

Zen-Dub: AI Voice Dubbing and Localization

Technical Whitepaper v2025.03

Zen LM Research Team

research@zenlm.org

papers.zenlm.org

March 2025

Abstract

Zen-Dub is a specialized 3B parameter model for AI-powered voice dubbing that combines neural machine translation with voice cloning to produce naturalistic dubbed audio preserving the original speaker’s voice characteristics, emotional cadence, and lip-sync timing across 50+ languages. Built on the Zen MoDE (Mixture of Distilled Experts) architecture with dedicated acoustic modeling components, Zen-Dub achieves speaker similarity MOS 4.3/5, translation BLEU 42.3, and lip-sync accuracy 87.4% on standard dubbing evaluation corpora. The model produces studio-quality dubbed audio in real-time on A10G hardware, enabling full-length feature film dubbing at 10–15× faster than real-time with a single GPU.

Contents

1	Introduction	2
2	Architecture	2
2.1	Source Audio Processing	2
2.2	Context-Aware Neural Machine Translation	3
2.3	Prosody Transfer	3
2.4	Lip-Sync Adaptation	3
2.5	Voice Synthesis with Speaker Cloning	4
3	Training	4
3.1	Training Data	4
3.2	Multi-Task Training	4
3.3	Human Evaluation Loop	4
4	Evaluation	5
4.1	Speaker Similarity	5
4.2	Translation Quality	5
4.3	Lip-Sync Accuracy	5
4.4	Processing Speed	5
5	Related Work	6
5.1	Neural Dubbing	6
5.2	Voice Cloning	6
5.3	Prosody Transfer	6
6	Conclusion	6

1 Introduction

Professional voice dubbing is one of the most labor-intensive tasks in media localization. For a 90-minute film, professional dubbing requires:

- Script translation and adaptation by localization experts
- Casting native-language voice actors
- Recording sessions in acoustic studios
- Audio post-production (mixing, timing adjustment, EQ matching)

This pipeline costs \$50,000–\$500,000 per language for feature-length content and takes 4–12 weeks per language. The result: most content is localized into fewer than 10 languages, leaving global audiences underserved.

Zen-Dub compresses this pipeline into a single model inference call, producing dubbed audio that preserves:

1. **Speaker voice identity** — Timbre, pitch range, resonance, and speaking style characteristics extracted from source audio, cloned into the target language.
2. **Emotional cadence** — Prosodic features (speech rate variation, emphasis patterns, pause placement) are detected in source audio and reproduced in target synthesis with language-appropriate phoneme mapping.
3. **Lip-sync timing** — Phoneme duration in the target language synthesis is adjusted to match source mouth movement timing, achieving natural lip-sync without visible audio lag.
4. **Background audio preservation** — Source non-speech audio (music, ambient sound, sound effects) is automatically separated and preserved in the output mix.

2 Architecture

Zen-Dub is a multi-component pipeline with an integrated 3B parameter neural backbone. The pipeline consists of five stages:

1. Source audio processing (ASR + speaker characterization)
2. Neural machine translation (context-aware)
3. Prosody transfer and lip-sync adaptation
4. Voice synthesis (TTS with speaker cloning)
5. Audio post-processing and mixing

2.1 Source Audio Processing

Automatic Speech Recognition (ASR): A streaming CTC-based ASR module transcribes source audio with word-level timestamps. Architecture: 6-layer conformer encoder, 2-layer CTC decoder. Word error rate $\leq 4\%$ across the 50 source languages.

Speaker characterization: A speaker embedding network extracts a 192-dimensional speaker identity vector $\mathbf{s} \in \mathbb{R}^{192}$ per speaking segment, capturing:

- F0 (fundamental frequency) range and distribution

- Spectral envelope (vocal tract shape)
- Speaking rate and rhythm
- Breathiness, roughness, and strain parameters (voice quality dimensions)

Emotion detection: A 12-class emotion classifier produces soft probability distributions over: neutral, happy, sad, angry, fearful, disgusted, surprised, excited, calm, contemptuous, bored, amused.

Background separation: A lightweight source separation module (UNet architecture, 200M parameters) separates speech from music and ambient sound, preserving the non-speech audio for the final mix.

