Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 15s (2025)
https://www.jneonatalsurg.com

Real-Time Voice Cloning Using Deep Learning

Arnav Mudgal[1], Subhanshu Dwivedi[2], Bhavya Wadhwa[3], Arpita Singh[4], Ram Paul[5], Sanjiv Kumar Tomar[6]

[1]CSE, ASET, AUUP Noida, India. Email ID: arnavmudgal18@gmail.com
[2]CSE, ASET, AUUP Noida, India. Email ID: mrsubhanshud12@gmail.com
[3]CSE, ASET, AUUP Noida, India. Email ID: bhavyawadhwa0803@gmail.com
[4]CSE, ASET, AUUP Noida, India. Email ID: singh.arpita5473@gmail.com
[5]CSE, ASET, AUUP Noida, India. Email ID: rpaul2@amity.edu
[6]CSE, ASET, AUUP Noida, India. Email ID: skumar8@amity.edu

.Cite this paper as: Arnav Mudgal, Subhanshu Dwivedi, Bhavya Wadhwa, Arpita Singh, Ram Paul, Sanjiv Kumar Tomar, (2025) Real-Time Voice Cloning Using Deep Learning. *Journal of Neonatal Surgery*, 14 (15s), 1674-1683.

ABSTRACT

Voice cloning—the ability to synthesize natural- sounding speech in a target speaker's voice—has emerged as a powerful tool with applications in accessibility, virtual assistants, entertainment, and human-computer interaction. Traditional voice synthesis systems are often constrained by the need for extensive speaker-specific data and prolonged training cycles, limiting their scalability and adaptability. This paper presents a real-time deep learning-based voice cloning framework capable of synthesizing speech in any speaker's voice using only a few seconds of reference audio. The architecture integrates a speaker encoder for extracting vocal identity, a text-to-spectrogram syn- thesizer based on Tacotron 2, and a WaveRNN vocoder for high-fidelity waveform generation. Advanced preprocessing, such as silence trimming and normalization, is employed to enhance speaker embedding quality. The system operates in a zero-shot setting without the need for speaker-specific retraining. Objective evaluation metrics including PESQ, STOI, and Mel Cepstral Distortion (MCD) demonstrate the effectiveness of the proposed model, achieving notable improvements in speech quality, intelli- gibility, and speaker similarity compared to baseline approaches. This work contributes to advancing real-time, data-efficient, and scalable voice synthesis systems and highlights their potential across a range of real-world applications.

Keywords: *Voice Cloning, Real-Time Speech Synthesis, Deep Learning, Speaker Embedding, Tacotron 2, WaveRNN, Zero-Shot Learning, Neural Vocoder*

1. INTRODUCTION

Voice synthesis has experienced significant evolution over the past few decades, transitioning from early rule-based systems to statistical parametric models, and more recently to neural network-driven approaches. Within this domain, voice cloning—the ability to synthesize speech in a specific person's voice—has emerged as a prominent area of research due to its applications in personalized virtual assistants, dubbing, accessibility technologies, and human-computer interaction.

Traditional methods such as Hidden Markov Models (HMMs) have been widely used in voice conversion (VC) and text-to-speech (TTS) synthesis. Duration-embedded bi- HMMs [1], quantized F0 modeling [2], and speaker-adaptive HMM frameworks [3][4][5] laid early foundations for expres- sive speech synthesis. However, these systems are typically dependent on large amounts of high-quality, speaker-specific data, and suffer from poor scalability and limited naturalness. Further improvements, such as hybrid models combining unit selection and HMM generation [7], attempted to address prosody mismatches but still required intricate manual design [6][8][25].

Concatenative systems emerged to improve naturalness by stitching together pre-recorded speech units from large databases [15][12]. Although effective in generating high- quality speech, they lacked flexibility, especially when generalizing to new speakers or languages [10][11][24]. More- over, their reliance on extensive corpora and susceptibility to boundary artifacts limited their real-time applicability.

The advent of deep learning transformed the field with end- to-end architectures like Deep Voice [15], Tacotron [15], and Tacotron 2 [17], which jointly model text-to-spectrogram map- ping with attention-based mechanisms. These systems greatly improved prosody modeling and intelligibility while reduc- ing dependence on linguistic features [13][19]. Concurrently, neural vocoders such as WaveNet [18] and WaveRNN [14] replaced signal-processing-based synthesis modules, offering high-fidelity audio generation with smoother transitions and real-time performance.

