Unsupervised TTS Acoustic Modeling for TTS With Conditional Disentangled Sequential VAE

Jiachen Lian, *Member, IEEE*, Chunlei Zhang, *Member, IEEE*, Gopala K. Anumanchipalli, *Member, IEEE*, and Dong Yu, *Fellow, IEEE*

***Abstract*—In this paper, we propose a novel unsupervised text-to-speech acoustic model training scheme, named UTTS, which does not require text-audio pairs. UTTS is a multi-speaker speech synthesizer that supports zero-shot voice cloning, it is developed from a perspective of disentangled speech representation learning. The framework offers a flexible choice of a speaker's duration model, timbre feature (identity) and content for TTS inference. We leverage recent advancements in self-supervised speech representation learning as well as speech synthesis front-end techniques for system development. Specifically, we employ our recently formulated Conditional Disentangled Sequential Variational Auto-encoder (C-DSVAE) as the backbone UTTS AM, which offers well-structured content representations given unsupervised alignment (UA) as condition during training. For UTTS inference, we utilize a lexicon to map input text to the phoneme sequence, which is expanded to the frame-level forced alignment (FA) with a speaker-dependent duration model. Then, we develop an alignment mapping module that converts FA to UA. Finally, the C-DSVAE, serving as the self-supervised TTS AM, takes the predicted UA and a target speaker embedding to generate the mel spectrogram, which is ultimately converted to waveform with a neural vocoder. We show how our method enables speech synthesis without using a paired TTS corpus in AM development stage. Experiments demonstrate that UTTS can synthesize speech of high naturalness and intelligibility measured by human and objective evaluations. Audio samples are available at our demo page.**

***Index Terms*—Unsupervised TTS acoustic modeling, self-supervised learning, wavLM, C-DSVAE, representation learning.**

I. Introduction

TEXT-TO-SPEECH (TTS) synthesis plays an important role for human computer interaction. With the continuous development of neural-based TTS systems, e.g., Tacotron [1], DurIAN [2], FastSpeech [3], [4], [5] or more recent Glow-TTS series [6], [7], [8], [9], high-fidelity synthetic speech has been produced and the gap between machine generated speech and real speech is reduced. This is especially true for languages with rich resources, which refers to a sizeable high quality parallel speech and textual data. Usually, the supervised TTS system requires dozens of hours of single-speaker high quality data to retain a good performance. However, collecting and labeling such data is a non-trivial task, which is time-consuming and expensive. In that sense, current supervised solutions are still have their limitations on the demanding needs of ubiquitous deployment of customized speech synthesizers for AI assistants, gaming or entertainment industries. Natural, flexible and controllable TTS pathways become more essential when facing these diverse needs.

One possible way to relax the aforementioned limitation is to develop text-to-speech synthesis that does not require high quality parallel speech and textual data for training TTS acoustic models (AM). In this preliminary study, we propose a novel UTTS framework based on the recent advancement of self-supervised learning (SSL) techniques for speech representation [10], [11], [12], [13]. Like many conventional supervised TTS systems, our proposed UTTS framework is essentially a two-stage TTS process, namely a model that converts a raw text to a mel spectrogram and a vocoder which transforms the mel spectrogram to the speech waveform. While the supervised TTS AM achieves the text to mel spectrogram transformation by explicitly composing an external aligner or by learning a latent alignment [2], [3], [4], [8], the UTTS AM must resolve this mapping with an indirect way since no paired TTS AM training data is available.

For conceptual comparison purpose, Fig. 1 illustrates the necessary components of a typical supervised TTS AM (e.g., DurIAN [2] or FastSpeech2 [4] which shares a similar structure) and our UTTS AM. Although UTTS follows most of the procedures of supervised AM during inference, the UTTS AM differs the supervised AM in three major ways: 1) For supervised AM, the neural network maps forced alignment of phoneme sequence to frame-by-frame acoustic features. We replace forced alignment with unsupervised alignment, which is derived from acoustic partitioning with the SSL features (e.g., kmeans labels of WavLM [11] features). 2) supervised AMs (DurIAN and FastSpeech2) are only trained with TTS speech data, since UTTS investigates the Conditional Disentangled Sequential Variational Auto-encoder (C-DSVAE [12]) as the backbone, the self-supervised/unsupervised model possibly enables more generic speech data to build a more generalized AM instead of being restricted with the TTS data only. 3) UTTS

