DiTTo-TTS: Diffusion Transformers for Scalable Text-To-Speech without domain-specific factors

Keon Lee, Dong Won Kim, Jaehyeon Kim, Seungjun Chung, Jaewoong Cho

Implémenté par: HALIMI Abdelkrim, GHENAIET Walid, ATTABI Melissa Dahlia

Introduction

Les modèles de diffusion latente (LDMs) se sont montrés très performants dans diverses tâches comme la génération d'images, d'audios ou encore de vidéos. Appliqué à la synthèse vocale (Text-To-Speech, TTS), ces modèles nécessitent l'utilisation de facteurs spécifiques au domaine afin d'assurer un bon alignement temporel entre texte et parole. Cette dépendance complique la préparation des données et limite la scalabilité des modèles.

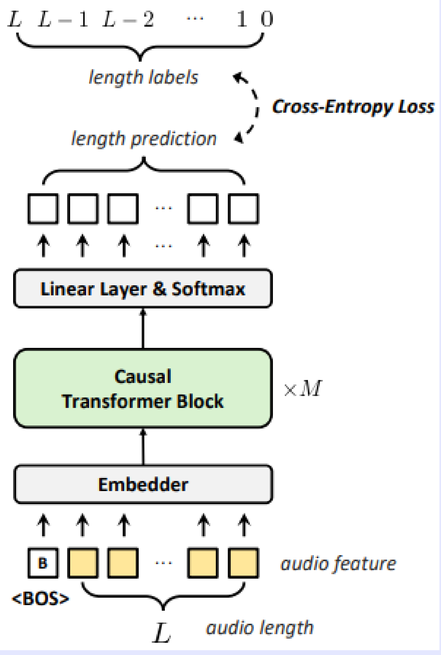
DiTTo-TTS est une méthode qui propose d'aller au delà de ces limitations tout en atteignant de bonnes performances. C'est une architecture basé sur un Diffusion Transformer (DiT), et intégrant un prédicteur de durée.

Modèle

Speech Length Predictor

Prédit la longueur totale du signal audio pour un texte donnée.
L'encodeur transforme le texte de manière bidirectionnelle.
Le décodeur prend en entrée l'audio encodé (NAC) et applique un masque causale.
La cross-attention entre les textes encodés et l'audio permet de prédire la longueur.

Entraîné séparément en utilisant l'entropie croisée.



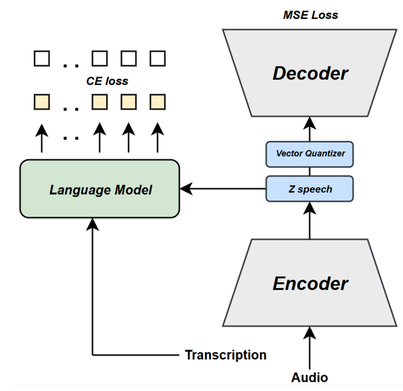
Neural Audio Codec

Encode les signaux audios en représentations latentes alignées avec le texte, les quantifie puis les décode.

4 éléments:

- Encodeur
- Quantificateur vectoriel
- Décodeur
- Modèle de langage

Loss: (1)

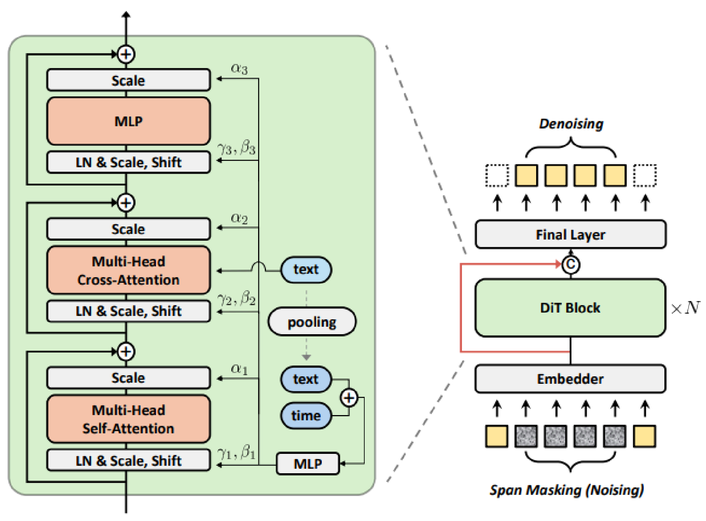


(1) $\mathcal{L}(\psi) = \mathcal{L}_{NAC}(\psi) + \lambda \mathcal{L}_{LM}(\psi),\ \mathcal{L}_{LM}(\psi) = -\log p_\phi(\boldsymbol{x}|f(\boldsymbol{z}_{speech}))$

Diffusion Model

Génère la parole à partir des représentations textuelles z_{text} et audio z_{speech} en utilisant un processus de diffusion.

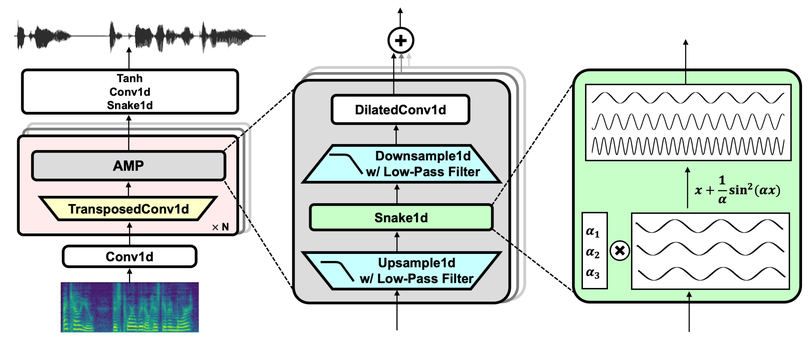
Entraîné à minimiser la fonction de perte suivante: (2)



(2) $\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{t\sim\mathcal{U}(1,T),\epsilon\sim\mathcal{N}(0,I)}\left[\|\boldsymbol{v}^{(t)} - \boldsymbol{v}_\theta(\boldsymbol{z}^{(t)}, \boldsymbol{x}, t)\|^2\right]$

Pipeline Audio-Textuel

Jeu de données: MLS Librispeech - 10000 audios gardés en français + transcription texte.
Preprocessing: Texte - Tokenization (GPT2, ByT5).
Audio - Rééchantillonnage à 24khz
Reconstruction du signal audio: BigVGAN



Résultats

Modèle	WER	CER	SIM-o
YourTTS	7.92 (7.7*)	3.18	0.3755 (0.337*)
VALL-E	5.9*	-	-
SPEAR-TTS	-	1.92*	-
Voicebox	1.9*	-	0.662*
CLaM-TTS	5.11*	2.87*	0.4951*
Simple-TTS	4.09 (3.4*)	2.11	0.5026
Ditto-TTS (Taille Fixe)	1.7180	1.0113	0.1886

Conclusion

Nos résultats confirment la pertinence de l'approche validant à la fois l'alignement texte-parole et l'architecture DiT.
Limitations:

- Speech Length Predictor non testé faute de ressources.
- Modèle de diffusion entraîné beaucoup plus petit et sur un jeu de données restreint.

SORBONNE UNIVERSITÉ