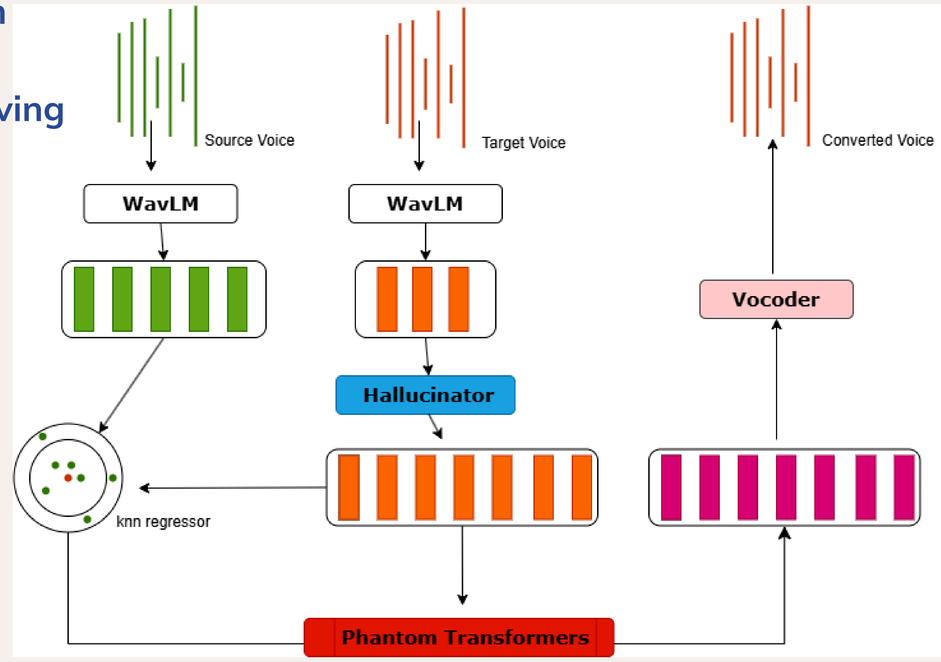
Anshul Thakur (B21CS085) Samay Mehar(B22AI048)

Zero-Shot Voice Conversion with Hallucinator and Phantom Embeddings

We present a novel zero-shot voice conversion (VC) framework that enhances audio conversion quality through innovative components. Our system incorporates: (i) a diffusion-based hallucinator that generates high-fidelity embeddings to replace traditional VAE models, and (ii) a phantom transformer that injects human-like filler sounds during silences by learning from secondary speaker embeddings. Drawing inspiration from the Phoneme Hallucinator, our approach addresses limitations in existing methods, particularly KNN-based VC, by improving intelligibility and speaker similarity.



Methodology

Our pipeline begins by extracting frame-level source and target embeddings from WavLM. If the target set is small, a diffusion hallucinator generates additional embeddings by denoising Gaussian noise conditioned on the target samples. A phantom transformer then injects natural fillers into silent segments to smooth transitions. We combine original, target, and hallucinated embeddings, feed them into a HiFiGAN vocoder, and synthesize coherent, high-fidelity speech while maintaining semantic and temporal coherence.

Dataset Used

VCTK

Expected Outcomes

1. Significantly improved intelligibility and speaker similarity in zero-shot voice conversion, achieving lower WER and higher MOS compared to baseline methods.
2. Natural, human-like speech with seamless transitions and realistic fillers, resulting in outputs that closely mimic target speaker prosody and reduce robotic artifacts.

References

1. Voice Conversion With Just Nearest Neighbors [https://arxiv.org/abs/2406.03816]
2. Phoneme Hallucinator: One-shot Voice Conversion via Set Expansion [https://arxiv.org/pdf/2305.14078]