Manuscript received 11 October 2022; revised 4 April 2023 and 24 May 2023; accepted 17 June 2023. Date of publication 28 June 2023; date of current version 7 July 2023. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yu Tsao. *(Jiachen Lian and Chunlei Zhang contributed equally to this work.) (Corresponding author: Chunlei Zhang.)*

Jiachen Lian was with the Tencent AI Lab Bellevue, Bellevue, WA 98004 USA. He is now with the EECS, UC Berkeley, Berkeley, CA 94720 USA (e-mail: jiachenlian@berkeley.edu).

Chunlei Zhang and Dong Yu are with the Tencent AI Lab Bellevue, Bellevue, WA 98004 USA (e-mail: cleizhang@global.tencent.com; dyu@tencent.com).

Gopala K. Anumanchipalli is with the EECS, UC Berkeley, Berkeley, CA 94720 USA (e-mail: gopala@berkeley.edu).

Digital Object Identifier 10.1109/TASLP.2023.3290423

[Online]. Available: https://neurtts.github.io/utts_demo/

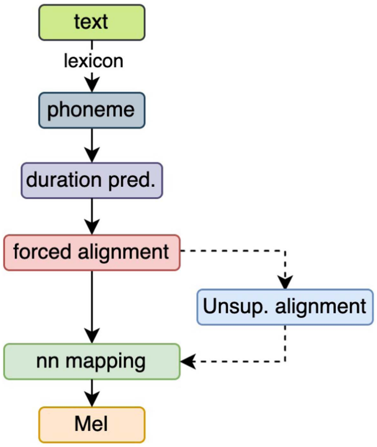


Fig. 1. Text to mel spectrogram converion with supervised TTS AM to UTTS AM.

uses external speech recognition data to train the multi-speaker duration model and the projection between forced alignment to unsupervised alignment (FA2UA), while traditional TTS AMs rarely introduces such variability.

For system implementation and experimental studies, the multi-speaker VCTK and LibriTTS dataset [14], [15] are utilized to demonstrate the effectiveness of the proposed method. We show the possibility that UTTS can generate speech of high naturalness and intelligibility in terms of objective and subjective tests.

Although more detailed explanations and analysis can be found throughout this article, the main contributions of this work are summarized as follows:

- By exploring the conditions to end to end disentangled speech representation learning, we justify that unsupervised TTS AM is feasible.
- TTS and voice conversion are unified under a generative model (C-DSVAE), showing the potential of employing self-supervised (disentangled) representation learning for voice generation tasks.
- With disentangled representations, UTTS offers a flexible framework that can bridge control signal (speaker identity, text, style etc.) and generated voice, which enables more fine grained processes for controllable and customized TTS.

The rest of this article is organised as follows. Section II introduces the background and related work. Section III illustrates the C-DSVAE and its development for voice generation. The UTTS framework is presented in Section IV. Experimental setup and results are shown in Section V, with conclusions in Section VI.

II. Related Work

In this section, we introduce the background and relation to our proposed method.

Supervised TTS: The neural network based supervised TTS systems can be generally categorised into two forms, the end-to-end methods [4], [8], [16], [17], [18] that directly maps text to speech waveform and the two-stage solutions [1], [2], [3], [4], [5], [7], [17] that utilize an acoustic model to convert text to frame by frame acoustic features (e.g., mel spectrogram) and a vocoder transforms the intermediate features to the waveform. While the end-to-end TTS draws an increased attention for some tasks, most of the current studies still follow the two-stage methods, mainly because of the less constraint from the data and optimization. As mentioned in the Introduction, the proposed UTTS system assembles like the two-stage supervised TTS, with the differences in some of the critical building blocks.

