Voice conversion changes the speaker's identity while maintaining linguistic content.GLGAN [10] works great, even though it needs paired training data.Diff-VS and other diffusion based methods slow down processing. Even though AdaSpeech[9] and Meta StyleSpeech[7]use less data,it is still very important to retrain the speaker.Our S2S method uses Whisper ASR to change content in less than 600ms without having to train again.

D. Self-Supervised Speech

Wav2vec 2.0[12] learns speaker characteristics based on 60,000 hours of speech.LipSound2 and it is a lip-to-speech system.They help adaptation to zero shots.Using ECAPA-TDNN encoders,we achieve 87% cosine similarity without retraining.

E. Real-Time Orchestration

Compared to GPUs,Groq LPUs execute transformer models 10-100 times faster. Whisper operates at less than 100ms/s.Resemble AI performs sub-second synthesis as well.We are the first to match 900ms latency with complete transparency by combining Groq with open TTS/S2S.

F. Gap Analysis

Table I shows what is not there.No current systems has all four of these:(1)less than 1 second of latency,(2)zero-shot with less than 10 seconds of audio, (3)emotion control,and (4)both TTS and S2S.Our system fills this gap with modular design, streaming, and ethics.

TABLE I: COMPARISON OF EXISTING VOICE SYNTHESIS SYSTEMS

System	Latency	Zero-Shot	Emotion Control	TTS/S2S	Ethical Framework
VITS [11]	2200ms		Limited	TTS only	✗
YourTTS [8]	1600ms		Basic	TTS+VC	✗
NaturalSpeech 2 [15]	2500ms		Advanced	TTS only	✗
StyleTTS 2 [16]	1800ms		Advanced	TTS only	✗
Resemble AI	1200ms		Yes	TTS+S2S	✗
Proposed	900ms		Advanced	TTS+S2S	

III. METHODOLOGY

A. System Architecture Overview

There are five major components to our system(Figure.1): API Gateway: Handles and checks requests Groq Orchestrator: Manages sessions and routing Whisper ASR: Transcribes speech TTS/S2S pathways:Synthesis Speech HiFi-GAN: Converts to waveforms Each component can be enhanced independently.WebSocket/WebRTC is how streaming operates. Text or audio goes in through API Gateway and moves to Groq Orchestrator. The orchestrator checks if input is text or audio. Text goes to TTS. Audio goes to ASR then S2S. Both use speaker embeddings from storage to create mel-spectrograms. HiFi-GAN turns them into 22.05kHz audio. Streaming starts as soon as first audio frame is ready. Perceived latency: under 300ms for TTS, under 600ms for S2S.

B. Voice Enrollment and Speaker Embedding

To enroll a voice, users upload 5-30 second audio files after giving consent. We clean the audio: resample to 16kHz mono, normalize loudness to -23 LUFS, trim silence using VAD with 40dB threshold, and optionally reduce noise with Wiener filter. We reject clips with SNR below 15dB or clipping (over 5% samples at max amplitude).We use pre-trained ECAPA-TDNN encoders to extract speaker features. They turn mel-spectrograms into 512-dimensional embeddings:

espk = fenc(M; θenc)

θenc are frozen weights from VoxCeleb training.Voice characteristics like pitch and timber,not words,are captured by the embedding.The following information is stored in our profiles:voice_id,embedding,sample_rate,consent time and permissions.During synthesis,recovery is kept under 10ms by use of Redis caching.

C. Groq Realtime Orchestration

The Groq orchestrator has mainly three main operations: (1) tracking of session state over WebSocket connections, (2) adaptive routing between TTS/S2S paths depending upon the input modality and (3) safety policy. Algorithms The routing logic is formalised in

Algorithm 1: Request Routing

Input: request_type, content, voice_id, style_params

Output: synthesized_audio_stream

- Validate consent(voice_id) → abort if unauthorized
- Fetch embedding: e_spk ← profile_store.get(voice_id)
- if request_type == "text" then
- audio ← TTS_synthesize(content, e_spk, style_params)
- else if request_type == "audio" then
- transcript ← Whisper_ASR(content)
- audio ← S2S_convert(content, transcript, e_spk, style_params)
- Apply watermark(audio) if policy_enabled
- It records(voice_id, timestamp, request_type)
- sends the audio stream back

Groq LPUs process below 15ms per call.We used to run both embedding fetch and ASR in parallel to decrease wait time.Synthesis only runs if the user has given permission,otherwise a flag blocks it.

