

子計畫名稱：適用於智慧服務的可信賴AI先進技術研究

子計畫主持人：李祈均教授/ 國立清華大學

展示技術：強健性/ 隱私性/ 公平性

強健性 (Robustness)

SWiBE: A Parameterized Stochastic Diffusion Process for Noise-Robust Bandwidth Expansion



Goal

To develop a noise-robust **bandwidth expansion (BWE)** method that restores high-frequency components in low-quality audio, improving speech intelligibility and quality

Methodology

- **SWiBE (Step-Wised Bandwidth Expansion)**: A novel BWE framework based on score-based generative modeling (SGM) and a stochastic diffusion process

- **Noise Robustness**: Outperforms state-of-the-art (SOTA) models in handling noisy environments

- **Control of Expansion**: Allows parameterized control over the expansion process for optimal perceptual quality and spectral reconstruction

Results: PESQ: 2.83 ; ESTOI: 0.852 ; LSD: 1.04

隱私性 (Privacy)

RW-VoiceShield: Raw Waveform-based Adversarial Attack on One-shot Voice Conversion



Goal

RW-VoiceShield: a system that protects speaker identities by adding adversarial noise, preventing voice conversion models from accurately replicating voices and safeguarding privacy

Methodology

- Tested on the **Free-VC** voice conversion model

- Adversarial noise is generated directly on the **raw waveform**, avoiding multi-stage processing issues

- Conducted experiments under both **white-box** (full model access) and **black-box** (limited model access) scenarios

Results

- White-box testing: 94% attack success rate, minimal impact on audio quality

- Black-box testing: 85% attack success rate, still effective

公平性 (Fairness)

An Inter-Speaker Fairness-Aware Speech Emotion Regression Framework



Goal

Developed a fairness-aware SER framework focusing on inter-speaker fairness

Methodology

- Used **ECAPA-TDNN** for utterance-level embeddings

- Applied **HDBSCAN** clustering to create pseudo-speaker IDs

- Evaluated using **Concordance Correlation Coefficient (CCC)** and **discrimination metrics** for fairness across speakers, clusters, and gender

Results

The framework improves **inter-speaker fairness** while maintaining competitive SER performance, especially on datasets without speaker IDs

