Disentangling Speech Representations Learning with Latent Diffusion for Speaker Verification

Zhe Li, *Student Member, IEEE*, Man-Wai Mak, *Senior Member, IEEE*, Jen-Tzung Chien, *Senior Member, IEEE*, Mert Pilanci, *Member, IEEE*, Zezhong Jin, *Student Member, IEEE* Helen Meng, *Fellow, IEEE*

***Abstract*—Disentangled speech representation learning for speaker verification aims to separate spoken content and speaker timbre into distinct representations. However, existing variational autoencoder (VAE)–based methods for speech disentanglement rely on latent variables that lack semantic meaning, limiting their effectiveness for speaker verification. To address this limitation, we propose a diffusion-based method that disentangles and separates speaker features and speech content in the latent space. Building upon the VAE framework, we employ a speaker encoder to learn latent variables representing speaker features while using frame-specific latent variables to capture content. Unlike previous sequential VAE approaches, our method utilizes a conditional diffusion model in the latent space to derive speaker-aware representations. Experiments on the VoxCeleb and CN-Celeb datasets demonstrate that our method effectively isolates speaker features from speech content using pre-trained speech representations. The learned embeddings are robust to language mismatches since the speaker embeddings become content-invariant after content removal. Additionally, we design contrastive learning experiments showing that our training objective can enhance the learning of speaker-discriminative embeddings without relying on classification-based loss.**

***Index Terms*—Speaker verification, disentangled speech representation, latent diffusion model, variational autoencoder, pre-trained speech model**

I. Introduction

A speech signal is represented as a one-dimensional waveform. Despite its apparent simplicity, a speech waveform encodes a wealth of high-level information such as phonemes, tone, emotion, gender, and speaker identity. However, attributes like speaking style, prosody, recording conditions, and noise levels are challenging to annotate [1], [2].

To extract accurate speaker representations and mitigate the impact of speech content variation, existing methods employ phonetic content representations as a reference for speaker embeddings. Specifically, these methods include: (1) leveraging pre-trained automatic speech recognition (ASR) models [3], [4], [5], [6], [7] and (2) utilizing jointly trained multi-task models with additional modules for content representation [8], [9], [10]. These approaches demonstrate that incorporating content representations enhances speaker recognition performance.

However, both strategies face limitations in practical applications. Pre-trained ASR models significantly increase model size and computational complexity during inference, often by one or two orders of magnitude compared to speaker recognition models. Meanwhile, joint training with additional modules requires either a separate dataset with both text labels and speaker identities or a unified dataset containing both, which is often costly and challenging to obtain.

Disentangled representations have garnered considerable attention in recent research due to their ability to capture distinct variations in data generation. These variations often carry semantic meaning, facilitating the removal of irrelevant factors and reducing sample complexity for downstream learning tasks. In speaker verification, an ideal disentangled representation can isolate time-invariant features (e.g., speaker characteristics) from dynamic information (e.g., speech content). Moreover, downstream tasks such as speech recognition and speaker classification can benefit significantly from these representations by utilizing the separated components to improve representation learning.

Recent studies have investigated disentangled representation learning through frameworks such as variational autoencoders (VAEs) [11], [12] and generative adversarial networks (GANs). Some approaches, like β-VAE [13], proposed new objective functions that constrain the information encoded in content and speaker representations. Models such as SpeechTripleNet [14], AnnealVAE [15], and JointVAE [16] set channel capacity for distinct latent variables to promote disentanglement. InfoGAN [17] divided the latent space and incorporated a mutual information regularization term into the standard GAN loss to enhance disentanglement. Similarly, Mathieu *et al.* [18] partitioned the encoding space into style and content components, employing adversarial training to encourage data points within the same class to share content representations while maintaining diverse style features.

However, Gaussian VAE-based models suffer from several limitations, including poor reconstruction quality and diminished generative performance when handling complex data distributions. These models tend to produce samples that collapse toward the distribution center and frequently fail to achieve effective disentanglement. A major contributing factor is that the training objective often prioritizes optimizing the inference network at the cost of the generative model,

Prof. Man-Wai Mak is the corresponding author. Zhe Li, Man-Wai Mak, and Zezhong Jin are with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong SAR. Jen-Tzung Chien is with the Institute of Electrical and Computer Engineering, National Yang-Ming Chiao Tung University. Mert Pilanci is with the Department of Electrical Engineering at Stanford University, USA. Helen Meng is with the Department of Systems Engineering & Engineering Management at The Chinese University of Hong Kong, Hong Kong SAR.

causing the encoder to over-regularize the decoder [19], [20], [21]. Additionally, it is challenging to balance the retention of rich information in the latent code while ensuring high sampling quality in VAEs [22], [23], [24], [21]. Although GAN-based models are powerful, they are difficult to train and require careful hyperparameter tuning [25], [26]. GAN-based models are also prone to training instability [27] and mode collapse [28].

Denoising diffusion models have recently demonstrated superior performance in disentangled representation learning, offering more stable training and higher representation fidelity compared to GANs and VAE-based models. Diffusion models address the challenge of representing complex, high-dimensional probability distributions by decomposing the problem into T incremental steps. At each step, the model transforms the noise data from a simpler distribution (e.g., the simplest Gaussian prior at $t = T$) to a more complex one (e.g., the real data distribution at $t = 0$). This iterative inference and denoising paradigm enables the model to map a simple distribution to a complex one through gradual refinement over many steps. However, the latent variables produced by diffusion models often lack high-level semantics and other desirable properties, such as speaker features.

To overcome the aforementioned limitations, we condition a denoising diffusion implicit model (DDIM) [29] on speaker features and propose a disentangled sequential model that leverages the capabilities of the DDIM to learn multi-level representations. Specifically, we employ a learnable speaker encoder to capture utterance-level speaker characteristics while the DDIM decodes and models the Gaussian variations in data. A latent vector represents the speaker features, and additional frame-level latent vectors capture dynamic information such as speech content. The DDIM's forward and generative processes are conducted within the joint latent space of speaker features and speech content.

We applied our proposed disentanglement method to the speech representations generated by the pre-trained models WavLM [30] and HuBERT [31]. This paper substantially extends our earlier work in [32]. Compared to the earlier conference version, we conducted comprehensive experiments on two speaker verification datasets, VoxCeleb and CN-Celeb, demonstrate that our method effectively extracts accurate speaker embeddings. Even in cases of language mismatches, the model continues to produce discriminative speaker embeddings. Additionally, we performed contrastive learning experiments, showing that without relying on classification loss, our disentanglement method enables the contrastive loss to learn a feature space in which embeddings of the same speaker are compact.

The contributions of this paper are summarized as follows:

1) We propose a disentangled latent diffusion autoencoder (DLD–AE) that aims at separating static (speaker characteristics) and dynamic (speech content) factors in sequential audio data. Our method extends the sequential VAE framework by incorporating a diffusion process into the context modeling, which enhances the learning of the speaker discriminative space.
2) We implement the DDIM within the joint latent space of speaker and speech features, which simplifies and accelerates the denoising process. Conditioning on speaker features allows for generating meaningful latent vectors for decoding.
3) We demonstrate that removing content information from a pre-trained Transformer model enhances the robustness of the embedding vectors against language mismatches.

II. Background

Denoising diffusion probabilistic models (DDPMs) and score-based generative models represent generative approaches that model a target distribution by reversely removing noise at different noise levels. In the denoising process, a Gaussian noise sample from a prior distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ are iteratively refined through T denoising steps to reconstruct a clean sample. Ho *et al.* [33] introduces a noise approximator $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$, which takes in a noisy input $\boldsymbol{x}_t$ at step t and uses a U-Net to predict the noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ that was added to the clean data $\boldsymbol{x}_0$ to yield $\boldsymbol{x}_t$. The training objective involves minimizing the discrepancy $||\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{\epsilon}||_2^2$. This framework effectively simplifies the variational lower bound for the marginal log-likelihood, leading to the widespread adoption of DDPMs in recent studies [34], [35].

In this work, we introduce a diffusion model in the latent space of an VAE-style model, where we define a Gaussian diffusion process at step t (out of a total of T steps) that progressively introduces noise into the input latent audio representation $\boldsymbol{z}_0$. Specifically, the forward diffusion process is defined as

$$q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = \mathcal{N}(\boldsymbol{z}_t; \sqrt{1-\beta_{t-1}}\boldsymbol{z}_{t-1}, \beta_t \boldsymbol{I}), \tag{1}$$

where β_t is a hyperparameter representing the noise level at step t. As a result of this Gaussian diffusion, the noisy version of the original audio $\boldsymbol{z}_0$ at step t is given by another Gaussian:

$$q(\boldsymbol{z}_t|\boldsymbol{z}_0) = \mathcal{N}(\boldsymbol{z}_t; \sqrt{\alpha_t}\boldsymbol{z}_0, (1-\alpha_t)\boldsymbol{I}), \tag{2}$$

where $\alpha_t = \prod_{s=1}^{t}(1-\beta_s)$. We aim to learn the reverse process, which involves estimating the distribution $p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t)$. This distribution is complex unless the gap between $t-1$ and t is infinitesimally small, i.e., $T \to \infty$. In this extreme case, the distribution $p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t)$ can be modeled as $\mathcal{N}(\boldsymbol{z}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{z}_t, t), \boldsymbol{\sigma}_\theta(\boldsymbol{z}_t, t))$ [33]. Among the methods to approximate this distribution, an effective approach is to use $\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t)$, which we discussed earlier. In practice, the assumption of $T \to \infty$ is not achievable, implying that a DDPM can only produce approximate representations.