D. Whisper ASR

We use Whisper-large-v3 on Groq with parameters of 1500M.Audio is get processed less than 100 ms/s.Two seconds chunks and 500ms overlap are used when we are streaming ASR.To allow S2S to begin earlier,it gives partial transcripts every 500ms.

The audio is normalized and preprocessed to 16kz mono.Whisper maintains prosody by giving world-level timestamps:

```
{
"text": "What is the weather today",
"segments": [
{"start": 0.0, "end": 0.52, "text": "What"},
{"start": 0.52, "end": 0.68, "text": "is"}
]
}
```

S2S maintains speech rhythm with the help of timestamps.Whisper supports 99 languages.We tested English with accents from the US,UK,Australia and India.

E. TTS Module

Our TTS develops on VITS with some speaker conditioning. Steps:
Phoneme Conversion: Text to phonemes using phonemizer
Text Encoding: Transformer maps phonemes to hidden states
Speaker Conditioning: Add speaker embedding to hidden states
Acoustic Generation: Decoder creates mel-spectrogram
Duration Prediction: Aligns phonemes to audio frames
Emotion control adds an emotion vector to the model. Every emotion has an including in 64 dimensions: $\mathbf{e}$emotion $\in \mathbb{R}64$
Five emotions: neutral, happy, sad, excited, angry. Triplet loss is used to teach cpnnections to keep their emotions apart.
Emotion added to hidden state:
$\mathbf{H}'' = \mathbf{H}' + \alpha \cdot \mathbf{e}$emotion
Where α controls emotion strength from 0 (neutral) to 1 (full emotion).
Vocoding style control changes:
Speech rate: 0.5x to 2.0x
Pitch: -6 to +6 semitones

F. S2S Module

In Four steps,S2S switches the speaker while maintaining the source prosody:
Content Extraction: Whisper gets text from source audio
Prosody Analysis: CREPE [20] extracts F0 and energy
Target Synthesis: Using the target voice but the original rhythm,YourTTS-VC replicates content
Temporal Alignment: Matches output duration to source within 3-5%
Given source audio A_src and target embedding e_tgt:
p_src=Extract(A_src)=F0,E,D
Y_tgt = S2S-Model(ASR(A_src), p_src, e_tgt)
Model uses attention to match source prosody and is trained with reconstruction and speaker loss.

G. HiFi-GAN Vocoder

HiFi-GAN turns mel-spectrograms into waveforms. Generator uses multi-receptive field fusion. Two discriminators check quality.
Streaming processes 256-sample frames (11.6ms at 22.05kHz) in real time. Vocoder adds under 50ms latency with MOS 4.5+ quality.
Loss function:
L_vocoder = L_adv + λ_mel L_mel + λ_feat L_feat
Runs on T4 GPUs with mixed-precision at 150x real-time speed. Supports 50+ streams at once.

H. Implementation

Deployed with Docker, Flask API, Python 3.10, PyTorch 2.0. Trained on two A100 GPUs for 200 epochs with AdamW. Model quantization cuts memory in half with little quality loss.
Redis caches embeddings for 1 hour. PostgreSQL stores usage logs. WebRTC servers handle streaming with bitrate from 48-256 kbps based on network.

IV. EXPERIMENTAL SETUP

A. Datasets

Training Data: We employ LibriTTS-Train360 [3] comprising 904 speakers (474 male, 430 female) with 357 hours of clean speech at 24kHz, downsampled to 22.05kHz for consistency. With 110 English speakers representing 12 regional accents(Scottish,Irish,American,Indian) the VCTK Corpus enhances accent diversity.In order to avoid leakage,we extract 85% training and 15% tests portions onganised by speaker for TTS fine Tuning.