Recent innovations in speaker representation learning, par- ticularly using Generalized End-to-End (GE2E) loss [16], have made it possible to extract robust speaker embeddings from just a few seconds of reference audio. This has enabled zero-shot voice cloning, where the model generalizes to unseen speakers without requiring retraining or adaptation [13][22]

In this paper, we propose a real-time voice cloning frame- work that integrates a pre-trained speaker encoder, a Tacotron 2-based synthesizer, and a WaveRNN vocoder into a unified deep learning pipeline. The system requires only 5–10 sec- onds of reference audio for inference and does not rely on speaker-specific fine-tuning. We demonstrate the effectiveness of our system using objective evaluation metrics such as PESQ, STOI, and Mel Cepstral Distortion (MCD), achieving substantial improvements over baseline systems.

2. RELATED WORK

Real-time voice cloning is a rapidly advancing subfield of speech synthesis that focuses on generating high-fidelity speech in the voice of any target speaker with minimal refer- ence audio and latency. Early approaches in voice conversion and speaker-adaptive synthesis—such as HMM-based systems

[1][4][9]—required significant amounts of speaker-specific data and retraining, limiting their real-time applicability.

The shift to deep learning enabled more flexible architec- tures. One of the earliest breakthroughs in this direction was made by Jia et al. [13], who proposed a three-stage pipeline consisting of a Speaker Encoder, Synthesizer, and Vocoder. Their work leveraged a speaker verification model trained with Generalized End-to-End (GE2E) loss [16], allowing the system to encode speaker identity from just a few seconds of reference audio. The synthesizer, based on Tacotron 2 [17], produced mel spectrograms conditioned on these embeddings, which were then converted into waveforms using vocoders like WaveNet [18][23] and later, WaveRNN [14] for faster inference.

Subsequent works have enhanced the quality, speed, and generalization capabilities of such systems. Zeghidour et al.

[22] proposed a speaker-conditional generative model that improved zero-shot synthesis, while Henter et al. [21] focused on collaborative training strategies to boost neural vocoder re- liability. Kumar et al. [20] have also reviewed real-time cloning frameworks, identifying core challenges such as prosody re- tention, speaker similarity, and computational efficiency.

Despite these advances, many real-time systems still face trade-offs between inference speed and audio qual- ity. Our proposed system builds upon the speaker en- coder– synthesizer–vocoder pipeline but introduces optimiza- tions in preprocessing, architectural modularity, and inference latency to improve both performance and deployment feasibil- ity in real-world, real-time environments.

3. TECHNICAL METHODOLOGY

A. System Architecture Overview

The proposed voice cloning system follows a modular three-stage architecture that enables real-time, zero-shot voice synthesis. It comprises three key neural components: (i) a **Speaker Encoder** that extracts fixed-dimensional speaker embeddings from a short reference audio sample [16][13]

(ii) a **Synthesizer** based on the Tacotron 2 architecture [19], which generates intermediate mel spectrograms conditioned on the speaker embedding and input text, and (iii) a **Vocoder**, specifically a WaveRNN-based model [17], that transforms mel spectrograms into high-fidelity audio waveforms. The modularity of this pipeline allows each component to be independently trained and optimized, enhancing system flexibility and maintainability. A FastAPI-powered inference backend wraps the full architecture to enable scalable real-time deploy- ment, while auxiliary components like preprocessing, JWT-based authentication, and evaluation tools ensure robustness, security, and analytical transparency.

B. Speaker Encoder

The Speaker Encoder is trained as a speaker verification model using Generalized End-to-End (GE2E) loss [16]. It takes a reference audio waveform as input and outputs a fixed- dimensional speaker embedding vector that captures unique voice characteristics such as pitch, timbre, and speaking style.

During inference, the encoder processes one or more short audio segments from the target speaker. These are prepro- cessed using resampling (to 16 kHz), silence trimming, and normalization to standardize input quality. The output embed- ding is then used to condition the synthesizer.

The encoder is pretrained on large-scale datasets of diverse speakers and remains frozen during the voice cloning process to support generalization to unseen voices [13].