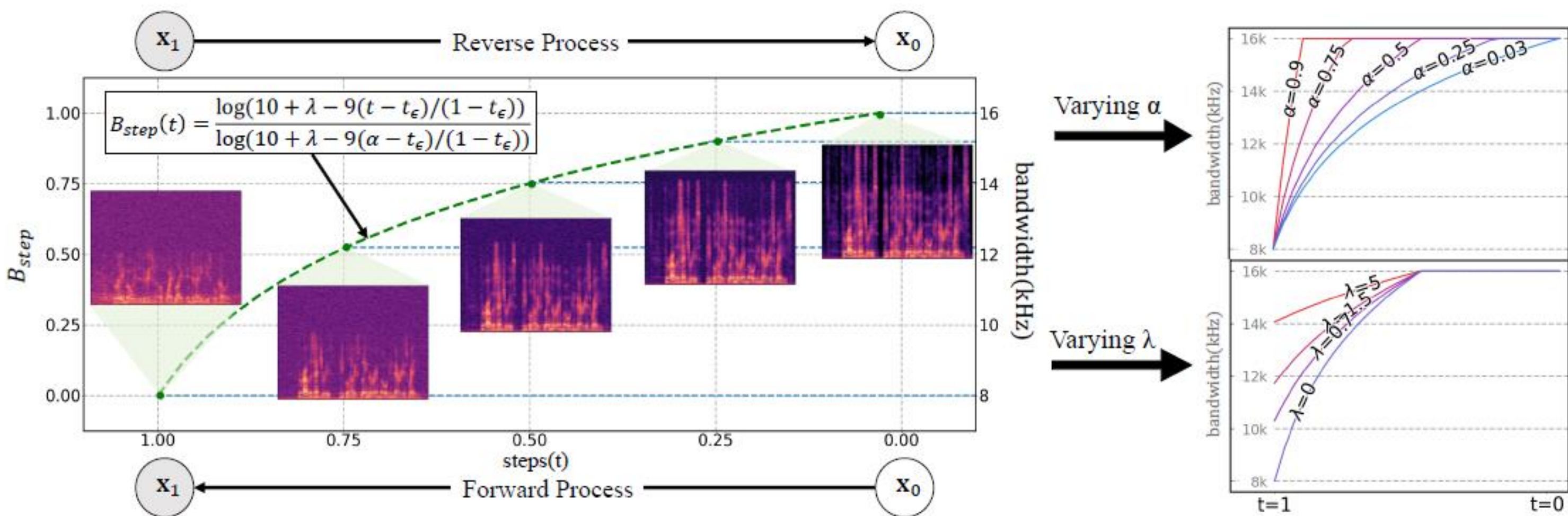


Figure 1: A visualization of the proposed B_{step} in the diffusion process of SWiBE, where x_1 and x_0 denote noise-perturbed 8k signal and clean 16k signal, respectively.

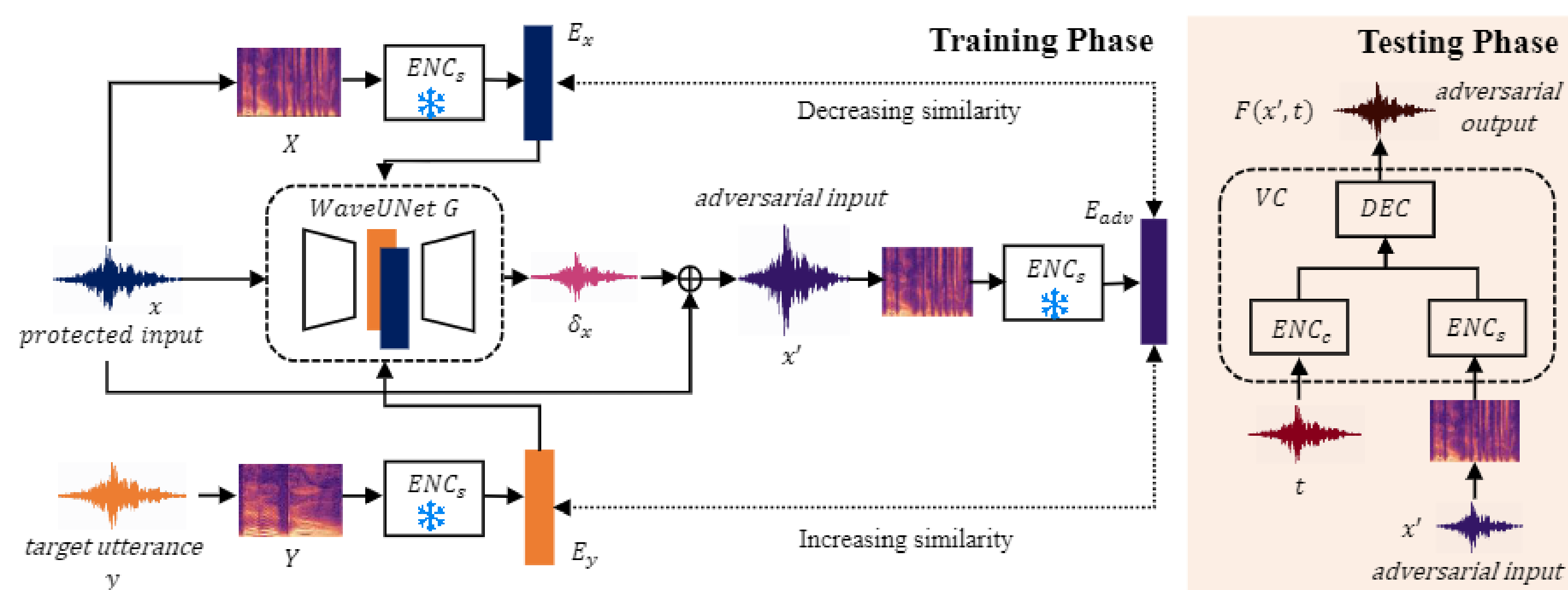


Figure 1: Training and Testing phase of RW-VoiceShield. The dashed double arrow section represents the target of the loss update.