As a latent-variable model, a DDPM produces latent variables $\boldsymbol{z}_{1:T}$ during the forward diffusion process. However, these variables are stochastic and encapsulate Gaussian noise, representing a gradual degradation of the audio rather than containing significant semantic content. To address this limitation, Song *et al.* [29] introduced a variant known as DDIM, which uses $\boldsymbol{\epsilon}_\theta(\boldsymbol{z}, t)$ to generate $\boldsymbol{z}_{t-1}$ from $\boldsymbol{z}_t$:

$$\boldsymbol{z}_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{\boldsymbol{z}_t - \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t) + \sigma_t\boldsymbol{\epsilon}_t, \tag{3}$$

where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is standard Gaussian noise that is independent of $\boldsymbol{z}_t$. In the special case when $\sigma_t = 0$ for all t in Eq. 3, the reverse process is deterministic. Unlike traditional DDPMs, a DDIM utilizes a deterministic approach for the generative (reverse) process, providing more structured control over the diffusion and denoising stages.

The DDIM maintains the marginal distribution in the DDPM, i.e., $q(\boldsymbol{z}_t|\boldsymbol{z}_0) = \mathcal{N}(\boldsymbol{z}_t; \sqrt{\alpha_t}\boldsymbol{z}_0, (1-\alpha_t)\boldsymbol{I})$. In this way, the DDIM shares the same objective function as the DDPM, differing only in the sample generation methods. Specifically, in DDIM, the reverse diffusion process $p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t)$ is implemented by a Gaussian distribution with zero covariance (a deterministic process) using the predicted noise $\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t)$, i.e., Eq. 3 with $\sigma_t = 0$.

Using the DDIM framework, the generative (reverse) process becomes deterministic. From this perspective, the DDIM functions as a decoder, mapping the latent variable $\boldsymbol{z}_T$ back to the original input sample $\boldsymbol{z}_0$. Although this approach provides an accurate input reconstruction, $\boldsymbol{z}_T$ lacks high-level semantic features that characterize a meaningful representation. This observation motivates the development of novel strategies to enhance the DDIM, incorporating mechanisms that enrich the semantic content of their latent variables, as we propose in this work.

III. Methodology

We denote the input speech sequence as $\boldsymbol{x}_{1:N} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)$, where $\boldsymbol{x}_i$ is the filter-bank feature vector corresponding to the i-th frame, and N is the number of frames in the sequence. To facilitate an informative global latent representation $\boldsymbol{z}_T$ for the decoding process, we introduce a conditional DDIM decoder, represented by $p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t, \boldsymbol{f}_s)$. This decoder is conditioned on an auxiliary latent variable $\boldsymbol{f}_s$ obtained through a speaker encoder $\boldsymbol{f}_s = \text{Enc}_\phi(\boldsymbol{x}_{1:N})$, which maps the entire input sequence $\boldsymbol{x}_{1:N}$ to speaker representation $\boldsymbol{f}_s$. $\boldsymbol{f}_s$ is fed into two linear heads to produce the mean vector $\boldsymbol{\mu}_s$ and standard deviation vector $\boldsymbol{\sigma}_s$. Then, the speaker vector $\boldsymbol{s}$ is obtained by sampling the Gaussian distribution defined by these mean and standard deviation vectors. The module "Switch" in Fig. 1 changes the dimension of $\boldsymbol{s}$ by repeating it N times so that the resulting matrix can be concatenated with dynamic content latents $\boldsymbol{c}_{1:N}$. We refer to the network in Fig. 1 as **D**isentengled **L**atent **D**iffusion–**Auto****E**ncoder (DLD–AE).

A. Speaker Encoder

We utilize an ECAPA-TDNN model [36] to transform the input speech sequence $\boldsymbol{x}_{1:N}$ to a representative vector $\boldsymbol{f}_s$. This vector captures crucial speaker information for the DDIM decoder (the yellow boxes in Fig. 1), expressed as $p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t, \boldsymbol{f}_s)$, to perform the denoising process and predict the output latent vector $\tilde{\boldsymbol{z}}_0 \equiv (\boldsymbol{z}_0, \boldsymbol{f}_s)$. By conditioning DDIM on an enriched information vector $\boldsymbol{f}_s$, we enhance the efficiency and accuracy of the denoising operation, ultimately leading to a more reliable generation of latent representations.

B. Content Encoder

We employ an LSTM with two linear heads as the content encoder to transform $\boldsymbol{x}_{1:N}$ into $\boldsymbol{c}_{1:N}$, where $\boldsymbol{c}_i$ denotes the dynamic state learned at frame i. We assume that each $\boldsymbol{c}_i$ depends on the preceding dynamic variables, denoted as $\boldsymbol{c}_{<i} \equiv \{\boldsymbol{c}_0, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_{i-1}\}$, with $\boldsymbol{c}_0 = \mathbf{0}$.

C. Reverse Diffusion Process

Our proposed conditional DDIM's reverse process utilizes the input $\tilde{\boldsymbol{z}}_{t-1} \equiv (\boldsymbol{z}_t, \boldsymbol{f}_s)$, which comprises the DDIM encoder's output and speaker representation, to generate an output latent vector. Using a denoising U-Net, each block of the conditional DDIM decoder models the probability distribution $p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t, \boldsymbol{f}_s)$:

$$p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t, \boldsymbol{f}_s) = \begin{cases} \mathcal{N}(\boldsymbol{z}_{t-1}; f_\theta(\boldsymbol{z}_1, 1, \boldsymbol{f}_s), \sigma_1^2\boldsymbol{I}) & \text{if } t = 1, \\ q_\sigma(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t, f_\theta(\boldsymbol{z}_t, t, \boldsymbol{f}_s)) & \text{otherwise} \end{cases}, \tag{4}$$

where σ_1 is set to 0. Following the approach in Song *et al.* [29], the inference distribution q_σ in Eq. 4 is defined as follows:

$$q_\sigma = \mathcal{N}\Bigg(\boldsymbol{z}_{t-1}; \sqrt{\alpha_{t-1}} f_\theta(\boldsymbol{z}_t, t, \boldsymbol{f}_s) + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \frac{\boldsymbol{z}_t - \sqrt{\alpha_t} f_\theta(\boldsymbol{z}_t, t, \boldsymbol{f}_s)}{\sqrt{1-\alpha_t}}, \sigma_t^2\boldsymbol{I}\Bigg) \tag{5}$$

where σ_t is set to 0. We implement f_θ in Eqs. 4 and 5 using a noise prediction network $\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{f}_s)$:

$$f_\theta(\boldsymbol{z}_t, t, \boldsymbol{f}_s) \equiv \text{Denoise}(\boldsymbol{z}_t, t, \boldsymbol{f}_s) = \frac{\boldsymbol{z}_t - \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{f}_s)}{\sqrt{\alpha_t}}, \tag{6}$$

where $\boldsymbol{\epsilon}_\theta$ is implemented by a U-Net as shown in Fig 1.

The training process involves optimizing the $\mathcal{L}_{DDIM}$ loss with respect to parameters θ and ϕ:

$$\mathcal{L}_{DDIM} = \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{z}_0, \boldsymbol{\epsilon}_t}\left[\, \| \, \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{f}_s) - \boldsymbol{\epsilon}_t \, \|_2^2 \,\right], \tag{7}$$

where $\boldsymbol{f}_s = \text{Enc}_\phi(\boldsymbol{x}_{1:N})$, $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, $\boldsymbol{z}_t = \sqrt{\alpha_t}\boldsymbol{z}_0 + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t$, and T is an integer, e.g., 100. Note that this simplified loss function optimizes the DDIM but does not optimize the actual variational lower bound.

D. Disentangled Sequential Variational Autoencoder

We define the global latent representation as $\boldsymbol{z}_0 \equiv (\boldsymbol{s}, \boldsymbol{c}_{1:N})$. Our formulation is based on the intuition that sequence variations can be decomposed into time-dependent dynamic components $\{\boldsymbol{c}_i\}$ and a static component $\boldsymbol{s}$. We assume independence between the static variable $\boldsymbol{s}$ and the dynamic variables $\boldsymbol{c}_{1:N}$, implying $p(\boldsymbol{z}_0) = p(\boldsymbol{s}, \boldsymbol{c}_{1:N}) = p(\boldsymbol{s})p(\boldsymbol{c}_{1:N})$. The static component remains constant across all frames within a given utterance but differs across different utterances.