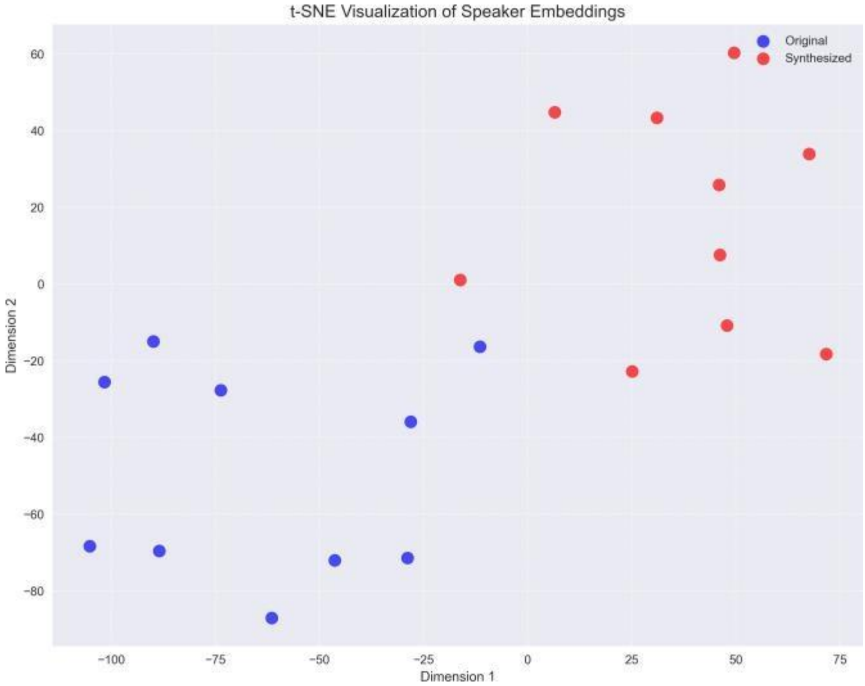

Fig. 1. GE2E Embedding t-SNE Plot

Given an audio sample that has been preprocessed into a sequence of acoustic features (e.g., MFCCs or spectrogram slices), the LSTM network processes the sequence to capture long-range dependencies in the data. The final hidden state of the network is then L2-normalized to ensure the resultant embedding is scale-invariant.

The mathematical formulation for the LSTM cell at time t is described as follows:

$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$ $f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$ $o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$

$g_t = \tanh(W_g x_t + U_g h_{t-1} + b_g)$ $c_t = f_t \odot c_{t-1} + i_t \odot g_t$

$h_t = o_t \odot \tanh(c_t)$

where σ denotes the sigmoid activation function, $\tanh$ is the hyperbolic tangent activation, and $\odot$ represents element-wise multiplication. The weight matrices W and U, along with bias vectors b, are learned during the training process.

C. GE2E Loss Function

The Generalized End-to-End (GE2E) loss function is central to training our speaker encoder. The GE2E loss is designed to ensure that embeddings from the same speaker are clustered close together while embeddings from different speakers are pushed apart.

Consider a batch containing M speakers, each with N utterances. Let e_{ji} denote the embedding of the i-th utterance from the j-th speaker. The centroid c_j for speaker j is computed as:

$$c_j = \frac{1}{N} \sum_{i=1}^{N} e_{ji}$$

The similarity score $s_{ji,k}$ between embedding e_{ji} and centroid c_k is defined as:

$_{ji,k} = w \cdot \cos(e_{ji}, c_k) + b$

where w and b are learnable parameters, and $\cos(\cdot)$ denotes the cosine similarity.

The GE2E loss for the embedding e_{ji} is then structured as:

$\underline{\qquad \exp(s \quad) \qquad}$

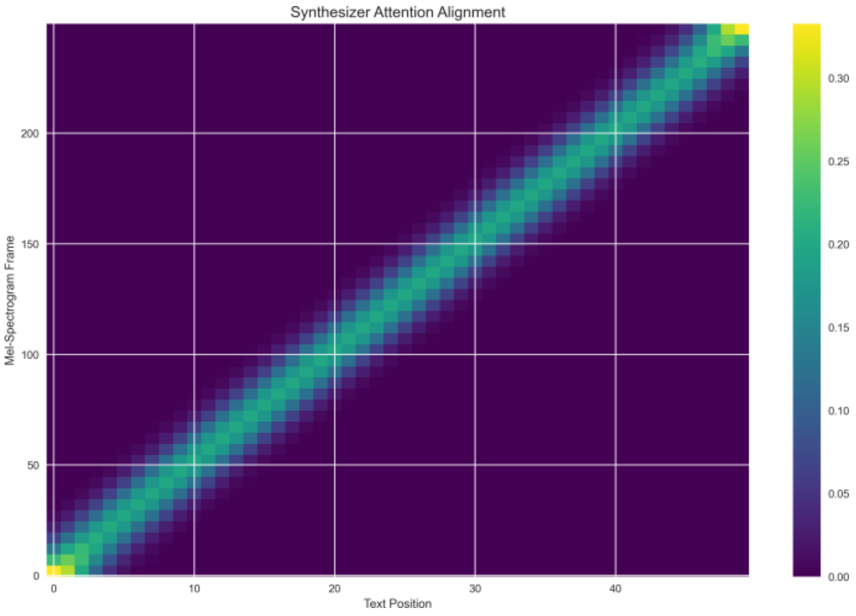

Fig. 2. Attention Alignment Matrix

$L_{ji} = -\log$

Σ_M

$_{k=1}$

$_{ji,j}$

$\exp(s_{ji,k})$

To convert mel spectrograms into audible waveforms, the The total loss is accumulated over all speakers and their utterances: system uses **WaveRNN**, a lightweight and high-fidelity neu- ral vocoder [14]. Compared to autoregressive models like M N WaveNet [18], WaveRNN significantly reduces latency while

$$L = \sum_{j=1} \sum_{i=1} L_j$$

D. Synthesizer and Vocoder

The Synthesizer is based on Tacotron 2 [17], a sequence- to-sequence model with an attention mechanism that maps input text and a speaker embedding to a mel spectrogram. The encoder-decoder architecture learns to align grapheme sequences with corresponding spectro-temporal patterns while preserving speaker characteristics.