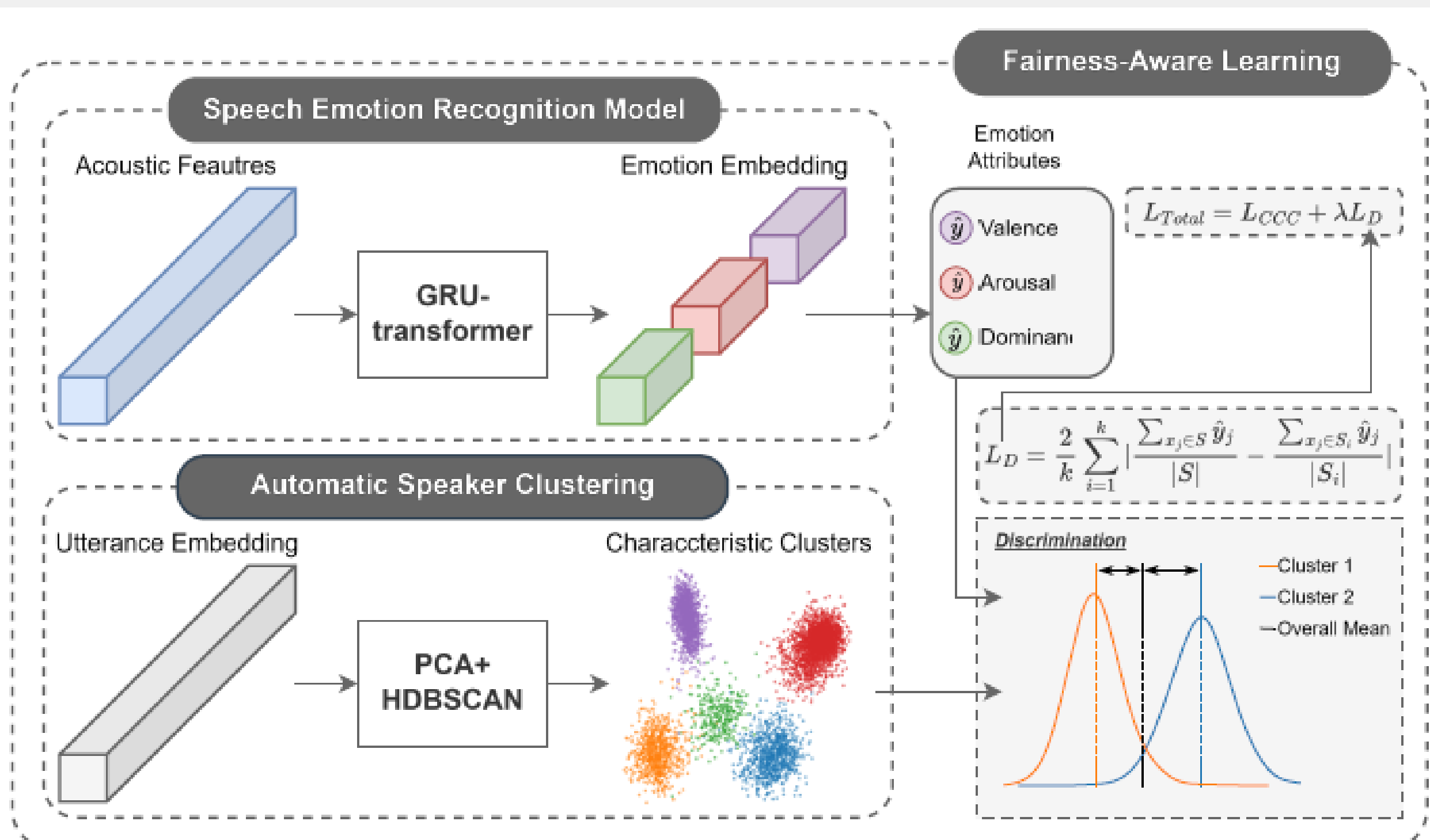


Figure 2: Overview of the speech characteristic fairness aware speech emotion recognition (SER) architecture.