In a speech signal, the phonetic transcription governs the movement of the vocal tract and the produced sounds over time, while the speaker's identity remains fixed throughout an

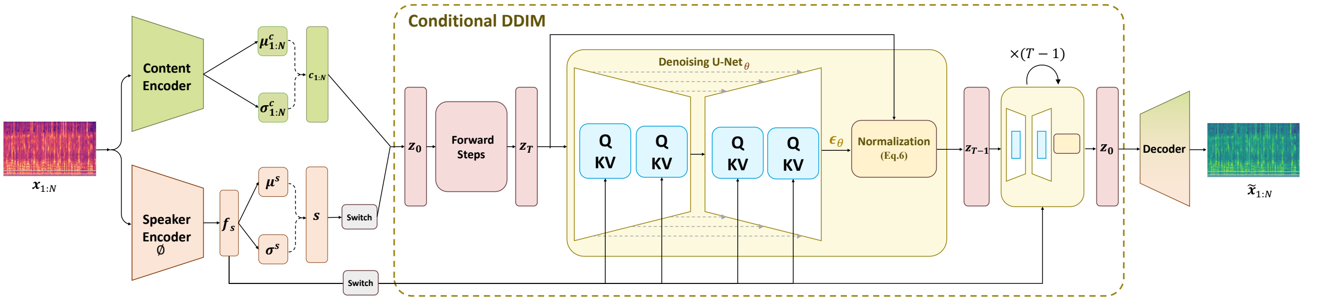


Fig. 1: The autoencoder comprises a speaker encoder, a content encoder, a conditional DDIM, and a speech decoder. The speaker encoder utilizes an ECAPA-TDNN [36] to transform the input speech $\boldsymbol{x}_{1:N}$ into a speaker representation $\boldsymbol{f}_s$, which is further transformed to $\boldsymbol{\mu}^s$ and $\boldsymbol{\sigma}^s$ through two linear heads. Similarly, $\boldsymbol{\mu}^c_{1:N}$ and $\boldsymbol{\sigma}^c_{1:N}$ can be obtained from a long short-term memory (LSTM) network with two linear heads. The "Switch" module changes the dimension of input vectors. For notational simplicity, we use the same symbols before and after the change of dimension. The dotted brace represents Gaussian sampling, which is performed by a reparameterization trick [37]. A conditional DDIM that serves as both a stochastic encoder $\boldsymbol{z}_0 \rightarrow \boldsymbol{z}_T$ and a deterministic decoder $\boldsymbol{z}_{t-1} = \text{Denoise}(\boldsymbol{z}_t, \boldsymbol{f}_s, t)$. $\boldsymbol{z}_0 \in \mathbb{R}^{2D\times N}$, where D is the dimension of $\boldsymbol{c}_i$ and $\boldsymbol{s}$. Similarly, $\boldsymbol{z}_T, \boldsymbol{z}_{T-1}, \ldots, \boldsymbol{z}_1$ have the same dimensions as $\boldsymbol{z}_0$.

utterance. Based on these assumptions, we derive the following complete likelihood [38]:

$$\begin{aligned} p(\boldsymbol{x}_{1:N}, \boldsymbol{z}_0) &= p(\boldsymbol{x}_{1:N}, \boldsymbol{s}, \boldsymbol{c}_{1:N}) \\ &= p(\boldsymbol{s}, \boldsymbol{c}_{1:N}) p(\boldsymbol{x}_{1:N}|\boldsymbol{s}, \boldsymbol{c}_{1:N}) \\ &= p(\boldsymbol{s}) \left[\prod_{i=1}^{N} p(\boldsymbol{c}_i|\boldsymbol{c}_{<i}) p(\boldsymbol{x}_i|\boldsymbol{s}, \boldsymbol{c}_i) \right]. \end{aligned} \tag{8}$$

We define $p(\boldsymbol{s})$ in Eq. 8 as a standard Gaussian $\mathcal{N}(\boldsymbol{s}; \boldsymbol{0}, \boldsymbol{I})$. We assume that $p(\boldsymbol{c}_i|\boldsymbol{c}_{<i})$ follows a Gaussian distribution:

$$p(\boldsymbol{c}_i|\boldsymbol{c}_{<i}) = \mathcal{N}(\boldsymbol{c}_i; \boldsymbol{\mu}_i(\boldsymbol{c}_{<i}), \text{diag}((\boldsymbol{\sigma}_i(\boldsymbol{c}_{<i}))^2)), \tag{9}$$

where $\boldsymbol{\mu}_i(\cdot)$ and $\boldsymbol{\sigma}_i(\cdot)$ can be modeled by an LSTM followed by two linear heads. Since both $\boldsymbol{\mu}_i(\cdot)$ and $\boldsymbol{\sigma}_i(\cdot)$ are conditioned on the temporal context $\boldsymbol{c}_{<i}$, their derivation at frame i requires access to the history $\boldsymbol{c}_{<i}$. To sample $\boldsymbol{c}_i$ from $p(\boldsymbol{c}_i|\boldsymbol{c}_{<i})$, $\boldsymbol{c}_{i-1}$ is first passed through the LSTM cells to forward one step, generating $\boldsymbol{\mu}_i(\cdot)$ and $\boldsymbol{\sigma}_i(\cdot)$ via linear transformation layers. The reparameterization trick is then applied to draw a sample from the resulting distribution [39], [40].

To derive latent representations solely from the observed data $\boldsymbol{x}_{1:N}$, where speaker characteristics and content are entangled, we aim to learn a posterior distribution $q(\boldsymbol{z}_0|\boldsymbol{x}_{1:N})$ that disentangles these two components. Specifically, we use variational inference:

$$\begin{aligned} q(\boldsymbol{z}_0|\boldsymbol{x}_{1:N}) &= q(\boldsymbol{c}_{1:N}, \boldsymbol{s}|\boldsymbol{x}_{1:N}) \\ &= q(\boldsymbol{c}_{1:N}|\boldsymbol{x}_{1:N}) q(\boldsymbol{s}|\boldsymbol{x}_{1:N}) \\ &= q(\boldsymbol{s}|\boldsymbol{x}_{1:N}) \prod_{i=1}^{N} q(\boldsymbol{c}_i|\boldsymbol{c}_{<i}, \boldsymbol{x}_{1:N}). \end{aligned} \tag{10}$$

The speaker latent posterior follows a Gaussian distribution:

$$q\left(\boldsymbol{s}|\boldsymbol{x}_{1:N}\right) = \mathcal{N}\left(\boldsymbol{s}; \boldsymbol{\mu}^s(\boldsymbol{x}_{1:N}), \text{diag}\{(\boldsymbol{\sigma}^s(\boldsymbol{x}_{1:N}))^2\}\right), \tag{11}$$

where the mean and standard deviation vectors, $\boldsymbol{\mu}^s(\cdot)$ and $\boldsymbol{\sigma}^s(\cdot)$, are modeled by an ECAPA-TDNN [36] with two linear layers. Similarly, we define:

$$q\left(\boldsymbol{c}_i|\boldsymbol{c}_{<i}, \boldsymbol{x}_{1:N}\right) = \mathcal{N}\left(\boldsymbol{c}_i; \boldsymbol{\mu}^c_i(\boldsymbol{x}_{1:N}, \boldsymbol{c}_{<i}), \text{diag}\{(\boldsymbol{\sigma}^c_i(\boldsymbol{x}_{1:N}, \boldsymbol{c}_{<i}))^2\}\right), \tag{12}$$

where $\boldsymbol{\mu}^c_i(\cdot)$ and $\boldsymbol{\sigma}^c_i(\cdot)$ are obtained by passing the inputs $\boldsymbol{c}_{<i}$ and $\boldsymbol{x}_{1:N}$ through bidirectional LSTMs, followed by an RNN and two linear layers. The reparameterization trick is applied to sample $\boldsymbol{s}$ and $\{\boldsymbol{c}_i\}_{i=1}^N$.

Previous studies have introduced similar parameterizations of dynamic variables through recurrent networks [39], [40]. The standard approach for learning latent representations is to maximize the evidence lower bound (ELBO) [11], [41]:

$$\max_{p,q} \mathbb{E}_{\boldsymbol{x}_{1:N}\sim p_{\mathrm{D}}(\boldsymbol{x}_{1:N})} \left[\underbrace{\mathbb{E}_{q(\boldsymbol{z}_0|\boldsymbol{x}_{1:N})} \log p(\boldsymbol{x}_{1:N}|\boldsymbol{z}_0)}_{\text{reconstruction term}} - \underbrace{\text{KL}[q(\boldsymbol{z}_0|\boldsymbol{x}_{1:N}) \parallel p(\boldsymbol{z}_0)]}_{\text{prior matching term}} \right], \tag{13}$$

where $p_{\mathrm{D}}(\boldsymbol{x}_{1:N})$ represents the empirical data distribution and $\text{KL}[\cdot||\cdot]$ denotes Kullback-Leibler (KL) divergence. Under the assumption that $\boldsymbol{s}$ and $\boldsymbol{c}_{1:N}$ are mutually independent in the posterior, the KL-divergence term is simplified as

$$\begin{aligned} &\text{KL}[q(\boldsymbol{z}_0|\boldsymbol{x}_{1:N}) \parallel p(\boldsymbol{z}_0)] = \\ &\text{KL}[q(\boldsymbol{s}|\boldsymbol{x}_{1:N}) \parallel p(\boldsymbol{s})] + \text{KL}[q(\boldsymbol{c}_{1:N}|\boldsymbol{x}_{1:N}) \parallel p(\boldsymbol{c}_{1:N})], \end{aligned} \tag{14}$$

where the second term is approximated using sampled trajectories of the dynamic variables $\boldsymbol{c}_{1:N}$.

