International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 3
May-Jun 2026, PP 01-8

Neural Codec Language Models for Unified Speech Generation and Transformation: A Review

Aboli Ashok Ugale, Associate Professor Vijay B. More
Department of Computer Engineering
MET, Institute of Engineering, Nashik, India

Abstract. Neural codec language models (NCLMs) have re- cently emerged as a powerful paradigm for unified speech generation and transformation. By modeling discrete acoustic tokens extracted from neural audio codecs, these systems enable scalable solutions for text-to-speech (TTS), voice conversion, speech enhancement, and editing within a single generative framework. This paper presents a comprehensive review of representative models including AudioLM, VALL-E, Voicebox, NaturalSpeech 2, and SpeechX, analyzing their architectural design, probabilistic modeling strategies, computational complex- ity, and task generalization capabilities. A comparative study highlights the tradeoff between perceptual quality and inference efficiency across autoregressive and diffusion-based approaches. Furthermore, existing research gaps in discrete representation fidelity, evaluation standardization, and multi-task optimization are identified. Finally, a conceptual extension termed SpeechX++ is discussed to address limitations through emotion conditioning, multilingual adaptation, and efficient inference strategies. The review demonstrates the ongoing transition toward general- purpose speech foundation models capable of robust, scalable, and ethically responsible deployment.

Keywords: Neural Codec Language Models, Speech Gen- eration, Text-to-Speech, Diffusion Models, Zero-Shot Learning, Transformer Architectures, Speech Enhancement, Multi-Task Learning

I. Introduction

Deep learning has significantly transformed modern speech generation and speech transformation systems. Early neural vocoders such as WaveNet [1] demonstrated that autoregres- sive modeling of raw audio waveforms can generate highly natural speech, establishing the foundation for neural speech synthesis. Later, HiFi-GAN [2] improved inference efficiency through adversarial training, enabling real-time high-fidelity waveform generation.

The introduction of the Transformer architecture [3] rev- olutionized sequence modeling by leveraging self-attention mechanisms for long-range dependency learning. In speech applications, the Conformer model [4] integrated convolutional operations with self-attention, significantly enhancing speech representation capability and recognition performance. These architectural advancements laid the groundwork for large-scale generative speech modeling.

A major paradigm shift occurred with the development of neural codec language models (NCLMs). Instead of directly predicting spectrograms or waveforms, these models operate over discrete acoustic tokens extracted from neural audio codecs. AudioLM [5] showed that language modeling over hierarchical discrete speech tokens can generate coherent long- form speech without explicit phoneme supervision. Building upon this framework, VALL-E [6] demonstrated zero-shot text-to-speech synthesis using short acoustic prompts, high- lighting the capability of codec-based autoregressive modeling. Voicebox [7] further extended this concept to multilingual speech generation and speech editing tasks, indicating the scalability of masked token modeling approaches.
The effectiveness of codec-based modeling relies heavily on high-fidelity discrete representation learning. EnCodec [8] introduced neural audio compression techniques that produce compact and perceptually robust discrete tokens suitable for language modeling. More recently, diffusion-based generative models such as NaturalSpeech 2 [9] improved speech natural- ness and robustness by modeling latent acoustic distributions through stochastic denoising processes.

Beyond speech synthesis, neural generative frameworks have been applied to speech enhancement and target speaker extraction. VoiceFilter [10] demonstrated speaker-conditioned speech separation in multi-speaker scenarios. Comprehensive overviews on neural target speech extraction [11] emphasize the importance of unified frameworks capable of handling noisy and overlapping speech environments. In this direction, SpeechX [13] proposed a unified neural codec language mod- eling framework capable of performing zero-shot TTS, speech enhancement, speech editing, and target speaker extraction within a single generative transformer architecture.

Despite these advancements, several challenges remain, including computational complexity of autoregressive infer- ence, stability of duration modeling, reconstruction artifacts introduced by discrete codecs, and inconsistencies between objective metrics such as DNSMOS [12] and perceived human quality. Therefore, a systematic review of neural codec lan- guage models is necessary to analyze architectural evolution, probabilistic modeling strategies, training objectives, evalua- tion methodologies, and emerging directions toward scalable speech foundation models.