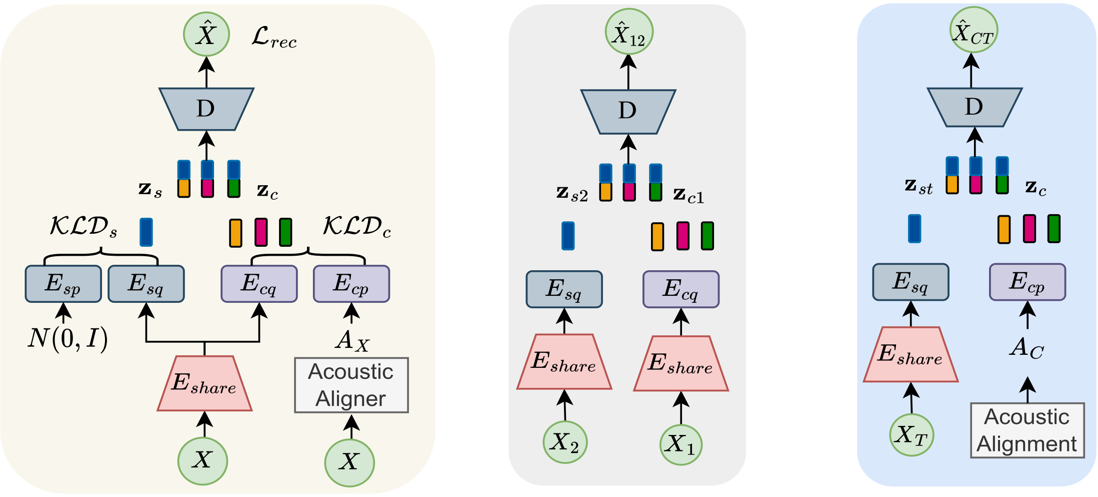
Unsupervised TTS: There are not so many successful tries of the unsupervised TTS system development until recently. Thanks to the recent development of self-supervised speech representation learning, both systems in [19], [20] used an unsupervised ASR model to transcribe the TTS speech data. With the pseudo labels, they once again can adopt the supervised TTS recipe to build the unsupervised TTS systems. The result indicates that it is possible to produce speech with high intelligibility even without paired TTS data. One notable concern for such unsupervised TTS systems remains in the development of unsupervised ASR, which still requires necessary improvement to close the gap between the supervised counterparts [21], [22]. Strictly speaking, current unsupervised ASR systems have to rely on either lexicons or alignment prediction to match the acoustics and text, such lexicon or aligner are built with supervisions (e.g., paired data is used in developing CMU dictionary). Regarding our unsupervised approach, instead of transcribing the training speech data and performing supervised network mapping, our TTS AM is a variant of variational auto-encoder (i.e., C-DSVAE [12]), with unsupervised alignment as the condition to regularize the learned latent representation to follow the phonetic structure. With this design, our UTTS system requires an additional mapping for text (forced alignment) space to the unsupervised alignment space. For that purpose, we just conduct a data driven process to map forced alignment to unsupervised alignment with the employment of an ASR system. Please refer to Section IV-B for more details about transforming text to unsupervised acoustic alignment for UTTS inference.

Disentangled speech representation for voice generation: Speech disentanglement algorithms, such as DSVAE and C-DSVAE [12], [13], [23], [24], provide a way to isolate global timbre/speaker feature and local content features. Our work is inspired from disentangled speech representation learning and its downstream tasks, such as reconstruction/resynthesis and voice conversion [12], [13], [25], [26]. The difference between those studies and UTTS is that, resyntheis or VC are speech driven, the process keeps the content unchanged, while UTTS has to make a linkage between raw text and the intermediate content embeddings in order to generate meaningful output speech. And how to connect the front-end text with unsupervised TTS AM (C-DSVAE) remains the key in building our UTTS system.