We define the **d**isentangled **s**equential **v**ariational **a**uto**e**ncoder (**DSVAE**) loss as the negative ELBO of the log-likelihood:

$$\begin{aligned} \mathcal{L}_{DSVAE} = &-\mathbb{E}_{p_{\mathrm{D}}(\boldsymbol{x}_{1:N})} \mathbb{E}_{q(\boldsymbol{z}_0|\boldsymbol{x}_{1:N})} \left[\log p(\boldsymbol{x}_{1:N}|\boldsymbol{z}_0) \right] \\ &+ \text{KL}\big[q(\boldsymbol{s}|\boldsymbol{x}_{1:N}) \parallel p(\boldsymbol{s})\big] \\ &+ \text{KL}\big[q(\boldsymbol{c}_{1:N}|\boldsymbol{x}_{1:N}) \parallel p(\boldsymbol{c}_{1:N})\big]. \end{aligned} \tag{15}$$

The first term in Eq. 15 corresponds to the reconstruction loss, while the next two terms correspond to the KL divergence between the posterior and prior distributions of the time-variant content embeddings $\{\boldsymbol{c}_i\}_{i=1}^N$ (Eq. 10 and 12) and the time-invariant speaker embeddings $\boldsymbol{s}$ (Eq. 11), respectively. Specifically, the reconstruction loss is computed through the mean squared error (MSE) between the decoder outputs and inputs. The KL divergence terms can be computed analytically

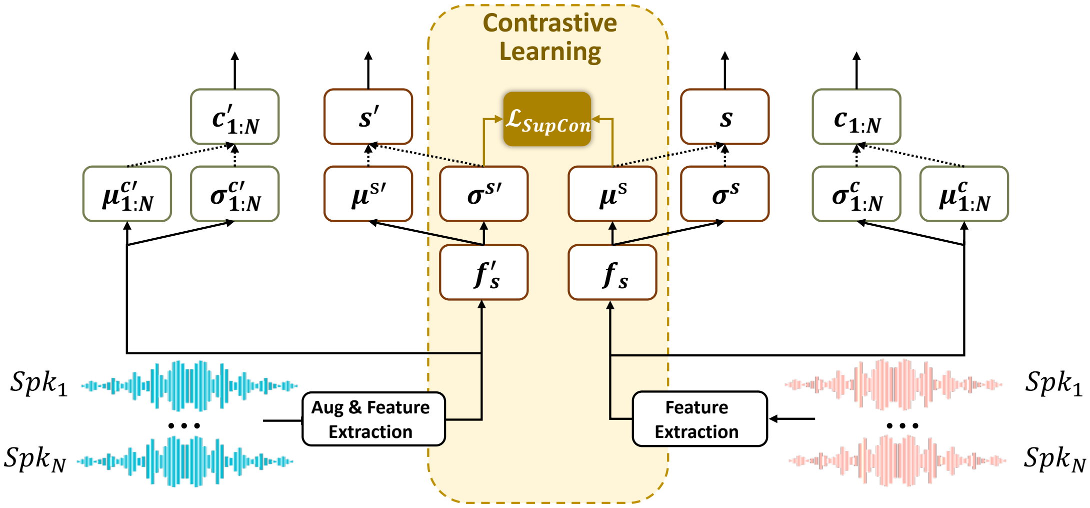


Fig. 2: The schematic illustrates the proposed contrastive speaker embedding approach with sequential disentanglement. “Aug” refers to the augmentation performed on the speech. Feature extraction involves the process of Fbank extraction and representation learning. $\boldsymbol{f}'_s$ is the speaker feature augmentation of $\boldsymbol{f}_s$ and $\boldsymbol{c}'_{1:N}$ is the content augmentation of $\boldsymbol{c}_{1:N}$. $\boldsymbol{f}'_s$ can be seen as the positive sample for the anchor $\boldsymbol{f}_s$. The dashed arrows inside the speaker encoder and content encoder indicate Gaussian sampling, performed using the reparameterization trick [37]. After training, the vectors produced by the $\boldsymbol{f}_s$ are utilized as speaker embeddings.

as both the priors and posteriors of $\boldsymbol{s}$ and $\{\boldsymbol{c}_i\}_{i=1}^N$ are assumed to be Gaussian distributed [42].

E. Contrastive Speaker Embedding with Sequential Disentanglement

We combine the sequential disentanglement method in [38] and contrastive learning in [43], [44], [45] to enhance the contrast between the speaker embeddings of different speakers without interfered by the content variations in their utterances. This method incorporates a disentanglement technique into a contrastive learning framework, adopting a SimCLR-like structure with a supervised contrastive loss. As depicted in Fig. 2, we utilize the supervised contrastive loss to distinguish positive examples of a particular class from negative examples belonging to other classes, using the provided labels. The original and augmented speaker embeddings are incorporated into the supervised contrastive loss:

$$\mathcal{L}_{SupCon} = \sum_{i=1}^{N} \frac{-1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(\mathrm{sim}(\boldsymbol{f}_s^i, \boldsymbol{f}_s^p)/\tau)}{\sum_{a \in \mathcal{A}(i)} \exp(\mathrm{sim}(\boldsymbol{f}_s^i, \boldsymbol{f}_s^a)/\tau)}, \tag{16}$$

where $\mathrm{sim}(\cdot)$ represents the cosine similarity. In Eq. 16, $\boldsymbol{f}_s^i$ is the anchor, $\boldsymbol{f}_s^a$ is a negative sample, $\mathcal{A}(i)$ denotes the set of negative sample indices relative to $\boldsymbol{f}_s^i$, $\boldsymbol{f}_s^p$ is a positive sample with respect to $\boldsymbol{f}_s^i$, and $\mathcal{P}(i)$ comprises the indices of positive samples within the augmented batch (comprising both original and augmented data). The scalar temperature parameter is denoted as $\tau \in \mathbb{R}^+$.

F. Model Training

To ensure a meaningful condition for the speaker embedding $\boldsymbol{s}$, we optimize the speaker encoder using AAM-Softmax [46]. To train the network, we define the total loss: the AAM-Softmax [46], DDIM loss (Eq. 7), and DSVAE loss (Eq. 15). The last one can be treated as regularization. The combination can be implemented as follows:

$$\mathcal{L}_{DLDAE} = \mathcal{L}_{AAM\text{-}Softmax} + \mathcal{L}_{DDIM} + \lambda \mathcal{L}_{DSVAE}. \tag{17}$$

In our contrastive learning experiment, we incorporate the SupCon loss (Eq. 16). Thus, the total loss becomes:

$$\mathcal{L}_{DLDAE\text{-}CL} = \mathcal{L}_{SupCon} + \mathcal{L}_{DDIM} + \lambda \mathcal{L}_{DSVAE}. \tag{18}$$

In Eq. 17 and Eq. 18, λ is a hyperparameter that regulates the impact of sequential disentanglement. During inference, only the speaker encoder is used to extract speaker embeddings.

IV. Experiments and Results

A. Implementation Details

We evaluated our proposed method on the CN-Celeb [47], [48] and VoxCeleb [49], [50] datasets. HuBERT Large [31] and WavLM Large [30] were chosen as the pre-trained models for extracting frame-level acoustic features, and SpecAugment [51] was applied to these features to create augmented features. The speaker encoder used in our experiments was ECAPA-TDNN [36]. We employed various augmentation techniques by following the Kaldi’s recipe [52], including adding noise, music, and background chatter using the MUSAN dataset [53]. Furthermore, we introduced reverberation effects by convolving the original waveforms with room impulse responses (RIR) from the RIR dataset [54]. Each augmentation type has a 60% chance of being selected.

Each training utterance was truncated to 3 seconds, and we used mini-batches of 256 utterances for training. AAM-Softmax [46] was employed with a margin of 0.2 and a scaling

TABLE I: Performance of the baseline models and the proposed DLD–AE on VoxCeleb1-O, VoxCeleb1-E, and VoxCeleb1-H. All experiments used ECAPA-TDNN as the speaker encoder and were trained on VoxCeleb2-dev. Results were obtained without AS-Norm [57], [58] nor quality-aware score calibration [59]. For RecXi, the results are based on the setting RecXi($\tilde{\phi}$, $\tilde{\phi}_{\text{lin}}$) in [56].

Row	Input Feature	Disentanglement Method	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
			EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
1	Fbank	None	1.12	0.145	1.25	0.142	2.43	0.239
2		RecXi + $\mathcal{L}_{\text{ssp}}$ [56]	1.19	0.107	1.29	0.141	2.46	0.227
3		DLD–AE (Ours)	1.01	0.101	1.28	0.153	2.35	0.213
4	HuBERT	None	0.91	0.119	0.99	1.146	2.35	0.252
5		DLD–AE (Ours)	0.88	0.088	**0.91**	1.011	2.05	0.231
6	WavLM	None	0.85	0.113	1.12	0.091	2.06	0.197
7		DLD–AE (Ours)	**0.78**	**0.081**	**0.91**	**0.090**	**1.83**	0.191

factor of 30. The learning rate was decreased by 3% after every epoch. We used the Adam optimizer combined with the CosineAnnealingWarmRestarts [55] learning rate scheduling strategy.