II. Literature Review

The evolution of neural speech generation systems can be broadly categorized into waveform modeling, transformer- based architectures, neural codec language models, diffusion- based synthesis, and unified multi-task frameworks.

Waveform and Neural Vocoder Models
Early neural speech synthesis systems directly modeled raw audio waveforms. WaveNet [1] introduced autoregressive convolutional modeling for waveform generation, achieving highly natural speech quality but at the cost of high inference latency. To address computational inefficiency, HiFi-GAN [2] proposed adversarial waveform generation, significantly re- ducing inference time while maintaining perceptual fidelity. These works established the foundation for neural vocoding and waveform modeling.

Transformer-Based Speech Modeling
The introduction of the Transformer architecture [3] en- abled improved long-range dependency modeling through self- attention mechanisms. In speech processing, the Conformer model [4] combined convolutional layers with self-attention to enhance local and global feature extraction. These ar- chitectures provided scalable sequence modeling capabilities essential for large speech generation systems.

Neural Codec Language Models
A paradigm shift occurred with the emergence of neural codec language models (NCLMs). AudioLM [5] demonstrated that hierarchical discrete acoustic token modeling can generate coherent long-form speech without explicit phoneme super- vision. VALL-E [6] extended this concept to zero-shot text- to-speech synthesis using short acoustic prompts. Voicebox [7] further expanded masked token modeling to multilingual speech generation and editing tasks.

The effectiveness of NCLMs relies heavily on discrete audio tokenization methods. EnCodec [8] introduced high- fidelity neural audio compression using vector quantization, enabling compact and robust discrete representations suitable for language modeling.

Diffusion-Based Speech Synthesis
Diffusion models have recently gained attention for speech synthesis. NaturalSpeech 2 [9] employed latent diffusion pro- cesses to improve perceptual quality and robustness. Unlike autoregressive models, diffusion-based approaches generate speech through iterative denoising, offering smoother acoustic transitions at increased computational cost.

Speech Enhancement and Target Speaker Extraction
Beyond synthesis, neural modeling has expanded to enhance- ment and extraction tasks. VoiceFilter [10] introduced speaker- conditioned spectrogram masking for target speech extraction in multi-speaker environments. A comprehensive overview of neural target speech extraction methods was presented in [11], highlighting the growing demand for unified and robust speech transformation systems.

Unified Multi-Task Frameworks
Recent research has focused on unifying multiple speech tasks within a single generative framework. SpeechX [13] proposed a versatile neural codec language model capable of performing zero-shot TTS, speech enhancement, speech editing, and target

speaker extraction using a shared autoregressive transformer backbone. Such unified systems indicate a shift toward speech foundation models that generalize across multi- ple domains.

Finally, objective evaluation frameworks such as DNS- MOS [12] have been developed to estimate perceptual quality without reference signals, facilitating large-scale bench- mark- ing of generative speech systems.

III. Taxonomy of Neural Codec Language Models

Table 1: Taxonomy of Neural Codec Language Modeling Approaches

Category	Representative Models	Strength	Limitation
Waveform Models	WaveNet, HiFi-GAN	High Natural- ness	High Compu- tational Cost
Autoregressiv NCLM	e AudioLM, VALL-E	Zero-Shot Adaptation	Sequential In- ference Delay
Masked To- ken Models	Voicebox	Robust Edit- ing	Complex Training
Diffusion Models	NaturalSpeech2	High Percep- tual Quality	Iterative Sampling Overhead
Unified Multi-Task	SpeechX	Task General- ization	Multi- Objective Optimiza- tion Complexity

Table I categorizes neural speech generation systems based on modeling paradigm and architectural design. The taxonomy highlights the transition from waveform-level mod- eling to discrete token-based language modeling, followed by diffusion and unified multi-task approaches. This classification provides a structured understanding of the research landscape and clari- fies the trade-offs between efficiency, generalization ca- pability, and perceptual quality.