III. C-DSVAE for Voice Generation

A. Overview

Disentangling speaker information from speech content is of great importance to speech tasks such as speaker recognition,



(a) C-DSVAE training (b) voice conversion (c) alignment driven voice generation

Fig. 2. System diagram of conditional DSVAE.

automatic speech recognition, voice conversion and text-to-speech [26], [27], [28], [29], [30]. Of the existed speech disentanglement methods, DSVAE [24] is a compelling framework that generates the disentangled speaker and content representation in an unsupervised manner. A recent DSVAE study in speech [13] presents impressive speaker verification and voice conversion performance by balancing the information flow between speaker and content. As the content prior being assumed independent of the input data, early versions of DSVAE suffer from discontinuous and unstable speech generation with damaged phonetic structures [24], [31], [32], [33]. More recently, C-DSVAE [12] was proposed to introduce the content-dependent acoustic alignment as a condition to regularize the DSVAE training, from which the phonetically discriminative content representations are learned and the stabilized vocalization is achieved, overcoming the architecture drawback of earlier explorations in DSVAE [31], [32], [33] and delivering enhanced zero-shot VC performances. In the following sections, we first present a systematic overview of the DSVAE framework, which serves as the backbone of the improved C-DSVAE. Then, we present detailed development methods of C-DSVAE by deliberating different types of content conditions as well as training objectives. Furthermore, we extend C-DSVAE by introducing the *masked prediction training* to capture the robust and long-range temporal relations over acoustic units. In the last two sections, we discuss how the voice conversion and alignment-driven voice generation are performed within the C-DSVAE framework, the latter exploration serves as the cornerstone of our UTTS system to be discussed in Section IV. The detailed model architectures can be found in Section V-B1.

B. DSVAE Framework

1) Backbone: Fig. 2(a) illustrates the training process of C-DSVAE, where DSVAE serves as the backbone architecture. The DSVAE consists of a shared encoder E_{share}, a posterior speaker encoder E_{sq}, a posterior content encoder E_{cq}, a prior speaker encoder E_{sp}, a prior content encoder E_{cp} and a decoder D. The input utterance X is passed into E_{share}, followed by E_{sq} and E_{cq} which encode the speaker posterior distribution $q(z_s|X)$ and the content posterior distribution $q(z_c|X)$, respectively. z_s denotes the speaker embedding and z_c denotes the content embedding. For the prior modeling, E_{sp} encodes the speaker prior $p(z_s)$ and E_{cp} encodes the content prior $p(z_c)$. Depending on different voice generation tasks, during the decoding/generation stage, the speaker embedding z_s and content embedding z_c are sampled from either the posteriors $q(z_s|X)$ and $q(z_c|X)$ or the priors $p(z_s)$ and $p(z_c)$, and the concatenation of them is passed into decoder D to generate the synthesized speech $\hat{X} = D(z_s, z_c)$.

2) Probabilistic Graphical Models and Training Objective: As introduced in [12], [13], [24], DSVAE is designed to follow independence assumption for both priors and posteriors to implicitly leverage disentanglement between speaker and content representations. The probabilistic graphical models are formulated in (1) and (2). Note that the factorization of $p_\theta(z_c)$ and $q_\theta(z_c|X)$ can either follow the streaming manner [13], [24], [33] or non-streaming manner [12], depending on the practical implementation. Equations (1) and (2) correspond to the latter one, where t is the frame index and θ denotes model parameters.

$$p_\theta(z_s, z_c) = p(z_s)p_\theta(z_c) = p(z_s)\prod_{t=1}^{T} p_\theta(z_{ct}|z_c) \tag{1}$$

$$\begin{aligned} q_\theta(z_s, z_c|X) &= q_\theta(z_s|X)q_\theta(z_c|X) \\ &= q_\theta(z_s|X)\prod_{t=1}^{T} q_\theta(z_{ct}|X) \end{aligned} \tag{2}$$

The overall training objective for DSVAE is a VAE loss, but with two latent variables i.e. z_s and z_c, as shown in (5). L_{DSVAE} forces strong disentanglement between z_s and z_c without additional constraints such as mutual information regularization [31], [32], [33], [34], as explained in [13]. KL divergence is denoted as $KLD, p_\theta(X|z_s, z_c))$ is modeled by the decoder D, representing the generation process. α and β are two balancing factors which are used to control the disentanglement [13].