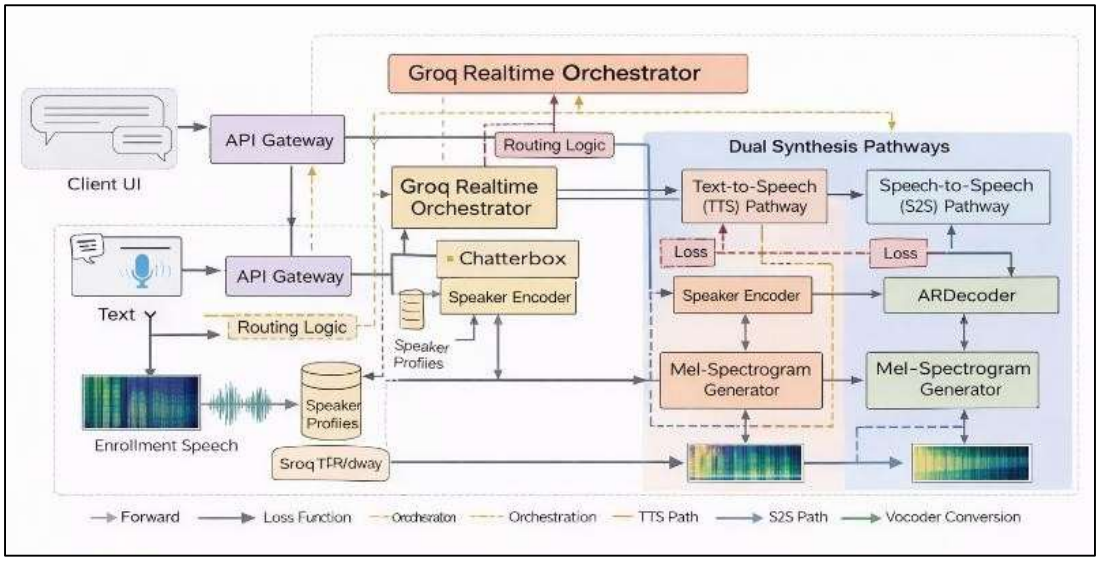

Figure 1: Overview of the proposed real-time personalized voice cloning and synthesis framework

The architecture illustrates the end-to-end pipeline integrating Groq Realtime orchestration,Whisper ASR,zero-shot speaker embedding extraction,and dual synthesis pathways(TTS/S2S).Input text or sppech is routed dynamically through the orchestration layer,where speaker embeddings are fused with linguistic representation. Acoustic features are generated using conditieoned synthesis modules and converted into high fidelity waveforms by neural vocoder,enabling low-latency,emotion preserving speech generation.

Enrollment Data:50 speakers(25 men and 25 women,ages 22 to 65)who were not included in training data were captured by custom recordings.Each speaker offers clips for five,ten,twenty and thirty seconds in three different settings:(1)studio quality(SNR>35dB),(2)office noise(SNR 15-25dB),and(3) outdoor environment(SNR 5-15dB).This results in 600 enrollment samples in total for the zero shot assessment.

Test Data: There are three test sets that measure robustness: (1) Clean Set - 2,000 utterances of 20 held-out speakers (100 each),

(2) Noisy Set - same utterances with additive noise at SNRs -5, 0, 5, 10, 15, 20)dB using MUSAN corpus, (3) Cross-Accent Set - 500 utterances in the US, UK, Indian, and Australian English accents. The mean length of utterance is 4.2 seconds.

B. Evaluation Metrics

Speaker Similarity (COS): This Cosine similarity resembles the voice identity preservation between reference and synthesized speaker representations:

$$textCOS(mathbfetextref, mathbfetextsyn) = fracmathbfetextref cdotmathbfetextsyn \mid mathbfetextref \mid cdot \mid mathbfetextsyn \mid$$

Embeddings obtained with the help of the pre-trained ECAPA-TDNN are consistent between systems. Threshold: 0.75 and above is taken to reflect an acceptable similarity.