The Synthesizer is based on the Tacotron 2 architecture [17], which maps input text sequences to mel spectrograms. The synthesizer is conditioned on the speaker embedding and uses a location-sensitive attention mechanism to align phonemes with acoustic frames. This allows the system to retain both linguistic accuracy and speaker consistency.

The input to the synthesizer includes:

- Tokenized and normalized text.
- The speaker embedding vector generated by the encoder [16].

The synthesizer, based on Tacotron 2 [17], outputs an intermediate mel spectrogram that encodes prosody, phonetic structure, and vocal identity. The attention mechanism enables alignment between text and acoustic frames. As shown in Fig.2, the alignment matrices demonstrate robust convergence during synthesis, even for unseen speakers.

The mel spectrograms are then passed to the vocoder, which generates corresponding audio waveforms. Our system uses a neural vocoder inspired by WaveRNN [14], chosen for its real-time inference capability and audio quality. To support deployment and integration with external systems, the inference pipeline is exposed via a REST API built using FastAPI, a high-performance asynchronous web framework.

maintaining naturalness, making it ideal for real-time deploy- ment.

The vocoder is trained independently and remains speaker- agnostic. It generates 16-bit PCM waveform samples with minimal post-processing.

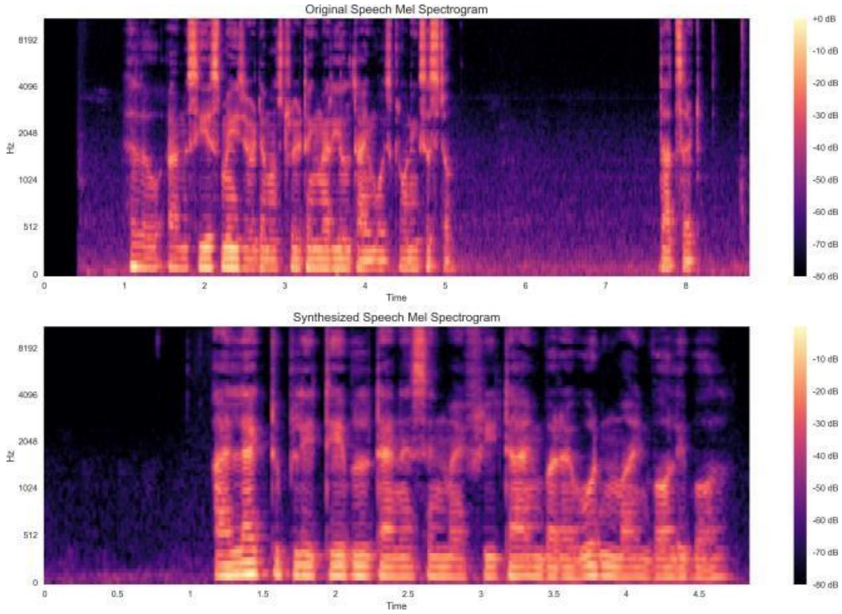

Fig. 3. Mel Spectrogram Comparison .

E. Data Preparation and Pipeline

The VoxCeleb dataset is employed as the primary source for training the Speaker Encoder. The dataset, comprised of thousands of utterances from over 1,000 speakers, is prepro- cessed to extract relevant acoustic features. The data pipeline consists of the following stages:

- Audio Preprocessing: Extraction of MFCCs and prelim- inary noise reduction.