2.2 Context-Aware Neural Machine Translation

Translation in dubbing differs from document translation: the output must be semantically accurate, phonetically compatible with the source timing, and natural in spoken form. Zen-Dub uses a modified Zen MoDE translation backbone with:

Isochrony constraints: The translation length (in phonemes) must approximately match the source length (in phonemes) to preserve lip-sync. A duration-aware decoding procedure biases beam search toward translations with matching phoneme counts:

$$\text{score}(y|x) = \log p_{\text{MT}}(y|x) - \lambda \left| \frac{|y|_\phi}{|x|_\phi} - 1 \right| \quad (1)$$

where $|y|_\phi$ and $|x|_\phi$ are the estimated phoneme counts of target and source strings, and $\lambda = 2.0$ is the isochrony penalty.

Spoken form optimization: A post-edit module rewrites written translations into spoken-language style: contractions, natural spoken discourse markers, and abbreviation expansion appropriate to the target language’s spoken register.

Character voice preservation: Named entities and proper nouns are handled with a character name preservation dictionary, ensuring consistency of character names across the dub.

2.3 Prosody Transfer

Prosody (speech rhythm, stress, intonation) is transferred from source to target via a prosody encoder-decoder:

$$\hat{\mathbf{p}}_{\text{target}} = f_{\text{prosody}}(\mathbf{p}_{\text{source}}, \mathbf{e}_{\text{emotion}}, \mathbf{l}_{\text{target}}) \quad (2)$$

where $\mathbf{p}_{\text{source}}$ is the source prosody representation (F0 contour, energy envelope, duration sequence), $\mathbf{e}_{\text{emotion}}$ is the emotion embedding, and $\mathbf{l}_{\text{target}}$ is a learned target language prosody prior.

The prosody decoder outputs word-level duration targets and phrase-level F0 contours in the target language, respecting the phonological constraints of the target language (e.g., tonal languages receive special treatment of F0 mapping).

2.4 Lip-Sync Adaptation

Lip-sync accuracy requires that the synthesized audio’s viseme sequence align with the source speaker’s mouth movements at a coarse level ($\pm 80\text{ms}$). The lip-sync adapter modifies phoneme durations in the synthesis plan:

$$d_i^* = \arg \min_{d_i} \left\{ |d_i - \hat{d}_i| + \mu \sum_j |\tau_j^{\text{sync}} - \tau_j^{\text{source}}| \right\} \quad (3)$$

where d_i are phoneme durations, $\hat{d}_i$ are the natural prosody targets, τ_j^{sync} are synthesized viseme boundary times, and τ_j^{source} are source mouth-open/close event times detected by a lightweight face motion detector.

2.5 Voice Synthesis with Speaker Cloning

The TTS backbone is a flow-matching vocoder conditioned on the speaker embedding:

$$\hat{A} = \text{Vocoder}(\mathbf{m}_{\text{phoneme}}, \mathbf{s}_{\text{speaker}}, \mathbf{p}_{\text{prosody}}) \quad (4)$$

where $\mathbf{m}_{\text{phoneme}}$ is the phoneme sequence with duration targets, $\mathbf{s}_{\text{speaker}}$ is the extracted speaker embedding, and $\mathbf{p}_{\text{prosody}}$ is the transferred prosody plan.

Zero-shot speaker cloning requires only 3–5 seconds of reference audio to produce a speaker embedding of sufficient quality for perceptually consistent synthesis. With 30+ seconds of reference audio, speaker similarity MOS exceeds 4.5/5.

3 Training

3.1 Training Data

Zen-Dub is trained on a multilingual dubbing corpus assembled from:

- Professional dub recordings (licensed): 48,000 hours across 50 languages
- Multilingual audiobooks (aligned with original): 120,000 hours
- Film and television subtitle-aligned audio: 800,000 hours (weakly supervised)
- Synthetic pairs generated by back-translation + TTS: 2M examples

3.2 Multi-Task Training

The model is trained jointly on:

- ASR: minimize CTC loss on source audio transcription
- MT: minimize cross-entropy on translation output
- TTS: minimize mel spectrogram reconstruction loss
- Speaker similarity: contrastive loss between same-speaker embeddings
- Lip-sync: minimize viseme boundary alignment error

The multi-task loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ASR}} + \mathcal{L}_{\text{MT}} + \mathcal{L}_{\text{TTS}} + \alpha \mathcal{L}_{\text{speaker}} + \beta \mathcal{L}_{\text{sync}} \quad (5)$$

with $\alpha = 0.5$, $\beta = 0.3$.