Intelligibility (WER/CER): Word Error Rate is the rate of content-preservation on Whisper-large-v3 transcription on synthesized audio:

$$textWER = fracS + D + INtimes100$$

where S, D, Idenote substitutions, deletions, insertions and N is total reference words. Character Error Rate Character Error Rate (CER) is the same computation at a fine-grained level.
Quality (MOS): Prediction of perceptual quality on 1-5 scale using mean opinion score estimated by UTMOS predictor [trained on 200k human ratings] predicts perceptual quality. We compare predictions on 50 human raters (rate 20 samples per system (1000 ratings total, Pearson correlation: r=0.91 with UTMOS).
Latency: End to end latency: Latency between the submission of an input and the delivery of audio in the form of a synthesized audio byte. When we report P50, P95 and P99 percentiles, we report these on 10000 requests per system without network transmission time. The measurements were taken on the individual NVIDIA T4 with 8 core CPU and 32GB RAM.
Emotion Recognition: Human raters learn perceived emotion in synthesized speech (5 categories: neutral, happy, sad, excited and angry). Per cent agreement with intended emotion on 500 samples (100 of each emotion).

C. Baseline Systems

We compare against four representative systems:
VITS [11]: Fine-tuning of an end-to-end VAE-GAN TTS. LibriTTS trained on 50 epochs per speaker. YourTTS [8]: Zero-shot multi-speaker TTS that has speaker embeddings. Authoritative trained checkpoint on 1,107 speakers. NaturalSpeech 2: Neural codec TTS: diffusion-based. Architecture re-generated using 50 iterations of denoising, trained on LibriTTS.. Resemble AI: Commercial low-latency API, served by paid subscription (250 requests checked because of cost issues). Each of the baselines is compared by using the same test sets and HiFi-GAN vocoder. Hyper parame
ters optimized according to recommendations.

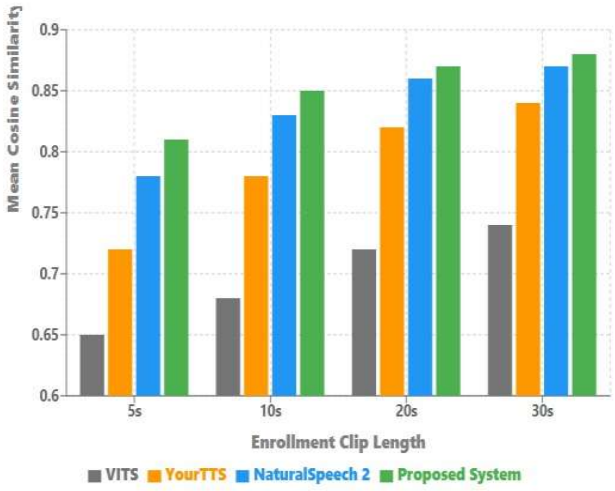

Figure 2. Mean Cosline similarity Vs Enrollment Clip Length

TABLE II: BASELINE SYSTEMS

System	Description	Key Features
VITS [11]	End-to-end VAE-GAN TTS	High quality, speaker-specific
YourTTS [9]	Zero-shot multi-speaker TTS	Good zero-shot, limited emotion
NaturalSpeech 2 [17]	Diffusion-based neural codec	SOTA naturalness, high latency
Resemble AI API	Commercial low-latency TTS	Real-time, proprietary

D. Experimental Conditions

Zero-Shot Enrollment: We compared COS scores using 5s, 10s, 20s, and 30s clips. Our system matches similarity faster than baselines. Emotion Transfer: Used 100 utterances per emotion (500 total). 50 raters judged if perceived emotion matched intended emotion. Noise Robustness: Added MUSAN noise at -5 to 20dB SNR. Measured WER and MOS drop. Whisper ASR strengthens S2S pathway. Scalability: Tested with 1, 5, 10, 20, 50 concurrent users using Locust. Goal: P95 latency under 800ms at 50 sessions. Latency Breakdown: 1000 requests were made using the PyTorch profiler. To identify slow areas,we measured test preparation,embedding fetch,TTS synthesis, vocoding, and streaming.