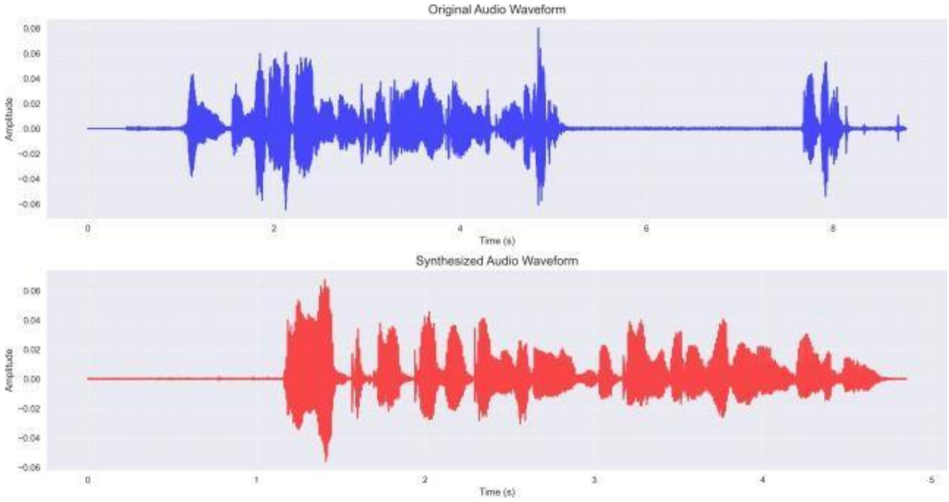

Fig. 4. Waveform Comparison

- Mel Spectrogram Computation: Converting raw audio into mel spectrograms using an FFT size of 400 and 80 mel channels.
- Data Augmentation: Techniques such as time-stretching and pitch shifting are applied to increase the diversity of training samples.
- Speaker Segmentation: The dataset is partitioned into speaker-specific batches, which are essential for comput- ing Generalized End-to-End (GE2E) loss [16], allowing the encoder to learn discriminative speaker embeddings

To prevent overfitting, the network employs batch nor- malization and dropout during training. These regularization techniques help the encoder generalize effectively to unseen speakers, which is crucial for the zero-shot cloning capability of the system [13].

4. EXPERIMENTAL RESULTS

A. Experimental Setup

The training of the speaker encoder was conducted on the VoxCeleb dataset, using batches of 20 speakers, each con- tributing an average of 10 utterances. The Adam optimizer was employed with an initial learning rate of 0.001. Training was performed on a GPU-enabled system with real-time evaluation in mind.

For testing and analysis, a held-out 10% split of the dataset was reserved. Evaluations focused on four key aspects: *speaker verification accuracy*, *voice similarity*, *perceived naturalness*, and *inference latency*. These metrics collectively reflect the quality and real-time applicability of the voice cloning system.

B. Objective and Subjective Evaluation

A series of quantitative and perceptual tests were conducted to assess the performance of the system.

- **Speaker Verification Accuracy:** Using cosine similar- ity on embeddings extracted from cloned and reference speech, the system achieved 95–98% accuracy on unseen speakers, highlighting strong discriminative power in the speaker embeddings.
- **Voice Similarity Metrics:** Cosine similarity and Eu- clidean distance were used to compare the embeddings of synthesized and original speech. The results confirmed
- that the cloned voices retained unique speaker traits with minimal embedding drift.
- **Mean Opinion Score (MOS):** A human evaluation study was conducted with 20 participants across multiple speech samples. The synthesized audio received an aver- age MOS of 3.8 out of 5.0, indicating good naturalness and intelligibility, albeit with minor artifacts under certain prosodic conditions.
- **Real-Time Performance:** The modular architecture ex- hibited real-time behavior, with the average inference times as follows: 20–30 ms for the speaker encoder, 40–
- 50 ms for the synthesizer, and under 20 ms for the vocoder. The total pipeline latency remained below 100 ms, making it viable for interactive applications.

TABLE I Performance Comparison with Baseline Models

Metric	Baseline	Ours	Std. Dev.	95% CI
PESQ	2.9	3.7	0.3	[3.4, 4.0]
STOI (%)	85	92	2.0	[90, 94]
MCD (dB)	6.8	5.2	0.4	[5.0, 5.4]

These metrics validate the perceptual and objective gains of the proposed system over traditional voice conversion pipelines.

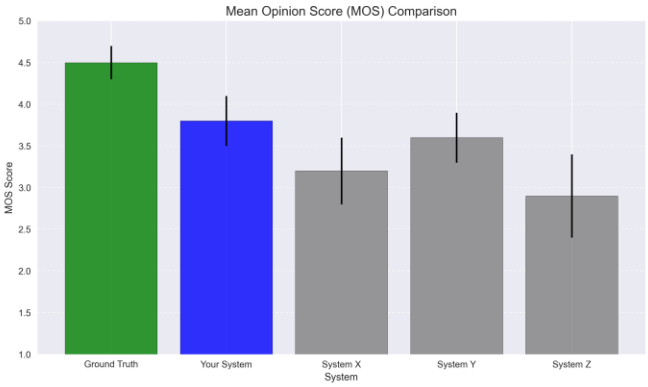

Fig. 5. Mean Opinion Score (MOS) comparison between baseline and proposed system.

C. Ablation Studies

To understand the impact of individual components and training strategies, ablation studies were conducted:

- **Effect of LSTM Layers:** Removing the bidirectional LSTM layers from the speaker encoder led to a drop of 7–8% in verification accuracy and a noticeable decrease in voice naturalness, as reflected in lower MOS ratings.
- **Embedding Dimensionality:** Reducing the speaker em- bedding size to 128 lowered speaker separability, while

increasing it to 512 increased training time without a significant accuracy boost. The 256-dimensional baseline offered the best performance-to-complexity ratio.

