Adapter-Based Extension of Multi-Speaker TTS Model for New Speakers

- Paper, Nvidia

- 기존 Multi-Speaker TTS모델을 Adapter 튜닝을 통해 새로운 Speaker에 대해 업데이트

- TTS 모델
  - FastPitch, GST-speaker embedding -> CLN

- Adapter
  - Transformer layer 뒤에 모두 삽입
  - Pitch predictor, aligner 등 모든 모듈에 삽입

- Full Finetuning 과의 성능 비교
  - 훈련 Set이 1시간일 때는 더 좋은 성능
  - 데이터셋에 따라 성능 차이가 있지만, 대부분 Full Finetuning과 유사한 성능을 보임

WhiSLU: End-to-End Spoken Language Understanding with Whisper

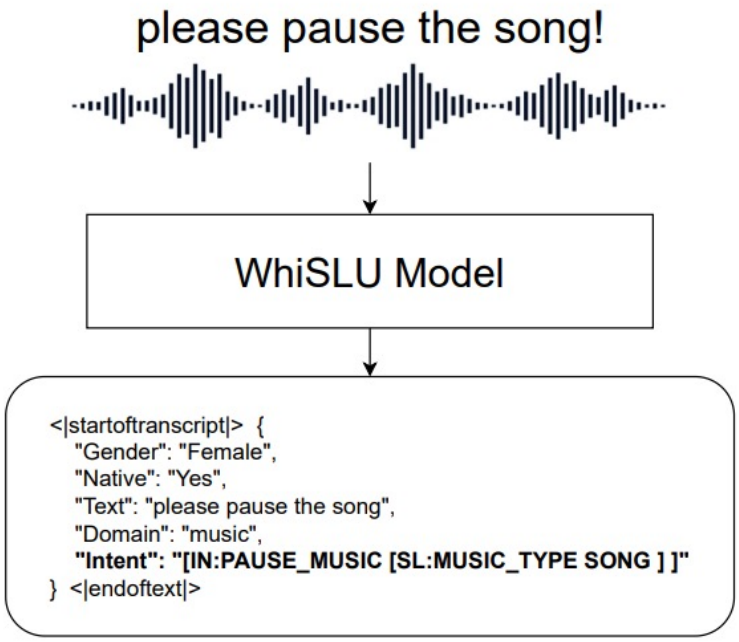


Figure 1: *This figure presents an example of WhiSLU's prediction. The model is trained to directly generate a well-formatted JSON string, learned from the sequence-level multitask learning strategy. The "Intent" entry represents the main task prediction, while the remaining entries correspond to the predictions of auxiliary tasks.*

- **Paper**
- **Whisper로 SLU를 훈련할 때, Auxiliary task를 활용**
- **여러가지 SLU task를 Json 형식으로 생성하도록 훈련**

Model	EM	EM-TREE	SLU-wer
wav2vec2.0 [5]	68.70	82.78	-
HuBERT [5]	69.23	82.87	-
cascade system	72.36	82.78	-
WhiSLU-large	74.49	84.89	6.8103
WhiSLU-large-SML	**76.68**	**86.37**	**6.1407**

Table 4: *Overall comparison result. The table illustrates a comparison between the performance of the baseline models and our proposed WhiSLU models. Notably, both the WhiSLU-large models with and without the sequence-level multitask learning (SML) strategy achieved significant improvements in EM, EM-TREE, and SLU-wer. This suggests that our proposed method of transferring knowledge from the ASR model to the NLU task is effective in enhancing the performance of the WhiSLU model.*

Whisper Features for Dysarthric Severity-Level Classification

- Paper
- **Whisper Encoder + CNN layers**

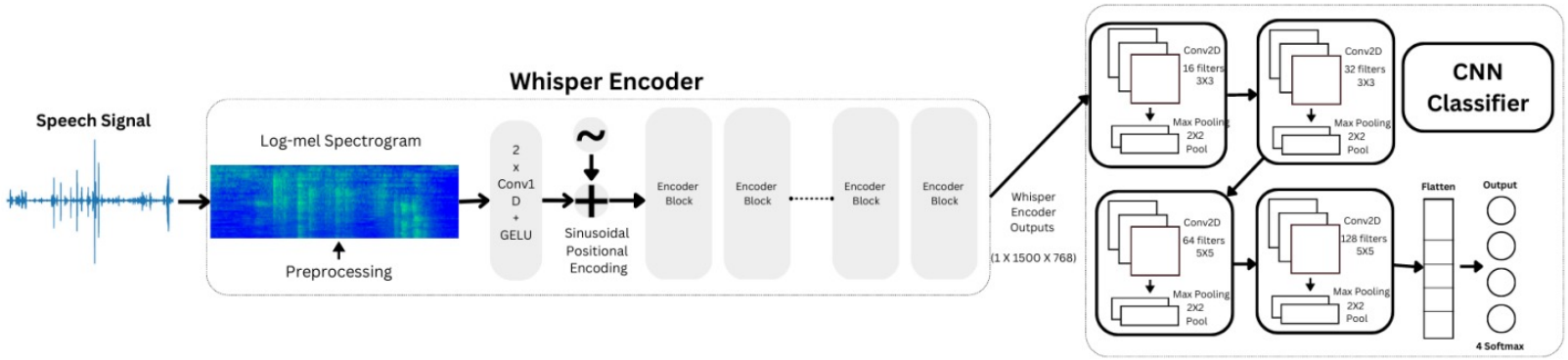


Figure 1: *Functional Block Diagram of Proposed Whisper Encoder Transfer Learning Pipeline in tandem with CNN classifier*