For experiments utilizing pre-trained models, we initially trained the speaker encoder with AAM-Softmax [46] for 20 epochs, followed by additional training using the objectives defined in Eq. 7 and Eq. 15. For contrastive learning experiments, the speaker encoder was first trained with Eq. 16 for 100 epochs and then further trained with Eq. 7 and Eq. 15.

B. Comparing with Existing Methods

To evaluate the effectiveness of our proposed DLD–AE, we compared its performance with existing disentanglement techniques. Table I shows the performance of models trained on VoxCeleb2 and tested on VoxCeleb1. Results are based on Fbank features and features extracted from the pre-trained models, with ECAPA-TDNN serving as the speaker embedding network.

RecXi [56] is a disentanglement framework designed to simultaneously model speaker traits and content variability in speech. This framework employs three Gaussian inference layers, each including a learnable transition model to extract distinct speech components. Additionally, a self-supervised loss, denoted as $\mathcal{L}_{ssp}$, is introduced to disentangle content dynamically using speaker identity labels only.

As shown in Table I, when using Fbank features, our DLD–AE (row 3) outperforms the baseline ECAPA-TDNN (row 1) and achieves competitive results compared to RecXi (row 2). This demonstrates the effectiveness of our disentanglement framework in improving speaker verification performance. The results also show that our disentanglement technique is particularly effective when applied to pre-trained features, including HuBERT and WavLM features. For example, with the WavLM features, DLD–AE (row 7) reduces the EER to 0.78% on VoxCeleb1-O, compared to 0.85% without disentanglement (row 6). A similar trend is observed for minDCF. This improvement is attributed to our framework's ability to disentangle static speaker components, enhancing speaker recognition effectively. The improvement highlights the importance of modeling the dynamic contents in speech and disentangling the speaker and content representations.

TABLE II: Performance of the baseline models and the proposed DLD–AE on VoxCeleb1-O using PTMs as frame-level feature extraction. The training dataset is VoxCeleb1-dev.

Row	Feature	System	VoxCeleb1-O	
			EER(%)	minDCF
1	Whisper	ECAPA-TDNN [62]	2.92	0.391
2		Whisper-SV [62]	2.22	0.307
3	wav2vec	EF-wav2vec-large-finetune [63]	3.42	-
4	HuBERT	EF-HuBERT-large-finetune [63]	2.36	-
5		ECAPA-TDNN	2.05	0.270
6		ECAPA-TDNN + DLD–AE	1.88	0.202
7	WavLM	ECAPA-TDNN	1.73	0.221
8		ECAPA-TDNN + DLD–AE	**1.61**	**0.142**

Table II shows the performance of our method and other state-of-the-art methods when using pre-trained models (PTMs) as frame-level feature extractors. Evidently, among all the PTMs, WavLM is the best frame-level feature extractor, followed by HuBERT and wav2vec.

Unlike our previous work on parameter-efficient fine-tuning [60], [61], we employed the Adam optimizer with the CosineAnnealingWarmRestarts [55] learning rate scheduler in this study. This combination results in enhanced performance. The proposed DLD–AE can effectively extract speaker features from HuBERT and WavLM features, highlighting its capacity for disentangling speaker representations from content representations.

C. Robustness to Language Mismatches

We also performed experiments on the CN-Celeb dataset. Comparing row 2 with row 1 and row 4 with row 1 in Table III reveals that utilizing English pre-trained models to extract input features for the speaker encoder can enhance the performance of Mandarin speaker verification. Additionally, the comparison between row 3 and row 2, as well as row 5 and row 4, demonstrates that the DLD–AE, which retains only speaker information, results in performance improvements.

The proposed method aims to mitigate the influence of content on pre-trained speech features. While removing content information may reduce phonotactic information in speech sequences, potentially affecting spoken language recognition, it enables the generation of language-invariant speaker embeddings.

D. Ablation Study

We conducted ablation experiments to investigate the importance of different components in the proposed DLD–AE.

TABLE III: Performance of the baseline models and the proposed DLD–AE on CN-Celeb. All experiments used ECAPA-TDNN as the speaker encoder and were trained on CN-Celeb1&2.

Row	Input Feature	Disentanglement Method	CN-Celeb	
			EER(%)	minDCF
1	Fbank	None	8.93	0.509
2	HuBERT	None	8.89	0.504
3		DLD–AE	8.57	0.467
4	WavLM	None	8.65	0.471
5		DLD–AE	**8.05**	**0.436**

For these experiments, we used VoxCeleb2 as the training set, VoxCeleb1-test as the testing set, and we used CN-Celeb1&2 as the training set and CN-Celeb1 as the test set. We also conducted experiments using DSVAE to perform the disentanglement, which is essentially a VAE-based disentanglement without the diffusion process. Results are shown in Table IV.

Comparing Row 6 with Row 5 in Table IV reveals that adding the VAE can slightly improve performance. However, a significant performance gain is observed when integrating the diffusion processes into the VAE (row 7). The best performance is achieved when the diffusion processes are conditioned on the speaker embeddings (row 8). The same conclusions are obtained regardless of which pre-trained models were used.

E. Comparison with ASR- or Contrastive Learning-based Systems

As mentioned in Section I, pre-trained ASR models can benefit speaker recognition tasks. We compare our proposed method with systems that leverage pre-trained ASR models in two aspects: (1) using ASR models to provide phonetic information for speaker recognition [65] and (2) initializing the speaker encoder with pre-trained ASR model weights [6]. We also compare contrastive learning-based disentanglement methods.

As shown in Table V, our method significantly outperforms the systems in rows 1, 2, and 3. The primary reason is that these systems do not use speaker labels. The system in row 4 incorporates visual information and pseudo labels, yet our proposed method performs better than it. Our proposed systems (#10) outperform the systems that use pre-trained ASR models (systems #5 and #6). A key advantage of our method is that it achieves competitive performance without requiring any ASR models.

We also applied the proposed method to supervised contrastive learning. Compared to row 7, the results in row 8 show that our method brings more noticeable improvements to contrastive learning-based approaches. The proposed method achieves the best results when combined with AAM-Softmax [46]. This is because contrastive speaker embedding assumes that the contrast between positive and negative pairs arises from speaker identity rather than other explanatory factors of variation [67], [68], such as linguistic content and languages. However, speaker embeddings often contain various types of information beyond speaker identity [69], [70], and non-speaker factors can also contribute to contrasting positive and negative pairs. This incorrect contrast introduces nuisance information into the embeddings, leading to performance degradation. Therefore, it is essential to disentangle speaker factors from other sources of variation and use only these factors in contrastive learning to ensure that the learned embeddings are speaker discriminative.

F. Impact of λ

The hyperparameter λ in Eq. 17 and Eq. 18 controls the extent of DLD–AE's contribution within the proposed framework. In text-independent speaker verification, text content is considered a nuisance, and ideally, this information should be removed during the embedding learning process. Lin *et al.* [71] proposed a frame shuffling approach to reduce content dependency in positive samples. In contrast, our method explicitly discards content information during contrastive learning to emphasize speaker-specific features. While either shuffling the frames in [71] or disentangling speaker information from content information can improve speaker verification performance, they contradict the methods that incorporate (rather than disentangle) phonetic information into speaker embeddings. For instance, Liu *et al.* [72] employed multi-task learning by integrating a phonetic classifier with a speaker classifier, resulting in improved performance. Similarly, Wang *et al.* [73] utilized phonetic information at both the segment (embedding) and frame levels, demonstrating that while phonetic content at the segment level could hinder SV performance, its use at the frame level was beneficial.

In this subsection, we analyze the impact of varying λ on SV performance. We selected λ ranging from 0.01 to 0.1, incrementing by 0.01 at each step. The results, shown in Fig. 3, indicate that for both EER and minDCF, when WavLM is used as the pre-trained model, the best performance is achieved at $\lambda = 0.01$, while the worst performance occurs at $\lambda = 0.1$. For HuBERT, the optimal result is observed at $\lambda = 0.02$, with performance declining as λ increases. These findings suggest that placing excessive emphasis on sequence decoupling may negatively impact the model's ability to learn discriminative speaker embeddings.

Overall, this analysis indicates that the DLD–AE effectively decouples speaker and content factors, aiding in the extraction of speaker-specific information. However, excessive separation of content information may hinder speaker verification performance.