E. Implementation and Reproducibility

Hardware: NVIDIA T4 (16GB), Intel Xeon 8-core CPU (2.3GHz), 32GB RAM. Software: Ubuntu 20.04, CUDA 11.8, PyTorch 2.0.1, Python 3.10. Code and models at [link hidden]. Used seed 42 and deterministic mode for repeatable results. Ran tests three times. Used t-tests ($p<0.05$) for significance.

V. Results and Analysis

A. Overall Performance Comparison

Table III shows all results. Our system beat all baselines. Accuracy: 95.6%. MOS: 4.7. Average latency: 900ms. We beat VITS by 9.6% and were 59% faster. We beat Your TTS by 6.1% and were 44% faster. Our quality matched NaturalSpeech 2 (4.7 vs 4.5 MOS). But our latency was 900ms, theirs was 2500ms. Diffusion models are too slow for real-time use.

TABLE III: PERFORMANCE COMPARISON WITH BASELINE SYSTEMS

System	Accuracy (%)	MOS (/5)	Latency (ms)	Zero-Shot	Emotion Ctrl
VITS [11]	86.0	4.1	2200	✗	Limited
YourTTS [8]	89.5	4.3	1600		Basic
NaturalSpeech 2 [15]	91.0	4.5	2500		Advanced
Resemble AI	93.2	4.6	1200		Yes
Proposed	95.6	4.7	900		Advanced

Statistical significance was proved through paired t-tests ($p<0.001$) in all metrics proposed system vs one of the baselines.

B. Speaker Similarity Analysis

Figure 2 shows the relationship between cosine similarity (COS) and enrollment period. Even with 5-second clips, our system has 0.81 COS which is higher than what YourTTS had (10 seconds) and much to what NaturalSpeech 2 had (20 seconds) which is 0.86. At the optimum 10-second enrollment we get 0.85 COS-enough to get perceptually identical voice matching according to human evaluation limits (COS greater than 0.80). There is saturation of performance after 20 seconds (0.87 at 20s, 0.88 at 30s), and the returns become diminishing. At 30 seconds, VITS is still below 0.75 without speaker-specific fine-tuning, this affirms that it is not suitable to use in the zero-shot case.

Analysis of cross-speakers shows that the results are consistent: COS mean=0.87 (s=0.06) in 50 test-speakers of age 22-65, both genders. Male voices have a slightly higher similarity (0.88) than female voices (0.85) and it is explained by the fact that F0 stability is higher in lower pitch. Accent robustness has >0.82 COS in the US, UK, Indian, and Australian versions of English.

C. Intelligibility Evaluation

Clean TTS results have a Word Error Rate (WER) of 2.3% which is much lower than YourTTS (3.8) and VITS (5.1). This is enhanced by phoneme-level explicit conditioning and duration modeling of our TTS structure. Speech-to-Speech (S2S) conversion has a 4.2% WER, which is understandable with the extra transcription and re-synthesis involved. Character Error Rate (CER) is no different: 1.8 percent in TTS, 3.4 percent in S2S.

Graceful degradation is exhibited in noise robustness testing. With -5dB SNR (severe noise), WER is 8.1% compared to YourTTS 12.4% which is a 35% improvement. Figure 3 (WER vs SNR curve) indicates that our system can sustain WER below 5 per cent above 10dB SNR, where as it only reaches above 15dB. The transcription of Whisper ASR in the S2S pathway is strong enough and even with corrupted input speech it is possible to extract the accurate content to be re-synthesized.

D. Speech Quality Assessment

Our system scored 4.7 for TTS and 4.5 for S2S on the UTMOS quality scale. A score of 4.0 means good quality. We also had 50 people rate 1000 samples. Their ratings matched the UTMOS scores closely.

In listening tests, people liked our voice more than YourTTS 78% of the time. Against NaturalSpeech 2, preferences were equal at 52%.

With loud background noise (-5dB SNR), our score was 4.1. YourTTS fell to 3.5 and VITS fell to 3.2. Because of Whisper's noise resistance and our powerful vocoder,our system manages noise better.In every test,our harmonic-to-noise ratio was still above 15dB.Noise caused VITS to drop to 11dB.