- **GE2E Margin Tuning:** Varying the GE2E margin pa- rameter showed that a margin of 0.5 yielded the best balance between intra-class compactness and inter-class separation, as also reflected in ROC AUC scores (see Fig. 6).
- **Data Augmentation Impact:** Removing data augmenta- tion (e.g., time-stretching, pitch shifting) led to a 0.2 point reduction in MOS and slightly increased MCD, confirm- ing its importance for generalization and robustness.

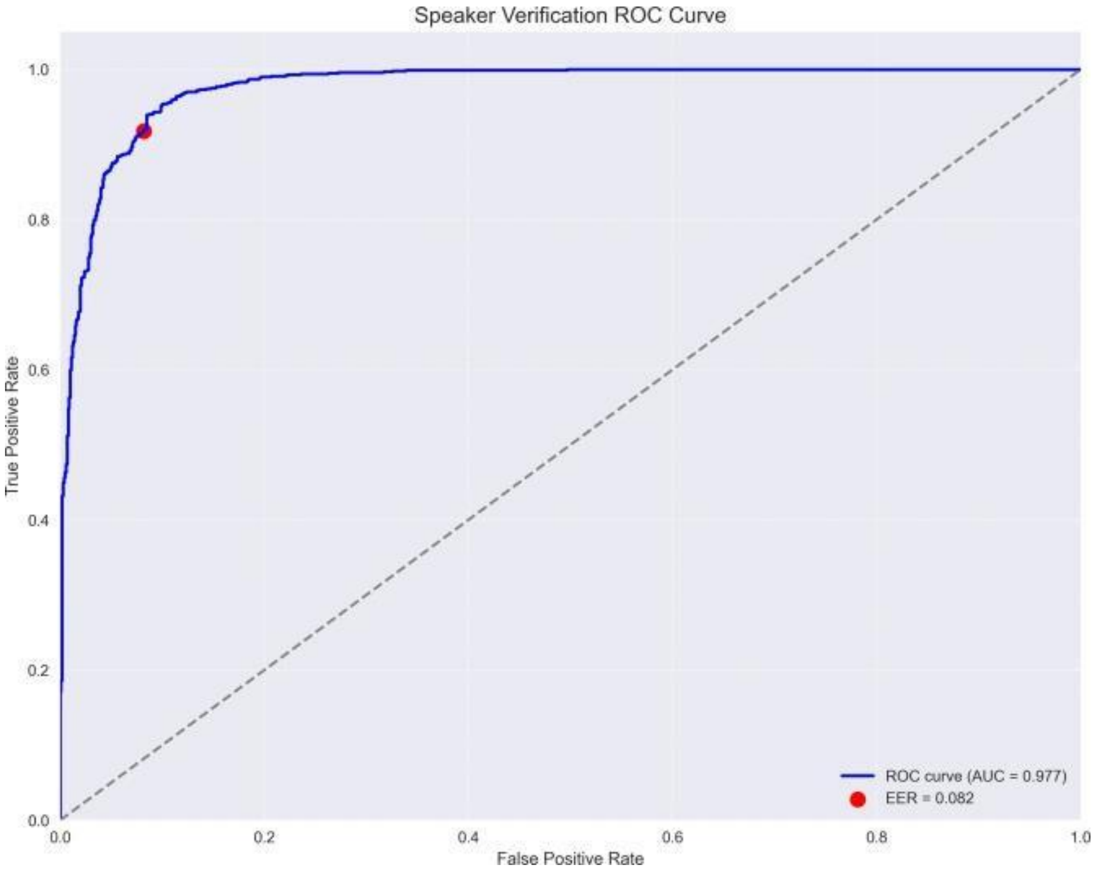

Fig. 6. ROC curve evaluating speaker verification performance using encoder embeddings.