Table 2: Evolution Timeline of Key Speech Generation Models

Year	Model	Key Contribution
2016	WaveNet	Autoregressive raw waveform modeling
2020	HiFi-GAN	Fast GAN-based neural vocoder
2023	AudioLM	Hierarchical codec tokenlanguage modeling
2023	VALL-E	Zero-shot TTS via codec prompts
2023	Voicebox	Masked token mod- eling + editing
2024	NaturalSpeech2	Latent diffusion speech synthesis

2024	SpeechX	Unified multi-task codec language modeling

IV. Datasets and Benchmarks

Neural codec language models are typically trained and evaluated on large-scale speech corpora that vary in size, language diversity, and speaker coverage. Table III summarizes commonly used benchmark datasets in generative speech modeling.

Table 3: Commonly Used Speech Datasets in Neural Codec Language Modeling

Dataset	Language	Hours	Use Case
LibriSpeech	English	1000+	TTS / ASR
Libri-Light	English	60000+	Self-Supervised Pretraining
Common Voice	Multilingual	15000+	Multilingual TTS
VCTK	English	44	Multi-Speaker TTS
VoxCeleb2	Multilingual	2400+	Speaker Adaptation

These datasets differ significantly in scale and linguistic diversity. Large-scale corpora such as Libri-Light enable self- supervised pretraining, while smaller multi-speaker datasets such as VCTK are commonly used for controlled speaker adaptation experiments. Multilingual datasets such as Com- mon Voice facilitate cross-lingual and zero-shot evaluation, which are critical for scalable speech foundation models.

V. Mathematical Formulation Of Neural Codec Language Models

Neural codec language models (NCLMs) operate by modeling discrete acoustic tokens rather than continuous waveforms. Given an input speech signal x(t), a neural audio codec encoder [8] transforms it into a sequence of discrete tokens:

$$z = \{z1, z2, \ldots, zT\}, (1)$$

where $z_t \in V$ and V denotes the discrete token vocabulary obtained via vector quantization.

Autoregressive Modeling
Autoregressive neural codec models such as AudioLM [5] and VALL-E [6] estimate the joint probability of token sequences as:

$$P(z) = Y\ P(z_t \mid z_{<t}), \quad (2)$$

where $z_{<t}$ represents all previous tokens.
The training objective minimizes the negative log- likelihood:

LAR

= − Σ log P (zt t=1 | z<t). (3)

This formulation enables coherent long-form speech gener- ation but introduces se- quential inference latency proportional to sequence length T .Conditioned Generation
For text-to-speech synthesis, conditional modeling is applied:

P (z | y) = Y P (zt | z<t, y), (4)

where y represents text embeddings derived using Trans- former encoders [3]. This approach enables zero-shot speaker adaptation as demonstrated in [6].

Diffusion-Based Modeling
Diffusion models such as NaturalSpeech 2 [9] model speech generation through a de- noising process. The forward diffusion adds Gaussian noise:

q(xt | xt−1) = N (√1 − βtxt−1, βtI), (5) while the reverse process learns to predict noise ϵ using a neural network:

Ldiff = Et,ϵ ϵ − ϵθ(xt, t) 2 . (6)

Diffusion-based generation improves perceptual quality but increases computational complexity due to iterative denoising steps.

Speech Enhancement and Extraction Objective
For target speaker extraction [10], the objective can be formu- lated as:
Lenh =

sˆ(t) − s(t) 2, (7)
where sˆ(t) is the enhanced signal and s(t) is the clean reference.

Evaluation Metrics
Objective evaluation commonly includes Word Error Rate (WER):
S + D + I WER = , (8)

N
where S, D, and I denote substitution, deletion, and inser- tion errors, and N is the total number of reference words.

Perceptual quality is often estimated using non-intrusive metrics such as DNSMOS [12], which predicts Mean Opinion Score (MOS) through learned regression models.