G. Impact of Diffusion Steps

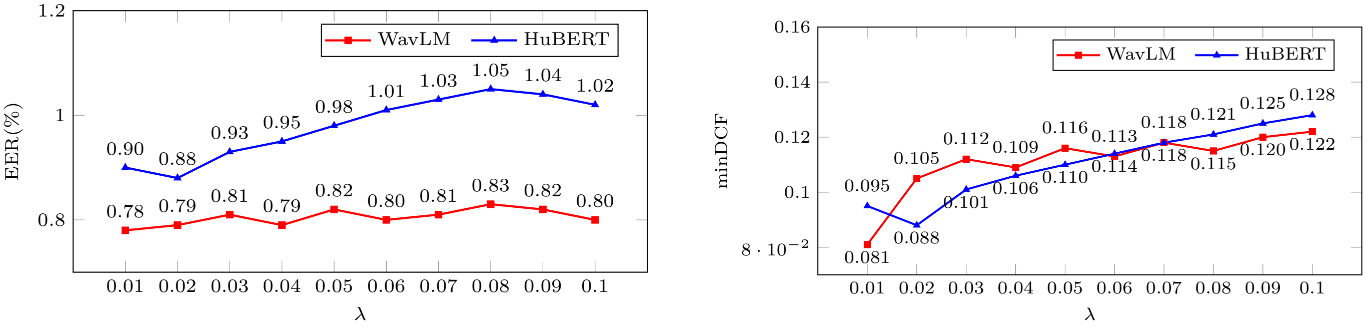
The fast generation speed is a notable advantage of making the denoising process conditions on speaker features. One major reason why standard diffusion models, such as DDPM, require numerous sampling steps is that they can only approximate $p(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t)$ with a Gaussian distribution when T is sufficiently large (typically around 1000). In our work, we employ DDIM for diffusion and denoising, which substantially reduces the number of steps. Unlike standard DDPM, which often requires hundreds or even thousands of iterative steps, DDIM can generate high-quality samples in just a few dozen steps. This efficiency is achieved through an explicit inference

TABLE IV: Performance of the baseline models and the proposed DLD–AE on VoxCeleb1-O by using PTMs as frame-level feature extractors. The speaker encoder is ECAPA-TDNN, and the training datasets for VoxCeleb and CN-Celeb evaluations are VoxCeleb1 and CN-Celeb1&2, respectively.

Row	Feature	Disentanglement Method	VoxCeleb1-O		CN-Celeb	
			EER(%)	minDCF	EER(%)	minDCF
1	HuBERT	None	0.91	0.119	8.89	0.504
2		DSVAE [11]	0.90	0.093	8.84	0.509
3		DLD–AE w/o condition	0.89	0.090	8.79	0.481
4		DLD–AE	0.88	0.088	8.57	0.467
5	WavLM	None	0.85	0.113	8.65	0.471
6		DSVAE [11]	0.83	0.094	8.59	0.471
7		DLD–AE w/o condition	0.81	0.084	8.37	0.452
8		DLD–AE	0.78	0.081	8.05	0.436

TABLE V: Performance of the proposed DLD–AE and the systems that leverage pre-trained ASR models or contrastive learning for SV. The training dataset is VoxCeleb2-dev. Results were obtained without AS-Norm nor quality-aware score calibration.

Row	System	Speaker Labels	Pre-trained ASR Model	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
				EER	minDCF	EER	minDCF	EER	minDCF
1	SimCLR + DSVAE [38]	✗	✗	6.37	0.533	7.36	0.574	11.72	0.677
2	MoCo + DSVAE [38]	✗	✗	6.29	0.534	7.17	0.567	11.42	0.668
3	MCL-DPP [64]	✗	✗	2.89	-	3.17	-	6.27	-
4	MCL-DPP-C [64]	Pseudo label	✗	1.44	-	1.77	-	3.27	-
5	IPA [65]	✓	✓	1.81	-	1.68	-	3.12	-
6	NEMO [6]	✓	✓	0.88	0.137	**1.08**	0.134	2.20	0.225
7	SupCon [66]	✓	✗	2.48	0.278	2.51	0.285	4.76	0.450
8	DLDAE-CL (Eq. 18)	✓	✗	2.41	0.241	2.50	0.240	4.18	0.376
9	AAM-Softmax [46]	✓	✗	1.12	0.145	1.25	0.142	2.43	0.239
10	DLDAE-CL (Eq. 18) + AAM-Softmax [46]	✓	✗	**0.95**	**0.099**	1.09	**0.118**	**2.18**	**0.201**



(a) EER(%)

(b) minDCF

Fig. 3: Results on VoxCeleb1-O for different λ in Eq. 17, using WavLM Large and HuBERT Large as the PTMs. The training dataset is VoxCeleb2-dev.

process that reduces the random noise term, making each step more efficient and accurate. As illustrated in Fig. 4, the optimal performance is achieved with 10 steps, while for HuBERT, the best results are obtained using 20 steps.

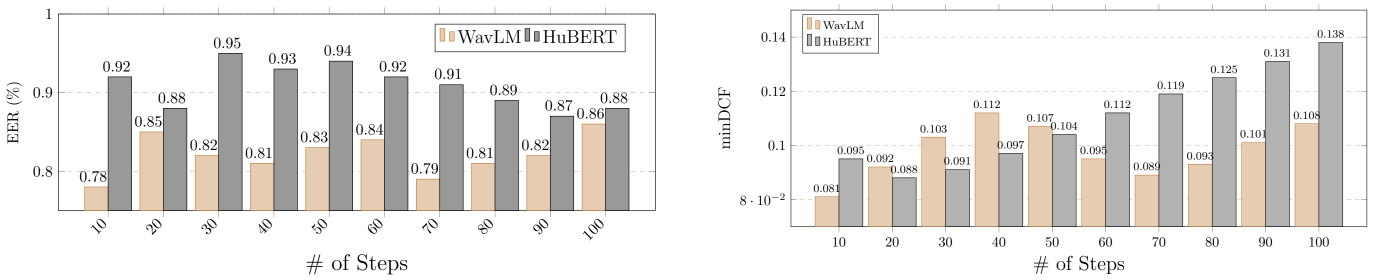
H. Visualization of Speaker Embedding

We validated the effectiveness of the proposed disentanglement method through visualization by analyzing the embedding vectors of 20 speakers from the VoxCeleb1 dataset, with each speaker having 100 utterances. To achieve this, we used t-SNE to project the high-dimensional embedding vectors to a two-dimensional space.

Fig. 5a shows the embeddings obtained using WavLM features, while Fig. 5b presents the visualization results after applying the disentanglement method. Additionally, Figs. 5c, 5d, 5e, 5f provide comparisons of the embedding projected with and without the application of DLD–AE. From these figures, it is evident that the embeddings become more tightly clustered after incorporating DLD–AE. When content information is removed, distinct boundaries emerge between different speakers. This improved clustering effect clearly demonstrates the effectiveness of our model in distinguishing between speakers.

V. Conclusions

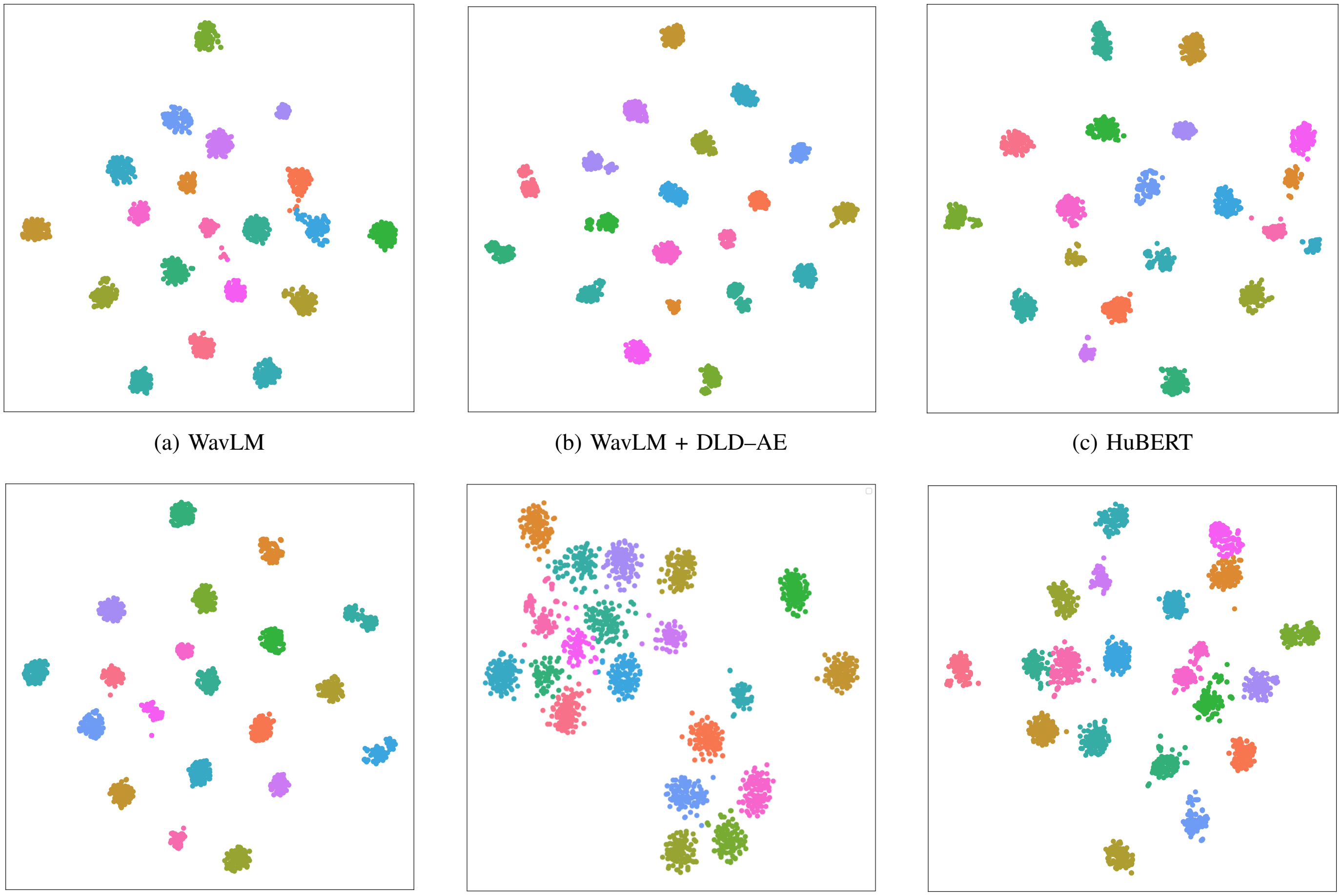
This paper introduces a sequential disentanglement framework based on a latent diffusion model, designed to separate speaker traits from content factors while utilizing only speaker traits for classification. In our experiments, WavLM and HuBERT were employed as pre-trained models to extract speech features. Additionally, we conducted experiments incorporating supervised contrastive learning. Evaluation results on the VoxCeleb1 test set and the CN-Celeb dataset indicate that the proposed method consistently outperforms



(a) EER(%)

(b) minDCF

Fig. 4: Results on VoxCeleb1-O for different numbers of diffusion steps in Eq. 17, using WavLM Large and HuBERT as the PTMs. The training dataset is VoxCeleb2-dev.