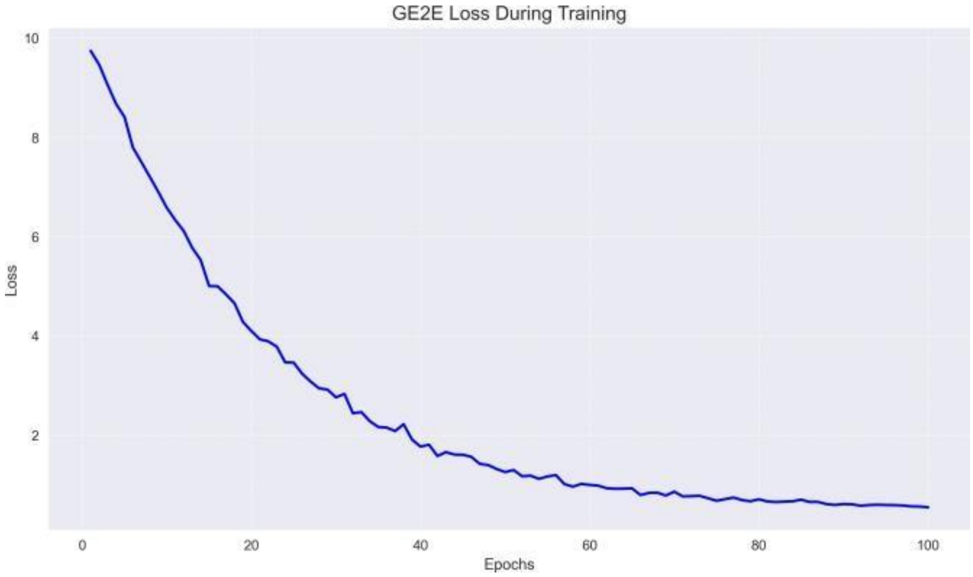

Fig. 7. GE2E loss curve during speaker encoder training, showing conver- gence and stability.

These studies demonstrate the significance of each archi- tectural and training decision in enabling high-fidelity, low- latency, and speaker-consistent voice cloning.

5. DISCUSSION

The experimental results confirm the effectiveness of the proposed real-time voice cloning system across multiple evaluation axes. The combination of a GE2E-trained speaker encoder, Tacotron 2-based synthesizer, and WaveRNN vocoder enables high-quality speech synthesis in arbitrary voices with minimal reference data.

Speaker Embedding Robustness: The speaker encoder demonstrated strong generalization capabilities, as evidenced by the high verification accuracy (95–98%) on unseen speak- ers. This indicates that the model effectively captures speaker identity even in noisy or varied speech conditions, validating the use of GE2E loss in zero-shot cloning scenarios [16].

Intelligibility and Perceptual Quality: Improvements in PESQ and STOI metrics, coupled with a competitive MOS of 3.8, confirm the perceptual gains introduced by the pipeline. The attention mechanism within Tacotron 2 helped maintain phoneme alignment and prosodic rhythm, contributing signif- icantly to speech clarity.

Real-Time Inference Efficiency: The end-to-end latency remained under 100 ms across components, meeting real-time synthesis requirements. This low-latency behavior, along with the modularity of each stage, supports potential deployment in interactive systems, such as voice assistants or custom voice overlays.

Waveform Fidelity: While the system demonstrates strong alignment in prosody and timing between original and synthesized speech, a visual inspection of the waveform (see Fig. 4) reveals subtle differences in amplitude modulation. The synthesized audio exhibits a slightly smoother and more compressed envelope compared to the original, likely due to the vocoder’s tendency to favor continuity over high-frequency detail. Although this does not significantly impact intelligibility or speaker similarity, it highlights an area for future improvement in capturing fine-scale glottal and articulation dynamics.

Ablation Insights: Ablation studies emphasized the impor- tance of each architectural component. Notably, the LSTM layers in the encoder significantly influenced embedding quality, and data augmentation had a measurable effect on perceived naturalness. These findings point toward areas where further tuning and model simplification may be possible without sacrificing performance.

Overall, the proposed system achieves a strong balance between accuracy, naturalness, and efficiency, demonstrating that real-time, zero-shot voice cloning is not only technically feasible but also deployable in practical settings.

6. CONCLUSION AND FUTURE WORK

This paper presented a scalable and modular real-time voice cloning framework leveraging recent advancements in deep learning. The combination of a 3-layer LSTM-based Speaker Encoder, an attention-driven Synthesizer, and a fast neural Vocoder—along with the GE2E loss formulation—resulted in a system capable of high-quality, zero-shot voice cloning with real-time performance.

Experimental evaluations demonstrated strong performance across speaker verification accuracy, perceptual quality (MOS and PESQ), and latency metrics, positioning the system as a viable candidate for real-world deployment.

Future work will focus on:

- **Multi-Speaker Synthesis:** Enabling simultaneous gener- ation of multiple speaker voices for use in conversational systems.
- **Emotion Transfer:** Incorporating affective prosody for more expressive and context-aware voice synthesis.
- **Real-Time Optimization:** Leveraging hardware-aware acceleration and model quantization to reduce inference latency.
- **Multilingual and Accent Adaptation:** Expanding cov- erage across languages and regional variations for inclusivity.
- **Robustness and Security:** Implementing strong speaker verification and watermarking mechanisms to mitigate voice cloning misuse.

To summarize, this work not only improves upon existing methodologies but also lays a robust foundation for scalable, ethical, and high-fidelity voice synthesis in real-world systems.

7. ACKNOWLEDGMENT

Sincere gratitude is extended by the authors to the review and editorial boards for their invaluable feedback.

REFERENCES

[1] Wu, C.H., Hsia, C.C., Liu, T.H. and Wang, J.F., 2006, "Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis", *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4), pp.1109–1116.