$$\mathcal{L}_{KLD_s} = \mathbb{E}_{p(X)}\left[KLD(q_\theta(z_s|X)||p(z_s))\right] \quad (3)$$

$$\mathcal{L}_{KLD_c} = \mathbb{E}_{p(X)}\left[KLD(q_\theta(z_c|X)||p_\theta(z_c))\right] \quad (4)$$

$$\begin{aligned}\mathcal{L}_{DSVAE} = {} & \mathbb{E}_{p(X)}\mathbb{E}_{q_\theta(z_s,z_c|X)}\left[-\log(p_\theta(X|z_s, z_c))\right] \\ & + \alpha\mathcal{L}_{KLD_s} + \beta\mathcal{L}_{KLD_c} \quad (5)\end{aligned}$$

C. C-DSVAE System

In the DSVAE systems [13], [24], [31], [33], the content prior $p(z_c)$ is independent of X. As formulated in (4), the training objective is to minimize the KL divergence between $p(z_c)$ and $q(z_c|X)$. Since $q(z_c|X)$ explicitly models the phonetic structure of the speech content, the content-independent prior will force the content representation to lose phonetic discriminativity in the latent space, which is detrimental to the disentanglement. In order to generate phonetically meaningful and continuous speech with stable vocalizations, we introduce the acoustic alignment A_X as the condition for content prior distribution. The motivations are two-folds. First, acoustic alignment carries the phonetic structure of the utterance, which will be encoded into the content prior model. Such alignment-aware prior acts as a "teacher" model to force the "student", which is the content posterior $q_\theta(z_c|X)$, to inherit the phonetic structure of the original utterance. Second, during the inference, the acoustic alignment A_X could be sampled from the alignment-aware prior so as to perform alignment-driven speech synthesis, which is the basis for our UTTS system to be introduced in Section IV.

1) Acoustic Alignment as Content Condition: We consider two types of acoustic alignment: forced alignment (**FA**) and unsupervised alignment (**UA**). For a specific utterance X, we denote the acoustic alignment as A_X^{FA} and A_X^{UA} respectively.

a) Forced alignment: A_X^{FA} is the forced alignment of the utterance X. We use *Montreal forced alignment* (MFA) [35] to extract the forced alignment given audio-text pair. In consistency with [5], the forced alignment is a sequence of monophones and there is a total number of 72 monophones. In the next phase, the content prior encoder E_{cp} takes the one-hot form of alignment sequence as input at each time step and predicts the frame-wise content prior distribution $p(z_c|A_X^{FA})$.

b) Unsupervised alignment: The forced alignment A_X^{FA} is not always available since audio-transcription pair does not naturally exist. And because of this, the existing supervised TTS models are not scaled-well with respect to audio samples under diverse acoustic conditions, which limits their potential in synthesizing natural and expressive speech in the generalized open domains. To further utilize the advantage of untranscribed audio samples that are largely existed, we switch to the unsupervised alignment that is derived via the self-supervised pre-training. Of the popularly proposed self-supervised speech pre-training models, WavLM [11] delivers state-of-art performance in "full stack" downstream speech tasks such as speaker recognition, speaker diarization, ASR and speech separation. WavLM employs a denoising masked speech prediction training strategy by feeding noisy and overlapped speech as inputs to predict the pseudo-labels of original speech on masked region, which improves both the speech content modeling capability and model robustness. The WavLM-Base model is pretrained from Librispeech-960h [36], and we adopt the WavLM-Base [11] model to extract the acoustic features. With WavLM features adopted, Kmeans++ [37] algorithm is performed on the features to generate frame-wise pseudo labels, denoting as unsupervised alignment A_X^{UA}, which is also assumed to capture the phonetic structure of the utterance X. The pretrained WavLM-Base model together with Kmeans algorithm acts as the *acoustic aligner* as indicated in Fig. 2(a).