(a) WavLM

(b) WavLM + DLD–AE

(c) HuBERT

(d) HuBERT + DLD–AE

(e) SupCon

(f) DLDAE-CL (SupCon+DDIM+DSVAE)

Fig. 5: Visualization of speaker-discriminative ability via t-SNE embeddings. Each color represents a distinct speaker.

traditional baselines, demonstrating the effectiveness of integrating sequential disentanglement with pre-trained models or contrastive learning for obtaining speaker-discriminative embeddings. Furthermore, the results on the CN-Celeb dataset highlight the proposed method's ability to effectively address language mismatch issues.

VI. Acknowledgment

This work was in part supported by the Research Grants Council of Hong Kong, Theme-based Research Scheme (Ref.: T45-407/19-N), a GRF (No. 15228223), and the Research Student Attachment Programme of HKPolyU.

References

[1] A. Tjandra, R. Pang, Y. Zhang, and S. Karita, "Unsupervised learning of disentangled speech content and style representation," in *Proc. of Interspeech*, 2021, pp. 4089–4093.

[2] W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen *et al.*, "Hierarchical generative modeling for controllable speech synthesis," in *Proc. of International Conference on Learning Representations (ICLR)*, 2021.

[3] Z. Jin, Y. Tu, and M.-W. Mak, "Joseph: phonetic-aware speaker embedding for far-field speaker verification," in *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2024.

[4] Q.-B. Hong, C.-H. Wu, and H.-M. Wang, "Decomposition and reorganization of phonetic information for speaker embedding learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1745–1757, 2023.

[5] D. Liao, T. Jiang, F. Wang, L. Li, and Q. Hong, "Towards a unified conformer structure: from asr to asv task," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[6] D. Cai, W. Wang, M. Li, R. Xia, and C. Huang, "Pretraining conformer with asr for speaker verification," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[7] Z. Song, L. He, P. Wang, Y. Hu, and H. Huang, "Introducing multilingual phonetic information to speaker embedding for speaker verification," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 091–10 095.

[8] T. Liu, M. C. Madhavi, R. K. Das, and H. Li, "A unified framework for speaker and utterance verification." in *Proc. of Interspeech*, 2019, pp. 4320–4324.

[9] T. Liu, R. K. Das, M. Madhavi, S. Shen, and H. Li, "Speaker-utterance dual attention for speaker and utterance verification," in *Proc. of Interspeech*, 2020, pp. 4293–4297.

[10] T. Liu, R. K. Das, K. A. Lee, and H. Li, "Neural acoustic-phonetic approach for speaker verification with phonetic attention mask," *IEEE Signal Processing Letters*, vol. 29, pp. 782–786, 2022.

[11] L. Yingzhen and S. Mandt, "Disentangled sequential autoencoder," in *Proc. of International Conference on Machine Learning (ICML)*, 2018, pp. 5670–5679.

[12] J. Bai, W. Wang, and C. P. Gomes, "Contrastively disentangled sequential variational autoencoder," *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 10 105–10 118, 2021.

[13] H. Lu, D. Wang, X. Wu, Z. Wu, X. Liu, and H. Meng, "Disentangled speech representation learning for one-shot cross-lingual voice conversion using ß-vae," in *Proc. of IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 814–821.

[14] H. Lu, X. Wu, Z. Wu, and H. Meng, "Speechtriplenet: End-to-end disentangled speech representation learning for content, timbre and prosody," in *Proc. of ACM International Conference on Multimedia*, 2023, pp. 2829–2837.

[15] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-vae," *arXiv e-prints*, p. arXiv:1804.03599, 2018.

[16] E. Dupont, "Learning disentangled joint continuous and discrete representations," *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.

[17] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, vol. 29, 2016.

[18] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun, "Disentangling factors of variation in deep representation using adversarial training," *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, vol. 29, 2016.

[19] C. Cremer, X. Li, and D. Duvenaud, "Inference suboptimality in variational autoencoders," in *Proc. of International Conference on Machine Learning (ICML)*, 2018, pp. 1078–1086.

[20] Y. Yacoby, W. Pan, and F. Doshi-Velez, "Failure modes of variational autoencoders and their effects on downstream tasks," *arXiv preprint arXiv:2007.07124*, 2020.

[21] X. Chen and S. Li, "Ph-vae: A polynomial hierarchical variational autoencoder towards disentangled representation learning," *arXiv preprint arXiv:2502.02856*, 2025.

[22] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[23] A. Sinha, J. Song, C. Meng, and S. Ermon, "D2C: Diffusion-decoding models for few-shot conditional generation," *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 12 533–12 548, 2021.

[24] O. Rybkin, K. Daniilidis, and S. Levine, "Simple and effective vae training with calibrated decoders," in *Proc. of International Conference on Machine Learning (ICML)*, 2021, pp. 9179–9189.

[25] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for GANs do actually converge?" in *Proc. of International Conference on Machine Learning (ICML)*, 2018, pp. 3481–3490.

[26] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, "Are GANs created equal? a large-scale study," *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.

[27] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. of International Conference on Learning Representations (ICLR)*, 2018.

[28] D. Bang and H. Shim, "MGGAN: Solving mode collapse using manifold-guided training," in *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 2347–2356.

[29] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. of International Conference on Learning Representations (ICLR)*.

[30] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[31] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[32] Z. Li, M.-W. Mak, J.-T. Chien, M. Pilanci, Z. Jin, and H. Meng, "Disentangling speaker and content in pre-trained speech models with latent diffusion for robust speaker verification," *Proc. of Interspeech*, 2025.

[33] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.

[34] J. Xue, Y. Deng, Y. Gao, and Y. Li, "Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation," *arXiv preprint arXiv:2401.01044*, 2024.

[35] H.-Y. Choi, S.-H. Lee, and S.-W. Lee, "DDDM-VC: Decoupled denoising diffusion models with disentangled representation and prior mixup for verified robust voice conversion," in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, 2024, pp. 17 862–17 870.

[36] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. of Interspeech*, 2020, pp. 3830–3834.

[37] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. of International Conference on Learning Representations (ICLR)*, 2014.

[38] Y. Tu, M.-W. Mak, and J.-T. Chien, "Contrastive self-supervised speaker embedding with sequential disentanglement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[39] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, 2015.

[40] A. Goyal, A. Sordoni, M.-A. Côté, N. R. Ke, and Y. Bengio, "Z-forcing: training stochastic recurrent networks," in *Proc. of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2017, pp. 6716–6726.

[41] Y. Zhu, M. R. Min, A. Kadav, and H. P. Graf, "S3VAE: Self-supervised sequential VAE for representation disentanglement and data generation," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6538–6547.

[42] J. Duchi, "Derivations for linear algebra and optimization," *Berkeley, California*, vol. 3, no. 1, pp. 2325–5870, 2007.

[43] Z. Li and M.-W. Mak, "Speaker representation learning via contrastive loss with maximal speaker separability," in *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 962–967.

[44] Z. Li, M.-W. Mak, and H. M.-L. Meng, "Discriminative speaker representation via contrastive learning with class-aware attention in angular

space," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[45] C.-X. Gan, M.-W. Mak, W. Lin, and J.-T. Chien, "Asymmetric clean segments-guided self-supervised learning for robust speaker verification," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 081–11 085.

[46] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4690–4699.

[47] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "CN-Celeb: A challenging chinese speaker recognition dataset," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7604–7608.

[48] L. Li, R. Liu, J. Kang, Y. Fan, H. Cui, Y. Cai, R. Vipperla, T. F. Zheng, and D. Wang, "CN-Celeb: multi-genre speaker recognition," *Speech Communication*, vol. 137, pp. 77–91, 2022.

[49] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," *Proc. of Interspeech 2017*, pp. 2616–2620, 2017.

[50] J. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," *Proc. of Interspeech 2018*, 2018.

[51] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. of Interspeech*, 2019, pp. 2613–2617.

[52] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5796–5800.

[53] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[54] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[55] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," in *Proc. of International Conference on Learning Representations (ICLR)*, 2022.

[56] T. Liu, K. A. Lee, Q. Wang, and H. Li, "Disentangling voice and content with self-supervision for speaker recognition," *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, pp. 50 221–50 236, 2023.