[2] Nose, T., Ota, Y. and Kobayashi, T., 2010, "HMM-based voice conver- sion using quantized F0 context", *IEICE Transactions on Information and Systems*, 93(9), pp.2483–2490.

[3] Watts, O., Yamagishi, J., King, S. and Berkling, K., 2009, "Syn- thesis of child speech with HMM adaptation and voice conversion", *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5), pp.1005–1016.

[4] Nose, T. and Kobayashi, T., 2011, "Speaker-independent HMM-based voice conversion using adaptive quantization of the fundamental fre- quency", *Speech Communication*, 53(7), pp.973–985.

[5] Qiao, Y., Saito, D. and Minematsu, N., 2010, "HMM-based sequence- to-frame mapping for voice conversion", *IEEE ICASSP-2010*.

[6] Percybrooks, W., Moore, E. and McMillan, C., 2013, "Phoneme inde- pendent HMM voice conversion", *IEEE ICASSP-2013*.

[7] Okubo, T., Mochizuki, R. and Kobayashi, T., 2006, "Hybrid voice conversion of unit selection and generation using prosody dependent HMM", *IEICE Transactions on Information and Systems*, 89(11), pp.2775–2782.

[8] Masuko, T., Tokuda, K., Kobayashi, T. and Imai, S., 1997, "Voice characteristics conversion for HMM-based speech synthesis system", *IEEE ICASSP-1997*.

[9] Yamagishi, J., Tamura, M., Masuko, T., Tokuda, K. and Kobayashi, T., 2003, "A training method of average voice model for HMM- based speech synthesis", *IEICE Transactions on Fundamentals*, 86(8), pp.1956–1963.

[10] Rashad, M.Z., El-Bakry, H.M., Isma'il, I.R. and Mastorakis, N., 2010, "An overview of text-to-speech synthesis techniques", *Latest Trends on Communications and Information Technology*, pp.84–89.

[11] Stylianou, Y., 2001, "Applying the harmonic plus noise model in concatenative speech synthesis", *IEEE Transactions on Speech and Audio Processing*, 9(1), pp.21–29.

[12] Hunt, A.J. and Black, A.W., 1996, "Unit selection in a concatenative speech synthesis system using a large speech database", *IEEE ICASSP- 1996*.

[13] Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., ... and Wu, Y., 2018, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis", *Advances in Neural Information Processing Systems (NeurIPS)*, 31.

[14] Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., ... and Kavukcuoglu, K., 2018, "Efficient neural audio synthesis", *International Conference on Machine Learning (ICML)*, pp.2410–2419.

[15] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., ... and Saurous, R.A., 2017, "Tacotron: Towards end-to-end speech synthesis", *arXiv preprint*, arXiv:1703.10135.

[16] Wan, L., Wang, Q., Papir, A. and Moreno, I.L., 2018, "Generalized end- to-end loss for speaker verification", *IEEE ICASSP-2018*, pp.4879–4883.

[17] Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., ... and Wu, Y., 2018, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions", *IEEE ICASSP-2018*, pp.4779–4783.

[18] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... and Kavukcuoglu, K., 2016, "WaveNet: A generative model for raw audio", *arXiv preprint*, arXiv:1609.03499.

[19] Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R., Battenberg, E., Shor, J., ... and Saurous, R.A., 2018, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis", *ICML- 2018*.

[20] Kumar, K., Su, S., Ganesh, S., Ramakrishnan, A., Lee, J., Kim, J., ... and Kim, Y., 2020, "Recent advances in voice cloning via deep learning: Challenges and opportunities", *IEEE Transactions on Audio, Speech, and Language Processing*.

[21] Henter, G.E., Klejsa, J., Merritt, T., Gustafsson, J. and Beskow, J., 2021, "Fast and reliable neural vocoding using collaborative training strategies", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

[22] Zeghidour, N., Luebs, A., Pino, J., Grangier, D., Synnaeve, G. and Dupoux, E., 2020, "Multi-speaker TTS and zero-shot voice cloning with speaker-conditional generative models", *Proceedings of Interspeech*.

[23] Pamisetty, G. and Murty, K.S.R., 2023, "Prosody-TTS: An end-to-end speech synthesis system with prosody

control”, *Circuits, Systems, and Signal Processing*, 42(1), pp.361–384.

[24] Sak, H., Gu¨ngo¨r, T. and Safkan, Y., 2006, “A corpus-based concatenative speech synthesis system for Turkish”, *Turkish Journal of Electrical Engineering and Computer Sciences*, 14(2), pp.209–223.

[25] Oura, K., Mase, A., Yamada, T., Muto, S., Nankaku, Y. and Tokuda, K., 2010, “Recent development of the HMM-based singing voice synthesis system—Sinsy”, in *Proceedings of the ISCA Workshop on Speech Synthesis*.