2) Masked unit prediction training: In order to capture the robust and long-range temporal relations over acoustic units and to generate more continuous speech, we adopt the *Masked Unit Prediction* (MUP) when training the content prior encoder E_{cp}. Denote $M(A_X) \subset [T]$ as the collection of T masked indices for a specific condition A_X, where the masking configuration is consistent with [10]. Let $\tilde{A_X}$ be a corrupted version of A_X, in which A_{X_t} will be masked out if $t \in M(A_X)$. Denote z_{cp} as the sample of output of the content prior encoder E_{cp}, i.e. $z_{cp} \sim E_{cp}(A_X)$. The negative loglikelihood loss (NLL) $\mathcal{L}_{MUP-C}$ for condition modeling is defined in (6), where $p(z_{cp_i}|\tilde{A_{X_i}})$ is the softmax categorical distribution. $\mathbb{E}_{A_X}$ denotes the expectation over all A_X. Since it is reported in [10] that masked-only prediction delivers impressive ASR performances, we do not incorporate the prediction loss for unmasked regions.

$$\mathcal{L}_{MUP-C} = -\mathbb{E}_{A_X}\Sigma_{i\in M(A_X)}\log p(z_{cp_i}|\tilde{A_{X_i}}) \quad (6)$$

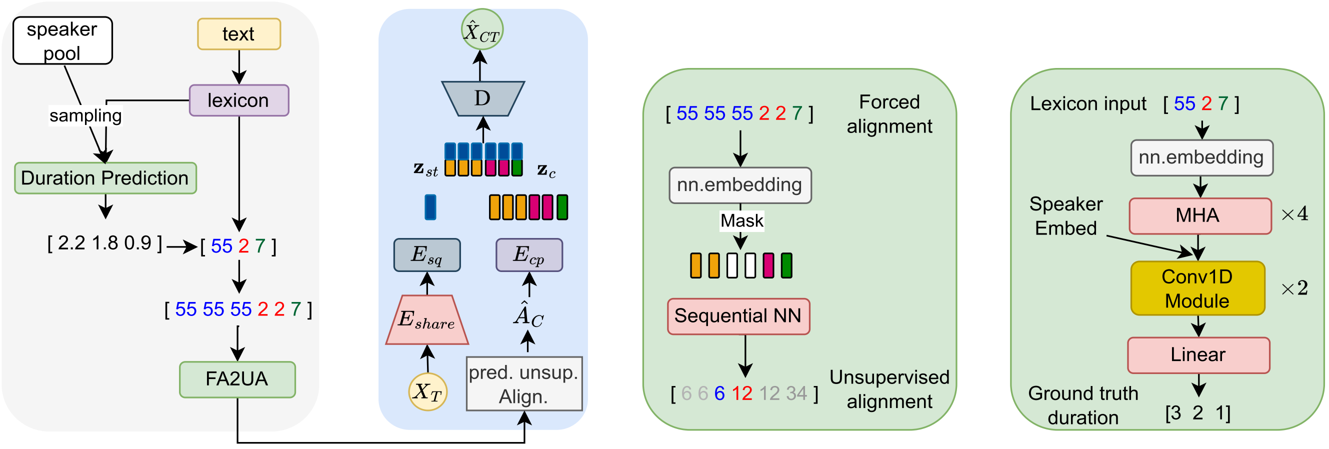
3) Training objective and configuration: We adopt the DSVAE training objective as expressed in (5), with the exception that the L_{KLD_c} is formulated with conditional content prior, as shown in (7). We design the content prior encoder E_{cp} such that $p_\theta(z_c|A_X)$ follows a non-streaming factorization. The masked unit prediction loss L_{MUP-C} is also incorporated. The overall loss objective for C-DSVAE is expressed in (8). When $\gamma = 0$, it is vanilla C-DSVAE training. We set $\alpha = 0.01$, $\beta = 10$ and $\gamma = 1$ if masked unit prediction training is involved. We adopt Adam optimizer [38] with an initial learning rate of 5e-4 which is decayed by 0.95 every 5 epochs. The batch size is fixed to be 256. The training configurations are in line with [12], [13].

$$\mathcal{L}_{KLD_c-C} = \mathbb{E}_{p(X)}\left[KLD(q_\theta(z_c|X)||p_\theta(z_c|A_X))\right] \quad (7)$$

$$\begin{aligned}\mathcal{L}_{C-DSVAE} = {} & \mathbb{E}_{p(X)}\mathbb{E}_{q_\theta(z_s,z_c|X)}\left[-log(p_\theta(X|z_s, z_c))\right] \\ & + \alpha\mathcal{L}_{KLD_s} + \beta\mathcal{L}_{KLD_c-C} + \gamma L_{MUP-C} \quad (8)\end{aligned}$$