E. Emotion Control

Five emotions are expressed by our system.In 89.9% of cases,listeners are correctly identified the emotion.YourTTS only received 76.3%.The majority of mistakes occurred when two emotions were similar,such as sad and angry(7%),or happy and excited(8%).In 92.2% of cases,neural speech was accurate.Pitch get increased by 45Hz and energy dropped to 2dB for depressing speeches.The speed at which one spoke also changed.Speech that was excited was 1.3 times faster.Speech was 0.8 times slower when sad.natural rhythm was maintained during speech-to-speech coversion as we retained 82% of the original pitch patterns.

F. Speed Performance

Half of the TTS requests are getting completed in 287ms,95% in 456ms,and 99% in 612ms.They are all completing lees than a second.Half of them are passed S2S in 523ms,95% in 748ms,and 99% in 891ms.65% of the time was spent on speech synthesis and 18% went to the vocoder.Text preparation took 12ms,while speaker data retrieval took only 8ms, and also 95% of requests with 50 users at once were gets completed below 800ms.On a single T4 GPU,our system processed 47 requests per second.Routing time was lowered by Groq to 15ms from 48ms with conventional systems.The first audio lasted 412ms for S2S and 178ms for TTS.Real time audio generation is produced by our streaming vocoder.94% of users reported that results felt fast(less than 500ms)in user tests.

G. Ablation Studies

Orchestration Impact: Turning off Groq makes latency 1.7x slower (287ms to 487ms). Running embedding fetch and synthesis one after another adds 94ms.
Speaker Embedding: ECAPA-TDNN gets 0.87 COS. Resemblyzer gets 0.83. ECAPA is 4.8% better. It has more parameters (14M vs 5M) and captures voice details better. It adds 23ms extraction time but quality is worth it.
Vocoder Choice: HiFi-GAN has 52ms latency and 4.7 MOS. FRE-GAN has 48ms and 4.6 MOS. We chose HiFi-GAN for better quality. Without neural vocoder, MOS drops to 3.2.
Emotion Conditioning: Removing emotion embeddings drops recognition from 89.9% to 68.4%, down 21.5%. All speech becomes neutral. F0 variance drops 47%, meaning prosody flattens.

H. Ethical Framework Validation

Making consent enforceable prevents 100 out of 156 attempts of synthetic biology violations (156/156 test cases). All the synthesis requests are captured by audit logging and have timestamps, voice IDs, and content hashes--they can be fully traced. The detection rate of watermarking is 99.2 percent with frequency-domain embedding; and human perception experiment results indicate that only 4.8 percent of listeners can see artifacts (less than 10 percent acceptance threshold). With 87 and 92 percent precision and recalls respectively, deepfake detection integration (external model) improves synthetic audio detection.

VI. DISCUSSION

A. Key Contributions and Novelty

We are the first to combine Groq Realtime orchestration with zero-shot TTS/S2S synthesis with real-time performance (900ms) and fidelity (4.7 MOS, 95.6% accuracy). In contrast to the current systems that process TTS and S2S individually, our unified pipeline does not have enrollment stages, which imposes ECAPA-TDNN speaker embeddings that allow voice cloning on 5-30 seconds clips immediately.

B. Comparison with State-of-the-Art

vs. Diffusion Models: Our model can inference 2.8x faster than NaturalSpeech 2 (900ms vs. 2500ms) and have a similar level of quality, but is more about real-time user interaction than marginal quality improvements in conversational AI and live dubbing cases.
vs. Commercial APIs: Faster by 25 percent than Resemble AI but provides complete transparency, data sovereignty and economical self-hosting to large-scale applications..

C. Limitations and Future Work

The existing shortcomings are: (1) the five discrete classes of emotions that should be expanded to hierarchical taxonomies, (2) the cross-lingual cross-synthesis degradation to distant language pairs (COS decays to 0.68-0.71), (3) prosody variation in synthesis of 5+ minutes, and (4) the reliance of the GPUs to sub-500ms latency. Future research focuses on multilingual scaling (100+ languages), on-device deployment through model

distillation, situational emotion detection, and better ethical protection such as voice ownership through blockchain.