[57] Z. N. Karam, W. M. Campbell, and N. Dehak, "Towards reduced false-alarms using cohorts," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4512–4515.

[58] S. Cumani, P. D. Batzu, D. Colibro, C. Vair, P. Laface, V. Vasilakakis *et al.*, "Comparison of speaker recognition approaches for real applications." in *Proc. of Interspeech*, 2011, pp. 2365–2368.

[59] J. Thienpondt, B. Desplanques, and K. Demuynck, "The idlab voxsrc-20 submission: Large margin fine-tuning and quality-aware score calibration in dnn based speaker verification," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5814–5818.

[60] Z. Li, M.-W. Mak, and H. Meng, "Dual parameter-efficient fine-tuning for speaker representation via speaker prompt tuning and adapters," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 751–10 755.

[61] Z. Li, M.-W. Mak, H.-Y. Lee, and H. Meng, "Parameter-efficient fine-tuning of speaker-aware dynamic prompts for speaker verification," in *Proc. of Interspeech*, Sept 2024.

[62] L. Zhang, N. Jiang, Q. Wang, Y. Li, Q. Lu, and L. Xie, "Whisper-sv: Adapting whisper for low-data-resource speaker verification," *Speech Communication*, vol. 163, p. 103103, 2024.

[63] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned Wav2vec 2.0/HuBERT benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[64] R. Tao, K. A. Lee, R. K. Das, V. Hautamäki, and H. Li, "Self-supervised training of speaker encoder with multi-modal diverse positive pairs," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1706–1719, 2023.

[65] T. Zhou, Y. Zhao, J. Li, Y. Gong, and J. Wu, "CNN with phonetic attention for text-independent speaker verification," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 718–725.

[66] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 18 661–18 673, 2020.

[67] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[68] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner, "Towards a definition of disentangled representations," *arXiv preprint arXiv:1812.02230*, 2018.

[69] D. Raj, D. Snyder, D. Povey, and S. Khudanpur, "Probing the information encoded in x-vectors," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 726–733.

[70] R. Peri, H. Li, K. Somandepalli, A. Jati, and S. Narayanan, "An empirical analysis of information encoded in disentangled neural speaker representations," in *Proc. of Odyssey 2020: The Speaker and Language Recognition Workshop*, Tokyo, Japan, 2020.

[71] W. Lin, L. Li, and D. Wang, "Shuffle is what you need," in *Proc. of 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2022, pp. 245–249.

[72] Y. Liu, L. He, J. Liu, and M. T. Johnson, "Speaker embedding extraction with phonetic information," *Proc. of Interspeech*, 2018.

[73] S. Wang, J. Rohdin, L. Burget, O. Plchot, Y. Qian, K. Yu, and J. Cernockỳ, "On the usage of phonetic information for text-independent speaker embedding extraction." in *Proc. of Interspeech*, 2019, pp. 1148–1152.

VII. Biography Section

Zhe LI (Student Member, IEEE) is a PhD candidate in the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University and visiting PhD researcher in the Department of Electrical Engineering, Stanford University. He received his B.Eng. degree in computer science from Qilu University of Technology (Shandong Academy of Sciences), Jinan, China, in 2016 and his M.Sc. degree in software engineering from Xinjiang University, Urumqi, China, in 2021. He is an IEEE student member.

Man-wai MAK (Senior Member, IEEE) received a BEng (Hons) degree in Electronic Engineering from Newcastle Upon Tyne Polytechnic in 1989 and a Ph.D. degree in Electronic Engineering from the University of Northumbria at Newcastle (now Northumbria University) in 1993. He joined the Department of Electronic and Information Engineering (EIE) at The Hong Kong Polytechnic University in 1993, served as Interim Head of EIE from 2021 to 2023, and is presently a Professor and Associate Head of the Department of Electrical and Electronic Engineering. He has authored more than 220 technical articles and books in speaker recognition, machine learning, bioinformatics, and biomedical engineering and served as a guest editor of international journals. He has been an associate editor of IEEE Trans. on Audio, Speech and Language Processing, Journal of Signal Processing Systems, Advances in Artificial Neural Systems, and IEEE Biometrics Compendium. He is a tutorial speaker in Interspeech '16. Dr. Mak is also a co-author of the postgraduate textbook ”Biometric Authentication: A Machine Learning Approach, Prentice-Hall, 2005.”, ”Machine Learning for Protein Subcellular Localization Prediction, De Gruyter, 2015”, and ”Machine Learning for Speaker Recognition, Cambridge University Press, 2020.” He has received three Faculty of Engineering Research Grant Achievement Awards and a Faculty Award for Outstanding Performance (Research and Scholarly Activities). Dr. Mak has been an Executive Committee member of the IEEE Hong Kong Section Computer Chapter from 1995-2007 and the Chairman of the IEEE Hong Kong Section Computer Chapter from 2003-2005. He also served as a member of the IEEE Machine Learning for Signal Processing Technical Committee in 2005–2007 and as a Technical Committee Member of the IEEE Computation Intelligence Society, Intelligent Systems Applications, in 2008. Prof. Mak has served as Area Chair of Interspeech'14 and ICTAI 2016, Steering Committee Member of ISCSLP and ISCSLP16, and Program Co-Chair of ISCSLP 2018 and ISCSLP 2021. Prof. Mak's research interests include speaker recognition, machine learning, spoken language processing, biomedical engineering, and bioinformatics.

Jen-Tzung Chien (Senior Member, IEEE) received his Ph.D. degree in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 1997. He is now with the Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu. His research interests include machine learning, deep learning, speech recognition, blind source separation, face recognition, and information retrieval. Dr. Chien served as the Associate Editor of the IEEE Signal Processing Letters in 2008-2011 and the Tutorial Speaker for the ICASSP in 2012, the INTERSPEECH in 2013, the APSIPA in 2013, the ISCSLP in 2014, the ICASSP in 2015, the INTERSPEECH in 2016, and the ICASSP in 2017. He has published extensively, including the book ”Bayesian Speech and Language Processing”, Cambridge University Press, 2015. He is currently serving as an elected member of the IEEE Machine Learning for Signal Processing Technical Committee. He is the general co-chair of the IEEE International Workshop on Machine Learning for Signal Processing in 2017.

Mert Pilanci He is an assistant professor in the Department of Electrical Engineering at Stanford University. Prior to joining Stanford, he was an assistant professor of Electrical Engineering and Computer Science at the University of Michigan. In 2017, he was a Math+X postdoctoral fellow working with Emmanuel Candès at Stanford University. He received my Ph.D. in Electrical Engineering and Computer Science from UC Berkeley in 2016. His studies were supported partially by a Microsoft Research PhD Fellowship. He obtained his B.S. and M.S. degrees in Electrical Engineering from Bilkent University. His research interests are in large-scale machine learning, optimization, and information theory.

Zezhong Jin Zezhong Jin received his B.Eng. degree in Electronic information engineering from Hebei University in 2021 and M.Sc. degree in Electronic and Information Engineering from The Hong Kong Polytechnic University in 2023. He is currently pursuing a Ph.D. degree in the Department of Electrical and Electronic Engineering at The Hong Kong Polytechnic University. His research interests include speaker verification, self-supervised learning, and knowledge distillation.

Helen Meng (Fellow Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, MA, USA. Helen Meng is Patrick Huen Wing Ming, a professor of systems engineering and engineering management at the Chinese University of Hong Kong (CUHK). She is the Founding Director of the CUHK Ministry of Education (MoE)-Microsoft Key Laboratory for Human-Centric Computing and Interface Technologies (since 2005), Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems (since 2006), and Stanley Ho Big Data Decision Analytics Research Center (since 2013). Previously, she served as CUHK Faculty of Engineering's Associate Dean (Research), Chairman of the Department of Systems Engineering and Engineering Management, Editor-in-Chief of the IEEE Transactions on Audio, Speech and Language Processing, Member of the IEEE Signal Processing Society Board of Governors, ISCA Board Member and presently member of the ISCA International Advisory Council. She was elected APSIPA's inaugural Distinguished Lecturer 2012-2013 and ISCA Distinguished Lecturer 2015-2016. Her awards include the Ministry of Education Higher Education Outstanding Scientific Research Output Award 2009, Hong Kong Computer Society's inaugural Outstanding ICT Woman Professional Award 2015, Microsoft Research Outstanding Collaborator Award 2016 (1 in 32 worldwide), IEEE ICME 2016 Best Paper Award, IBM Faculty Award 2016, HKPWE Outstanding Women Professionals and Entrepreneurs Award 2017 (1 in 20 since 1999), Hong Kong ICT Silver Award 2018 in Smart Inclusion, CogInfoComm2018 Best Paper Award and the 2019 IEEE SPS Leo L. Beranek Meritorious Service Award for exemplary service to and leadership in the Signal Processing Society. Her research interests include AI for speech and language technologies to support multilingual and multimodal human-computer interactions, eLearning and assistive technologies, and big data decision analytics using AI. Helen has served in numerous Government appointments, including memberships in the Research Grants Council and the Steering Committee of Hong Kong's Electronic Health Record Sharing. Helen is a Fellow of HKCS, HKIE, IEEE, and ISCA.