先端科学技術研究科　修士論文要旨

所属研究室 (主指導教員)	ヒューマンAIインタラクション (Sakriani Sakti （教授))		
学籍番号	2411808	提出日	令和 7年 1月 21日
学生氏名	WANG YINGJIE		
論文題目	Fast and High Fidelity Speech Synthesis via Flow Matching in Frequency Domain		

要旨

Neural vocoders based on Generative Adversarial Networks (GANs) have emerged as a predominant methodology for high-quality speech synthesis. Based on which inverse Short-Time Fourier Transform (iSTFT) approaches have further demonstrated superior inference speed and quality. Nevertheless, these methods remain constrained by training instability and substantial computational requirements introduced by discriminative training. On the other hand, while diffusion models exhibit considerable potential in generating high-fidelity samples, their iterative sampling process makes it difficult to integrate with real-time applications.

This thesis introduces a neural vocoder combining frequency domain modeling with Optimized Transportation Conditional Flow Matching (OT-CFM), addressing the trade-off between fast GAN-based iSTFT inference and high-quality but slow diffusion model generation. The vocoder uses inverse Short-Time Fourier Transform (iSTFT) to efficiently reconstruct time-domain signals from frequency domain representations.

A key contribution is the introduction of a fast dimension reduction method for OT-CFM training. Rather than using Variational Autoencoders (VAE), this thesis proposes a simpler approach by taking half of the amplitude and phase components and concatenating them, significantly reducing computational complexity and accelerating training.

The proposed vocoder achieves faster inference speed compared to traditional diffusion models while maintaining competitive performance with GAN-based iSTFT methods. Experiments demonstrate robustness to variations in speaker characteristics, making it suitable for synthesizing high-quality speech for unseen speakers.

The vocoder achieves competitive speech quality through automated MOS evaluation, performing on par with state-of-the-art solutions. This combination of fast inference, high fidelity, and robustness makes it promising for real-time speech synthesis applications.