D. Applications of C-DSVAE

1) Voice Conversion: VC seeks to convert the speaker information of source utterance to the target speaker while keeping the speech content unchanged. As presented in Fig. 2(b), given the arbitrary source utterance X_1 and arbitrary target utterance X_2, the speaker embedding z_{s2} is extracted from X_2 using the posterior speaker encoder E_{sq} and the content embedding z_{c1} is extracted from X_1 using the posterior content encoder E_{cq}. The decoder then takes the concatenation of z_{s2} and z_{c1}



(a) A flow diagram of the UTTS inference procedure (b) FA2UA masked prediction training (c) Speaker-ware Duration Prediction

Fig. 3. Core components of the proposed UTTS system.

and generates the converted mel spectrogram $\hat{X_{12}}$, which is converted to waveform via a pre-trained neural vocoder.

2) Alignment driven Voice Generation: Given acoustic alignment A_C (either forced alignment or unsupervised alignment) and arbitrary target utterance X_T, alignment-driven voice generation aims to synthesize the speech from the alignment with the target speaker embedding, as shown in Fig. 2(c). Specifically, the speaker embedding z_s is extracted from X_T using posterior speaker embedding E_{sq} and the content embedding is sampled from $p_\theta(z_c|A_C)$ generated via the content prior encoder E_{cp}, which takes A_C as input. The alignment-driven voice generation is an essential trick that bridge the textual condition to the acoustic representations in our study, which serves as a cornerstone of our UTTS framework.

IV. UTTS System

Our proposed UTTS pipeline is presented in Fig. 3. We first obtain the phoneme sequence of the text transcription with Librispeech Lexicon,[1] which is consistent with [5]. The phoneme sequence is then converted to a list of token indices. The phoneme sequence (token indices) together with a speaker embedding (from a speaker pool for duration model training) is passed into the **Speaker-Aware Duration Prediction** module which delivers the predicted forced alignment (FA). The **FA2UA** module takes FA as input and predicts unsuperviserd alignment (UA). The UA along with a target speaker embedding is fed into the C-DSVAE to generate mel spectrogram. A neural vocoder is then applied to convert mel spectrogram to waveform. It is noted that the proposed UTTS system performs zero-shot **voice cloning** for the target utterance, which is a very difficult setting for TTS. In the following sections, we present the **Speaker-Aware Duration Prediction** and **FA2UA** modules. Both the modules are trained separately, the detailed model architecture of each module are presented in Section V-B1.

A. Speaker-Aware Duration Prediction

The duration predictor takes the phoneme sequence as well as speaker embedding as input to predict the speaker-aware duration for each phoneme [2], [4], as shown in Fig. 3(c). Specifically, the phoneme sequence is first passed into a trainable look-up table to obtain the phoneme embeddings. Afterwards, a four-layer multi-head attention (MHA) module is followed to extract the latent phoneme representation. A two-layer conv-1D module is then used to take the summation of latent phoneme representation and speaker embedding sampled from the speaker pool. A linear layer is finally applied to generate the predicted duration in the logarithmic domain. During training, the Mean Squared Error (MSE) is utilized to calculate the difference between the predicted duration and the target duration. For training, the target duration of a phoneme sequence is obtained from the forced alignment extracted by Montreal Forced Alignment (MFA) [35]. Note that the target duration is also in the logarithmic scale. During inference, we round up the duration generated by the predictor and expand the phoneme sequence to form an estimated forced alignment. We also present an example of how the duration predictor works during inference. Assume a random speaker is sample and the phoneme sequence is [55 2 7]. The predicted duration sequence is [2.2 1.8 0.9], correspondingly. We first round up the duration to make it become [3 2 1]. The phoneme sequence is then expanded to [55 55 55 2 2 7] as the predicted forced alignment (FA). The duration predictor works similarly to [2], [4], [5].

