LungLDM: Prompt-Based Synthetic Lung Sound Generation Using Latent Diffusion Models for Respiratory Health Diagnostic

Mohammed Mosuily[1], *Jagmohan Chauhan*[2]

[1]University of Southampton, UK
[2]University College London, UK

mtm1g19@soton.ac.uk, jagmohan.chauhan@ucl.ac.uk

1. Introduction

Lung sounds provide valuable insights for diagnosing and monitoring respiratory diseases such as asthma and chronic obstructive pulmonary disease (COPD). However, acquiring large-scale, diverse respiratory audio datasets is difficult due to privacy concerns, ethical restrictions, and the cost of clinical-grade recordings. This scarcity hinders the performance of AI models trained for respiratory health monitoring. Inspired by recent advances in generative models, we present **LungLDM**, a domain-specific latent diffusion model that generates realistic lung sounds from structured clinical prompts, helping to overcome the bottleneck of real-world data collection.

LungLDM introduces three key innovations: (1) an exponential penalty loss to improve the fidelity of reconstructed signals, (2) a squeeze-and-excitation encoder for channel-aware feature extraction, and (3) a channel-wise gating mechanism to enhance feature modulation in the U-Net diffusion backbone. Our model is trained on 124,883 respiratory recordings from 9 public datasets. Evaluation results demonstrate that LungLDM significantly outperforms AudioLDM [1] in both objective and subjective metrics, including KL divergence (0.01 vs. 1.75), Fréchet Audio Distance (14.71 vs. 18.82), and listener-rated relevance (64.94 vs. 52.69).

2. System Overview

LungLDM is a text-to-audio latent diffusion model designed to synthesize lung sounds from clinical-style text prompts. The model consists of a Flan-T5 text encoder, a U-Net-based latent diffusion network, and a VAE with a vocoder for audio reconstruction. Prompts include metadata such as age, gender, and condition (e.g., `"A person is coughing, male, 40s, smoker"`). These are converted to embeddings and used to condition the audio generation process.

A. Loss Function Design

To enhance reconstruction fidelity, we propose a loss function combining L1 loss, LPIPS perceptual loss, and an exponential penalty:

$$\ell = \|x - \hat{x}\|_1 + \alpha\text{LPIPS} + \beta\mathbb{E}[\exp(\|x - \hat{x}\|_1)]$$

This penalizes large deviations in spectrogram reconstruction and improves clinical realism.

B. Feature Enhancement Modules

The model incorporates a squeeze-and-excitation (SE) encoder to learn per-channel attention scores and a channel-wise gating mechanism within the U-Net to modulate features more effectively.

3. Dataset and Training

LungLDM is trained on over 124k samples from datasets including COUGHVID, MMLung, UK COVID-19, and others. Audio is resampled to 16kHz, preprocessed, and categorized into 13 types (e.g., cough, breathing, vowels). Text prompts are derived from associated metadata.

Training is performed on NVIDIA A100 GPUs for 250,000 steps. The model uses AdamW optimization with a learning rate of 2×10^{-4} and guidance scale of 3.5. The final model consists of 96M parameters.

4. Results and Discussion

A. Objective Metrics

LungLDM demonstrates strong performance across standard audio generation metrics:

Table 1: *Performance Comparison with AudioLDM*

Model	FAD ↓	KL ↓	IS ↑
AudioLDM	18.82	1.75	1.01
LungLDM	**14.71**	**0.01**	1.01

B. Subjective Evaluation

34 participants rated 39 audio files (13 ground truth, 13 AudioLDM, 13 LungLDM). Scores for Overall Audio Quality (OVL) and Relevance to Text (REL) improved as follows:

- **OVL**: 63.63 → 65.89
- **REL**: 52.69 → 64.94

5. Conclusion

LungLDM is a domain-specialized generative model for synthesizing high-quality lung sounds from clinical prompts. By enhancing diffusion architectures with SE encoding, gating, and novel loss design, LungLDM achieves superior fidelity and perceptual alignment compared to general-purpose models. This work contributes toward scalable, privacy-respecting training datasets for AI-based respiratory diagnostics.

6. References

[1] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-Audio Generation with Latent Diffusion Models," in *ICML International Conference on Machine Learning*, 1 2023, pp. 21 450 – 21 474. [Online]. Available: http://arxiv.org/abs/2301.12503