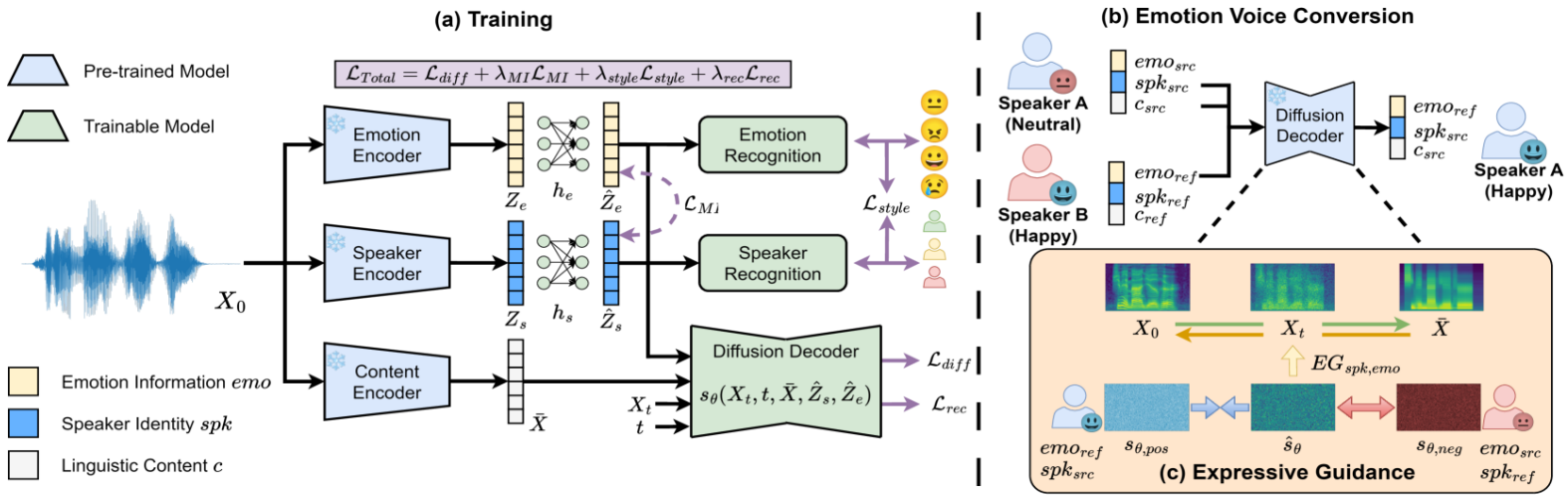
子計畫六： ZSDEVC: Zero-Shot Diffusion-based Emotional Voice Conversion with Disentangled Mechanism



Methods	MSP-Podcast					ESD				
	UTMOS	SECS	DNSMOS SIG	DNSMOS OVRL	ECA	UTMOS	SECS	DNSMOS SIG	DNSMOS OVRL	ECA
Source	2.830	1.000	3.401	2.892	0.621	3.927	1.000	3.479	3.191	0.951
Source (VO)	2.488	0.974	3.422	2.907	0.608	3.538	0.971	3.500	3.210	0.881
EMOCONV-DIFF	2.477	**0.837**	**3.528**	**3.096**	0.500	3.708	**0.822**	**3.572**	**3.315**	0.453
Our method ($\mathcal{L}_{Total}$)	2.427	0.773	3.505	3.073	0.584 (16.8 % ↑)	3.687	0.763	3.561	3.300	0.548 (21.1 % ↑)
w/ EG_{spk}	**2.493**	0.788	3.503	3.068	0.557 (11.4 % ↑)	**3.729**	0.774	3.562	3.302	0.504 (11.3 % ↑)
w/ EG_{emo}	2.353	0.744	3.485	3.040	**0.699** (40.0 % ↑)	3.665	0.747	3.560	3.298	**0.622** (37.4 % ↑)
w/ $EG_{spk,emo}$	2.383	0.766	3.484	3.032	0.672 (34.4 % ↑)	3.699	0.763	3.562	3.300	0.580 (28.1 % ↑)

Architecture: Encoders

Speech Representation Modules

- Lexical Content: Phoneme encoder (pretrained)
- Speaker Identity: Speaker encoder, speaker verification model
- Emotion Encoding: Emotion encoder, express encoder
- Input features: wav2vec with VAD (Voice Activity Detection)

Core Techniques

- Disentangled Representation: Separate speaker & emotion factors
- Guiding Mechanism: Improves expressiveness control

Diffusion-based Decoders

High quality samples

SDE

Training: predict added normal distribution noise by score model S_θ

Inference:

- Gradually remove noise
- Use sampling technique to effectively generate audios

Training Strategy: Training Objective

$L_{total} = L_{diff} + \lambda_{\mathrm{MI}} L_{MI} + \lambda_{style} L_{style} + \lambda_{rec} L_{rec}$

L_{diff}: Diffusion Loss

- MSE Loss between score function s_θ and prior noise

L_{MI}: Disentanglement Loss

- Minimize Mutual Information(MI) between speech representations

L_{style}: Style Loss

- SR/SER for style encoder

L_{rec}: Reconstruction Loss

- MSE Loss between source and estimation of source