

Zen-Dub-Live: Real-Time Streaming AI Dubbing for Live Video Content

Technical Whitepaper v2025.04

Zen LM Research Team
research@zenlm.org
papers.zenlm.org

April 2025

Abstract

Zen-Dub-Live is a 1 billion parameter streaming dubbing model designed for real-time AI dubbing of live video content, achieving end-to-end latency under 200ms (P95: 187ms) while supporting simultaneous translation and voice synthesis in 20+ languages. Unlike offline dubbing systems that process pre-recorded content with multi-second latency budgets, Zen-Dub-Live operates in a true streaming regime: it begins producing synthesized speech before the source speaker has finished the current utterance. The system achieves a speaker similarity mean opinion score of 4.0/5 and a livestream audio quality MOS of 3.9/5, demonstrating that streaming constraints do not require sacrificing perceptual quality. This paper describes the model architecture, streaming inference pipeline, latency decomposition, and quality evaluation.

Contents

1	Introduction	3
1.1	Model Overview	3
2	Architecture	3
2.1	Streaming Pipeline Overview	3
2.2	Causal Streaming ASR	4
2.3	Anticipatory Translation	4
2.4	Streaming Neural TTS	4
2.5	Zero-Shot Voice Cloning	5
2.6	Latency Decomposition	5
3	Training Methodology	5
3.1	Data	5
3.2	Streaming Training Objective	5
4	Evaluation	6
4.1	Latency	6
4.2	Speaker Similarity	6
4.3	Translation Quality	6
4.4	Livestream Quality	7
4.5	Language Support	7

5	Deployment	7
5.1	Infrastructure Requirements	7
5.2	Integration	7
6	Related Work	7
7	Conclusion	8

1 Introduction

Live video dubbing presents constraints that differ fundamentally from recorded content dubbing. Offline dubbing systems can analyze entire utterances before synthesis, align prosody with the full sentence structure, and post-process audio for quality. Live dubbing must:

1. Begin producing translated audio before the source utterance is complete.
2. Maintain temporal alignment with the original speaker’s lip movements to prevent jarring visual-audio desynchronization.
3. Handle barge-in, overlapping speech, and spontaneous corrections without introducing silence artifacts.
4. Maintain speaker voice identity across the session with no pre-enrollment.

Zen-Dub-Live addresses all four requirements through a streaming-first architecture that integrates speech recognition, machine translation, and voice synthesis in a unified, low-latency pipeline. The 1B parameter model is compact enough for real-time CPU-assisted inference on standard streaming infrastructure while maintaining the quality bar required for broadcast-grade live content.

1.1 Model Overview

Table 1: Zen-Dub-Live Model Specification

Parameter	Value
Architecture	Streaming encoder-decoder with causal attention
Total Parameters	1B
Supported Source Languages	20+
Supported Target Languages	20+
Language Pairs	400+
End-to-End Latency P50	142ms
End-to-End Latency P95	187ms
Speaker Similarity MOS	4.0/5
Livestream Quality MOS	3.9/5
Version	v2025.04
Release Date	April 2025

2 Architecture

2.1 Streaming Pipeline Overview

Zen-Dub-Live implements a fully streaming pipeline comprising four stages operating in a producer-consumer pipeline with bounded queues:

1. **Streaming ASR:** Causal acoustic encoder + streaming language model head produces partial transcripts on each incoming audio chunk (20ms frames).
2. **Anticipatory Translation:** A lightweight anticipatory module predicts likely continuations of partial transcripts, enabling early translation before the utterance completes.

3. **Prosody-Aligned Synthesis:** A streaming neural TTS synthesizer generates audio chunks synchronized to the translation output, with prosody conditioned on the source speaker’s pitch and rhythm.
4. **Voice Cloning:** A zero-shot voice cloning module adapts the synthesized audio to the source speaker’s voice characteristics extracted from a 3-second rolling context window.

2.2 Causal Streaming ASR

Standard ASR models use bidirectional attention, requiring the full utterance before producing a transcript. Zen-Dub-Live uses a causal conformer encoder [1] with a masked attention pattern that allows inference on streaming audio:

$$\text{Attn}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} + M_{\text{causal}} \right) V \quad (1)$$

where M_{causal} is a causal mask preventing attention to future frames. The encoder processes audio at 50Hz (20ms frames) with 1.5s of left context, balancing latency against acoustic model accuracy.

A streaming CTC decoder produces partial hypotheses on each frame, with confidence-gated emission: only token hypotheses exceeding a confidence threshold $\tau = 0.85$ are committed and forwarded to the translation stage.