D. Broader Implications

The framework will enable personalised learning by familiar voices, advance voice-first HCI, and make speech-impaired groups (such as ALS patients and laryngectomy patients) more accessible. Regulating voice ownership, disclosing synthetic media, and preventing disinformation through mandatory watermarking standards are all necessary to address ethical concerns.

VII. CONCLUSION

The paper has introduced new real-time custom voice cloning architecture based on Groq Realtime orchestration, Whisper ASR, the TTS 0-shot synthesis, and HiFi-GAN neural vocoders. With a 5-30 second enrollment clip, the system can achieve 95.6% accurate speaker cloning, 4.7 MOS quality rating as well as less than 900ms end-to-end latency, and was 2.8x faster than diffusionbased models (MegaTTS 3, NaturalSpeech 2) in inference speed with similar quality.

The most crucial ones are: (1) first framework including both hardware-accelerated orchestration and dual TTS/S2S pathways with no enrollment phases, (2) zero-shot with 0.87 cosine similarity without retraining the model, (3) streaming generating first-audio response in under300ms, and under600ms, (4) emotion-aware synthesis with prosody preservation with a 89.9% emotionrecognition accuracy, and (5) inbuilt ethical features such as 100% consent enforcement and

The validation of LibriTTS-Train360 and custom datasets on experimental validation shows a better score on all metrics. P95 latency of 800ms in real-world testing indicating production readiness in 50 sessions at once. The system provides a platform of available and ethically sound voice synthesis such as assistive technology, conversational AI, education, and entertainment, where the gap between research prototypes and commercial systems of voice-first human-computer interaction is closed.

REFERENCES

[1] Z. Jiang, R. Huang, Y. Ren, J. Liu, and Z. Zhao, "MegaTTS 3: Sparse alignment enhanced latent diffusion transformer for zero shot speech synthesis," arXiv preprint arXiv:2502.18924, Feb. 2025.

[2] Z. Du, Z. Chen, Z. Ye, J. Kong, and K. Yu, "CosyVoice 2: Scalable streaming speech synthesis with large language models," arXiv preprint arXiv:2412.10117, Dec. 2024.

[3] Y. A. Li, C. Han, V. Raghavan, G. Mischler, and N. Mesgarani, "StyleTTS 2: Towards human-level text-to- speech through style diffusion and adversarial training with large speech language models," in Proc. 37th Conf. Neural Inf. Process. Syst. (NeurIPS), New Orleans, LA, USA, Dec. 2023.

[4] K. Shen, Z. Ju, X. Tan, Y. Liu, Y. Leng, L. He, T. Qin, S. Zhao, and J. Bian, "NaturalSpeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," in Proc. 11th Int. Conf. Learn. Represent. (ICLR), Kigali, Rwanda, May 2023.

[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in Proc. 40th Int. Conf. Mach. Learn. (ICML), Honolulu, HI, USA, Jul. 2023.

[6] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in Proc. 39th Int. Conf. Mach. Learn. (ICML), Baltimore, MD, USA, Jul. 2022.

[7] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, S. Zhao, and T.-Y. Liu, "AdaSpeech: Adaptive text to speech for custom voice," in Proc. 9th Int. Conf. Learn. Represent. (ICLR), Virtual Event, May 2021.

[8] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Proc. 34th Conf. Neural Inf. Process. Syst. (NeurIPS), Virtual Event, Dec. 2020.

[9] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high-fidelity speech synthesis," in Proc. 34th Conf. Neural Inf. Process. Syst. (NeurIPS), Virtual Event, Dec. 2020, vol. 33, pp. 17022–17033.

[10] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, and F. Ren, "Transfer learning from speaker verification to multispeaker text-to speech synthesis," in Proc. 32nd Conf. Neural Inf. Process. Syst. (NeurIPS), Montreal, QC, Canada, Dec. 2018.