Computational Complexity
Autoregressive decoding has complexity O(T) sequential steps, whereas self-attention mechanisms scale as O(T 2). Diffusion models require K iterative denoising steps, re-

sulting in complexity O(K · T), highlighting the trade-off between quality and inference speed. Figure 1 illustrates the theoret-ical computational trends of autoregressive, attention-based, and diffusion-based speech generation models. Autoregressive decoding scales linearly with sequence length, whereas self-attention mechanisms scale quadratically. Diffusion models introduce additional iterative steps, increasing computational cost.

t=1

This formulation enables coherent long-form speech gener- ation but introduces sequential inference latency proportional to sequence length T .

Conditioned Generation
For text-to-speech synthesis, conditional modeling is applied:
P (z | y) = Y P (zt | z<t, y), (4)
t=1

V. Conclusion

This paper presented a comprehensive review of neural codec language models (NCLMs) for unified speech generation and transformation. The evolution from waveform-level autore- gressive models to discrete token-based language modeling and diffusion-based synthesis was systematically analyzed. Comparative evaluation indicates that autoregressive NCLMs provide strong intelligibility and zero-shot adaptation capabilities, whereas diffusion-based approaches achieve superior perceptual smoothness at increased computational cost. Uni- fied frameworks such as SpeechX demonstrate the feasibility of integrating multiple speech tasks, including text-to-speech, enhancement, and speaker extraction, within a single genera- tive architecture.

Despite significant progress, challenges remain in inference latency, discrete representation fidelity, multi-task optimiza- tion stability, and evaluation standardization. The proposed conceptual extension, SpeechX++, highlights future directions toward emotion-conditioned prompting, multilingual scalabil- ity, and real-time inference optimization. Overall, neural codec language models are evolving toward general-purpose speech foundation systems that balance quality, efficiency, robustness, and ethical deployment considerations.

References

1. Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu, “WaveNet: A generative model for raw audio,” arXiv:1609.03499, 2016.
2. Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, “HiFi-GAN: Genera- tive adversarial networks for efficient and high fidelity speech synthesis,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2020.

3. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017.
4. Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, “Conformer: Convolution-augmented transformer for speech recognition,” in Proc. Interspeech, 2020.
5. Zala´n Borsos, Anton Sharifi, Damien Vincent, Eugene Kharitonov, Jan Zeghidour, and Marco Tagliasacchi, “AudioLM: A language modeling approach to audio generation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2525–2536, 2023.
6. Chengyi Wang, Sanyuan Chen, Yu Wu, Zhuo Chen, Shujie Liu, Jinyu Li, and Furu Wei, “Neural codec language models are zero-shot text-to- speech synthesizers,” arXiv:2301.02111, 2023.
7. Tuan Le, Yannis Jiang, Shubham Cornu, Daniel Ba¨r, Paul-Henri Roy, and Meta AI Research Team, “Voicebox: Text-guided multilingual universal speech generation at scale,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2023.
8. Alexandre De´fossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, “High fidelity neural audio compression,” Transactions on Machine Learning Research, 2023.
9. Kai Shen, Xu Tan, Tao Qin, Shansong Liu, and Tie-Yan Liu, “Natu- ralSpeech 2: Latent diffusion models are natural and zero-shot speech synthesizers,” in Proc. International Conference on Learning Represen- tations (ICLR), 2024.
10. Quan Wang, Hannah Muckenhirn, Kevin Wilson, and Zhifeng Chen, “VoiceFilter: Targeted voice separation by speaker-conditioned spectro- gram masking,” in Proc. Interspeech, 2019.
11. Kateˇrina Zˇ mol´ıkova´, Marc Delcroix, Takuya Ochiai, and Shinji Watan- abe, “Neural target speech extraction: An overview,” IEEE Signal Processing Magazine, vol. 40, no. 3, 2023.
12. Chandan K. Reddy, Vishak Gopal, Richard Cutler, and Sriram Srini- vasan, “DNSMOS: A non-intrusive perceptual objective speech quality metric,” in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.
13. Xiaofei Wang, Manthan Thakker, Zhuo Chen, Naoyuki Kanda, Sefik Emre Eskimez, Sanyuan Chen, Min Tang, Shujie Liu, Jinyu Li, and Takuya Yoshioka, “SpeechX: Neural codec language model as a versatile speech transformer,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 3355–3364, 2024.