2.3 Anticipatory Translation

Translating partial utterances introduces the problem of syntactic incompleteness: source language sentences may be structurally incomplete mid-utterance, making direct translation ambiguous or incorrect. Zen-Dub-Live addresses this with an anticipatory translation module that:

1. Maintains a probability distribution over likely utterance completions based on the partial transcript and language model priors.
2. Produces a translation hypothesis for the most likely completion.
3. Updates the hypothesis as new tokens are confirmed by the ASR stage.
4. Commits tokens to synthesis only when they are stable across the top- k completion hypotheses.

This anticipatory approach reduces synthesis lag by 40ms on average (measured against an oracle that waits for full utterance confirmation) at the cost of a 3.1% retranslation rate when the hypothesis must be revised.

2.4 Streaming Neural TTS

The speech synthesis stage uses a streaming-compatible neural vocoder based on a modified LightSpeed architecture [2]. The model generates audio samples in 20ms chunks, matching the input frame rate. Prosody conditioning is applied from a parallel prosody predictor:

$$\hat{s}_t = \text{Vocoder}(c_t, p_t, v_t) \quad (2)$$

where c_t is the phoneme sequence, p_t is the predicted pitch contour (Hz), and v_t is the voice conditioning vector from the zero-shot cloning module.

2.5 Zero-Shot Voice Cloning

Speaker identity is preserved without pre-enrollment using a zero-shot voice cloning approach. A speaker encoder E_ϕ extracts a 256-dimensional voice embedding from a 3-second rolling window of original speech:

$$v_t = E_\phi(a_{t-3s:t}) \quad (3)$$

The voice embedding conditions the vocoder through cross-attention in the synthesis layers, transferring spectral envelope characteristics (timbre, formant patterns) while allowing the translated phoneme sequence to drive articulation. Speaker similarity is measured by cosine distance between voice embeddings of original and dubbed speech.

2.6 Latency Decomposition

The 187ms P95 end-to-end latency budget is allocated across pipeline stages:

Table 2: Latency Budget Decomposition

Stage	P50 Latency	P95 Latency
Audio buffering (20ms frame)	20ms	20ms
Streaming ASR	28ms	41ms
Confidence gating delay	12ms	22ms
Anticipatory translation	31ms	48ms
Prosody prediction	8ms	12ms
Neural TTS synthesis	31ms	44ms
Total end-to-end	142ms	187ms

3 Training Methodology

3.1 Data

Zen-Dub-Live is trained on a multilingual corpus of aligned speech and text:

Table 3: Training Data Composition

Source	Hours	Description
Broadcast speech (licensed)	180,000	News, documentary, sports commentary
Multilingual audiobooks	42,000	20+ languages, aligned text
Spontaneous conversation	28,000	Podcast, interview, lecture
Parallel speech corpus	15,000	Professional dubbing references
Synthetic TTS augmentation	60,000	120+ voice styles, 20 languages
Total	325,000	

3.2 Streaming Training Objective

The key training innovation is the *streaming consistency loss* that penalizes revisions to committed synthesis tokens. When the anticipatory translation module must retranslate a committed segment (because the utterance completion differed from the predicted hypothesis), a revision penalty is applied:

$$\mathcal{L}_{\text{stream}} = \mathcal{L}_{\text{TTS}} + \lambda_1 \mathcal{L}_{\text{speaker}} + \lambda_2 \mathcal{L}_{\text{revision}} \quad (4)$$

where $\mathcal{L}_{\text{revision}} = \mathbb{E}[|\hat{t}_{\text{commit}} - \hat{t}_{\text{final}}|_1]$ penalizes the L1 distance between committed and final translations for the same utterance segments.

4 Evaluation

4.1 Latency

Table 4: End-to-End Latency Benchmark (20+ language pairs)

Language Pair	P50	P95	P99
English → Spanish	138ms	181ms	207ms
English → Japanese	148ms	196ms	224ms
English → Arabic	145ms	192ms	219ms
Spanish → English	141ms	184ms	211ms
Mandarin → English	152ms	198ms	231ms
French → German	139ms	183ms	209ms
Average (all pairs)	142ms	187ms	213ms

4.2 Speaker Similarity

Speaker similarity is measured by cosine distance between speaker embeddings extracted from original and dubbed audio, converted to a 5-point MOS scale through a calibrated mapping validated on 500 human listener ratings.

