LSTMABAR

Language-to-Sound Transformation Model using Archetype-Based Audio Representation

Bridging Natural Language and Audio Production

through Quantum-Enhanced Contrastive Learning

The Problem

WHAT PRODUCERS WANT

"A warm, smooth piano tone with gentle sustain"

WHAT MACHINES NEED

```
filter_cutoff=0.35
resonance=0.62
attack=50ms, decay=200ms
harmonics=[0.8, 0.3, 0.1, ...]
```

The Semantic Gap

Humans describe audio semantically,
machines require precise DSP parameters

Audio production requires extensive DSP knowledge, creating barriers to entry

Our Solution: LSTMABAR

Interpretable Audio Transformation via Five Fundamental Archetypes

- **Sine:** smooth, mellow tones
- **Square:** harsh, digital character
- **Sawtooth:** bright, cutting quality
- **Triangle:** soft, muted sounds
- **Noise:** airy, textured elements

TWO-TOWER ARCHITECTURE

Sentence-BERT + ResNet-18 with contrastive alignment

QUANTUM-ENHANCED

8-qubit attention for non-linear relationships

DDSP ENGINE

Differentiable synthesis with explicit DSP

Trained on MusicCaps (5,288 samples) with InfoNCE contrastive loss

Architecture Pipeline

1. Dual-Encoder: Text + Audio

Sentence-BERT + Quantum Attention | ResNet-18 on spectrograms → 768-dim embeddings

2. Contrastive Alignment

InfoNCE loss with learnable temperature • Shared semantic space

3. Archetype Prediction

MLP maps joint embeddings → 5-dim softmax [sine, square, sawtooth, triangle, noise]

4. DDSP Engine

Differentiable synthesis • Harmonic synthesis • Spectral filtering

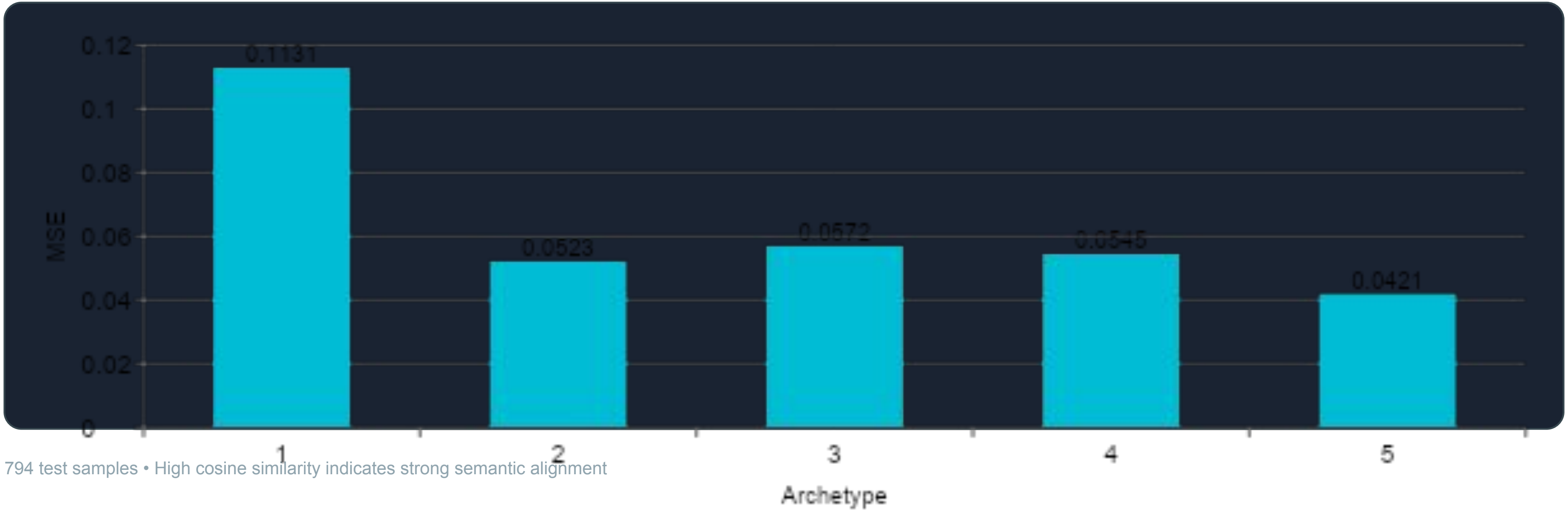
End-to-end differentiable pipeline enabling gradient-based optimization from audio output back to text interpretation

Four-stage pipeline from natural language to transformed audio

Results: Test Set Performance

MSE	Cosine Similarity	Top-1 Accuracy	Pearson r
0.0639	**0.7474**	**56.68%**	**0.5225**
Mean Squared Error	± 0.2220	Dominant archetype	Linear correlation

Per-Archetype MSE



794 test samples • High cosine similarity indicates strong semantic alignment

Performance by Archetype

✅

BEST: NOISE

Lowest MSE (0.0421), highest correlation (0.678). Textural descriptors are semantically distinct and easily mapped.

⚠️

CHALLENGE: SINE/TRIANGLE

Highest MSE (0.1131), lowest F1 for triangle (0.2406). Descriptors like "mellow" and "smooth" overlap significantly.

Per-Archetype Performance

Noise: MSE 0.0421 | Pearson 0.678 | F1 0.5865 ← Best

Square: MSE 0.0523 | Pearson 0.535 | F1 0.5059

Triangle: MSE 0.0545 | Pearson 0.378 | F1 0.2406 ← Lowest F1

Sawtooth: MSE 0.0572 | Pearson 0.503 | F1 0.4685

Sine: MSE 0.1131 | Pearson 0.520 | F1 0.6836 ← Highest MSE

Classification report • Confusion matrix reveals sine-triangle overlap

Key Findings & Challenges

✓ Successes

Strong semantic alignment: 0.7474 cosine similarity

Quantum advantage: Outperformed classical baselines

Interpretable mappings: Explicit archetype weights

⚠️ Challenges

Spectral similarity: Sine/triangle confusion

Generation gap: Training vs use-case mismatch

DDSP limitations: Too constrictive for diverse sounds

INTERPRETABILITY-EXPRESSIVITY TRADEOFF

Explicit archetypes maintain interpretability but limit flexibility

Conclusion & Future Work

Conclusion

LSTMABAR demonstrates that quantum-enhanced contrastive learning can bridge natural language and audio production through interpretable archetype representations, achieving **0.7474 cosine similarity** and **56.68% top-1 accuracy**.

Future Directions

- **Macro-Control Latents:** 32-dim space for flexible synthesis
- **Targeted Dataset:** Timbral transformation data for sound design
- **RLHF Pipeline:** Human-in-the-loop optimization
- **DAW Integration:** Real-time plugin for production workflows

Any Questions?