3.3 Human Evaluation Loop

Native speakers of each target language rate outputs using a 5-point MOS scale on three axes: naturalness, speaker similarity, and translation accuracy. Ratings below 3.5 trigger targeted fine-tuning on the flagged language pairs.

4 Evaluation

4.1 Speaker Similarity

Table 1: Speaker similarity MOS (5-point scale, native speaker raters)

Language Pair	Zen-Dub	3s Reference	30s Reference
EN → ES	4.3	4.1	4.6
EN → FR	4.2	4.0	4.5
EN → DE	4.1	3.9	4.4
EN → JA	4.0	3.8	4.3
EN → ZH	3.9	3.7	4.2
EN → AR	3.8	3.6	4.1
Average (50 pairs)	4.3	4.1	4.5

4.2 Translation Quality

Table 2: Translation quality (BLEU) on standard MT test sets

Language Pair	Zen-Dub BLEU	Isochrony-Constrained BLEU
EN → ES	46.1	43.8
EN → FR	44.3	42.1
EN → DE	38.2	36.4
EN → JA	29.4	27.8
EN → ZH	31.7	29.9
EN → AR	27.3	25.6
Average	42.3	40.2

4.3 Lip-Sync Accuracy

Table 3: Lip-sync accuracy metrics

Metric	Zen-Dub	Baseline (no sync)
Coarse sync accuracy (%)	87.4	61.2
Viseme boundary offset (ms)	68.3	184.7
Phrase boundary match (%)	91.2	73.4
Human MOS (lip-sync naturalness)	3.9	2.8

4.4 Processing Speed

Table 4: Dubbing throughput (minutes of dubbed output per minute of processing)

Hardware	Throughput (real-time ×)	GPU Memory
NVIDIA A10G (single GPU)	11.4×	22 GB
NVIDIA A100 (single GPU)	18.7×	42 GB
NVIDIA A100 (4× GPU)	62.3×	42 GB×4

5 Related Work

5.1 Neural Dubbing

Prior work on automatic dubbing focused primarily on isochrony constraints in machine translation [1] and prosody transfer [2]. End-to-end neural dubbing systems have emerged more recently, combining TTS, MT, and speaker adaptation in unified pipelines. Zen-Dub extends this work with native lip-sync adaptation and zero-shot multi-speaker cloning.

5.2 Voice Cloning

Zero-shot voice cloning from short reference audio has advanced rapidly through speaker embedding conditioning in neural vocoders. Flow-matching vocoders [3] achieve particularly high naturalness and speaker similarity, and form the backbone of Zen-Dub’s synthesis module.

5.3 Prosody Transfer

Cross-lingual prosody transfer is complicated by language-specific intonation systems. Work on prosody normalization and language-aware F0 transfer [4] has shown that speaker-independent prosody components (emphasis, rhythm) transfer better than language-dependent ones (tonal patterns in tonal languages).

6 Conclusion

Zen-Dub demonstrates that AI dubbing at studio-level quality across 50+ languages is achievable with a 3B parameter model operating at 11–18× real-time speed. The speaker similarity MOS of 4.3/5, BLEU 42.3, and lip-sync accuracy of 87.4% represent a substantial advance over prior automated dubbing approaches.

The economic impact is significant: content that previously required months and hundreds of thousands of dollars per language can now be localized in hours at a fraction of the cost, democratizing global media access.

Future work targets emotional consistency across scene cuts, multi-speaker separation for crowd scenes, and singing voice dubbing for musical content.

Acknowledgments

The authors thank the Zen LM linguistic team for multilingual evaluation support and native-speaker rater recruitment across all 50 languages.

References

- [1] E. Zetterholm, “Voice Imitation: A Phonetic Study of Perceptual Illusions and Acoustic Success,” *Lund University*, 2004.
- [2] M. Federico et al., “From Speech to Speech Translation,” *ICASSP*, 2020.
- [3] J. Kong et al., “VITS2: Improving Quality and Efficiency of Single-Stage Text-to-Speech with Adversarial Learning and Architecture Design,” *Interspeech*, 2023.
- [4] C. Wan et al., “Prosody Transfer in Neural Machine Translation for Dubbing,” *ACL*, 2023.