Table 5: Speaker Similarity MOS (1–5 scale)

System	Speaker Similarity MOS	95% CI
Zen-Dub-Live (streaming)	4.0	± 0.12
Offline dubbing baseline	4.4	± 0.09
Rule-based pitch shift	2.8	± 0.18
No voice cloning	1.9	± 0.21

The 0.4 MOS gap between Zen-Dub-Live and the offline baseline reflects the limited context available for zero-shot cloning in streaming mode.

4.3 Translation Quality

Table 6: Translation Quality (BLEU, case-sensitive)

Language Pair	Streaming BLEU	Offline BLEU	Gap
EN → ES	38.4	41.2	−2.8
EN → FR	42.1	44.7	−2.6
EN → DE	34.6	37.3	−2.7
EN → JA	29.8	32.4	−2.6
EN → ZH	31.2	34.1	−2.9
EN → AR	28.4	31.2	−2.8
Average	34.1	36.8	−2.7

The streaming mode incurs an average BLEU penalty of 2.7 points versus offline dubbing. This gap is attributable to anticipatory translation errors and the causal attention constraint in ASR.

4.4 Livestream Quality

Table 7: Livestream Quality MOS (1–5 scale, 200 listener panel)

Dimension	MOS Score	95% CI
Audio intelligibility	4.2	± 0.10
Naturalness	3.8	± 0.14
Temporal synchronization	3.9	± 0.12
Speaker consistency	4.0	± 0.11
Composite Livestream MOS	3.9	± 0.11

4.5 Language Support

Table 8: Supported Languages (source and target)

Region	Languages	Pairs
Western European	EN, ES, FR, DE, IT, PT, NL	42
Eastern European	PL, RU, CS, HU, RO, UK	30
East Asian	ZH, JA, KO	9
South/Southeast Asian	HI, ID, TH, VI, TR	20
Middle East / North Africa	AR, FA	2
Total	23 languages	400+

5 Deployment

5.1 Infrastructure Requirements

Zen-Dub-Live is designed for deployment on standard streaming infrastructure:

Table 9: Inference Hardware Requirements

Config	Hardware	Concurrent Streams	P95 Latency
GPU (recommended)	1 $\times$ A10G 24GB	48 streams	187ms
CPU-assisted	16-core CPU + GPU	24 streams	194ms
CPU-only (fallback)	32-core CPU	8 streams	312ms

5.2 Integration

Zen-Dub-Live exposes a WebSocket API that accepts a streaming audio input (PCM 16kHz mono) and configuration parameters (source language, target language, voice cloning mode), and returns a streaming audio output (PCM 24kHz stereo or AAC for direct RTMP insertion). Integration with major streaming platforms (RTMP, WebRTC, HLS) is supported through the Zen LM streaming SDK.

6 Related Work

Real-time speech translation has been addressed through cascaded [3] and end-to-end [4] approaches. Streaming ASR has advanced through CTC-based [5] and recurrent neural network transducers [6]. Zero-shot voice cloning for dubbing has been studied in [7]. Zen-Dub-Live is

the first system to integrate all components into a unified streaming model with sub-200ms end-to-end latency for live broadcast use.

7 Conclusion

Zen-Dub-Live achieves real-time AI dubbing of live video content with P95 end-to-end latency of 187ms, speaker similarity MOS of 4.0, and livestream quality MOS of 3.9 across 20+ languages and 400+ language pairs. The streaming-first architecture, anticipatory translation module, and zero-shot voice cloning combine to deliver broadcast-quality dubbing within live streaming latency constraints. Zen-Dub-Live opens practical AI dubbing to live sports, news, and entertainment broadcasting without requiring offline post-processing workflows.

References

- [1] Gulati, A. et al. (2020). Conformer: Convolution-augmented Transformer for Speech Recognition. Interspeech 2020.
- [2] Kim, J. et al. (2024). LightSpeed: Light and Fast Neural Vocoder using Dual WaveNet. ICASSP 2024.
- [3] Nakamura, M. et al. (2023). Cascade vs. Direct: What’s Best for Simultaneous Speech Translation? ACL 2023.
- [4] Lee, A. et al. (2022). Direct Speech-to-Speech Translation with Discrete Units. ACL 2022.
- [5] Graves, A. et al. (2006). Connectionist Temporal Classification. ICML 2006.
- [6] Graves, A. (2012). Sequence Transduction with Recurrent Neural Networks. arXiv:1211.3711.
- [7] Wang, C. et al. (2023). Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. arXiv:2301.02111.