- 성능

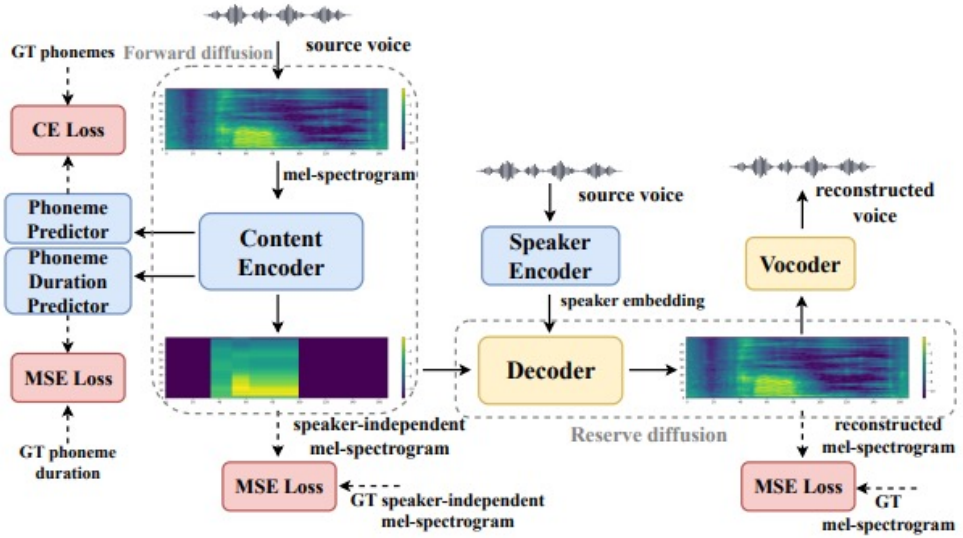
Table 5: *Performance Evaluation for Various Feature Sets*

Feature Set	Accuracy	F1-Score	MCC	Jaccard Index	Hamming Loss
MFCC	95.20	0.91	0.88	0.84	0.087
LFCC	96.05	0.96	0.96	0.93	0.034
Whisper	**98.02**	**0.98**	**0.96**	**0.97**	**0.019**

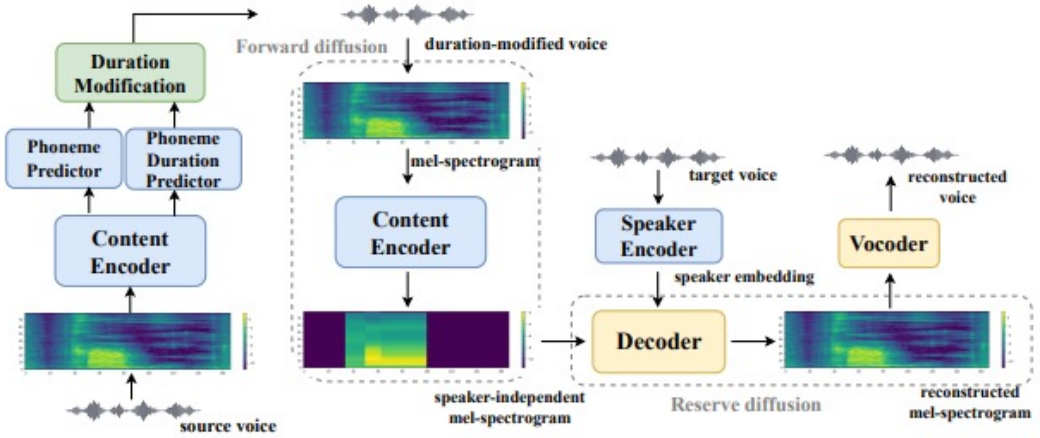
Related works

- **DuTa-VC: A Duration-aware Typical-to-atypical Voice Conversion Approach with Diffusion Probabilistic Model**
  - **Interspeech 2023, Paper, Github**

- **Target Speaker (Dysarthric speech)와 유사한 duration 정보를 가지도록 함**
- **Source Speaker에서 speaker-independent한 mel-spectrogram 을 추출**
- **의료진이 dysarthric 평가 feature 단위로 평가했을 때 유사한 지표를 보임**



* Training stage



* Inference stage

끝

감사합니다.