B. FA2UA With Masked Prediction Training

The FA2UA module takes forced alignment (FA) as input and predicts corresponding unsupervised alignment (UA). Specifically, FA is first passed into a learnable look-up table to obtain the FA embeddings. Subsequently, a 3-layer Bi-LSTM module is employed to predict the UA embeddings given FA embeddings. During training, we adopt a masked prediction training strategy to train our FA2UA module as masked prediction is expected to be good at capture the long-range time dependency across

[1] [Online]. Available: https://www.openslr.org/11/

tokens and encode more contextualized information for each token. Similar as Section III-C2, we denote $M(FA) \subset [T]$ as the collection of T masked indices for a specific unsupervised alignment FA. Let $\tilde{FA}$ be a corrupted version of FA, in which FA_i will be masked out if $i \in M(FA)$. Denote UA_i as the corresponding frame level unsupervised alignment of FA_i, which serves as the ground truth for the masked prediction training. The negative loglikelihood loss (NLL) $\mathcal{L}_{FA2UA}$ for masked prediction training is defined in (9), where $p(UA_i|\tilde{FA_i})$ is the softmax categorical distribution. $\mathbb{E}_{(FA,UA)}$ denotes the expectation over all (FA, UA) pairs. Following Section III-C2, masked-only prediction loss is incorporated for unsupervised alignment prediction. Fig. 3(b) presents an example, where [55 2] are masked out from the input FA and the model only predicts the corresponding UA targets [6 12] in the training phase. During the inference, we pick the token with the maximum probability $p(UA_i|FA_i)$ at each time step i to form the predicted UA sequence.

$$\mathcal{L}_{FA2UA} = -\mathbb{E}_{(FA,UA)}\Sigma_{i\in M(FA)} \log p\left(UA_i|\tilde{FA_i}\right) \quad (9)$$

C. Dual Content Encoders for Reconstruction

It is noted that the content representation to the decoder of C-DSVAE is sampled from the posterior content encoder E_{cq} in the original C-DSVAE training, which is suitable for the voice conversion tasks but not necessarily ideal for alignment-driven voice generation (i.e., UTTS inference). There exists a mismatch between the reconstruction based training and the UTTS inference. To alleviate this mismatch, we propose to use **dual content encoders for reconstruction** in the C-DSVAE training. Specifically, we run the normal C-DSVAE training guided by (8) until convergence, and then we start a second round of training where we also sample latent representation z_{cp} from the prior content encoder for reconstruction. And the second round of training is initialed with the pretrained C-DSVAE model. With this kind of training strategy, the dual content encoders for reconstruction training can be described as (10):

$$\begin{aligned}\mathcal{L}_{C-DSVAE} = \frac{1}{2}\{&\mathbb{E}_{p(X)}\mathbb{E}_{q_\theta(z_s,z_{cq}|X)}[-log(p_\theta(X|z_s,z_{cq}))]\\ &+\mathbb{E}_{p(X)}\mathbb{E}_{q_\theta(z_s,z_{cp}|X,A_X)}\\ &\times[-log(p_\theta(X|z_s,z_{cp}))]\}\\ &+\alpha\mathcal{L}_{KLD_s}+\beta\mathcal{L}_{KLD_c-C}+\gamma L_{MUP-C} \quad (10)\end{aligned}$$

where the new reconstruction term is formulated as an average loss with both posterior content embedding and prior content embedding for reconstruction. In our experiment, we find a clear quality improvement of the UTTS samples over the posterior content only reconstruction models.

V. Experiments

A. Datasets

a) VCTK: For VCTK corpus [14], we exclude speaker p315 since the transcriptions are not available. we then randomly select 90% of the speakers as training set and the remaining 10% as test set. The data is downsampled to 16 KHz. We extract mel spectrogram with the window size/hop size of 64 ms/ 16 ms, and the feature dimension is 80. 36900 trials are randomly generated for speaker verification subtasks. The settings are consistent with [12], [13].

b) LibriTTS: We downsample the original LibriTTS [15] data to 16 kHz. The original training split is taken as training set for C-DSVAE, FA2UA and duration predictor training. The mel spectrogram that is extracted for training follows the same configurations as the VCTK dataset.

B. Implementation Details

1) Model Architectures: Table I presents details of UTTS system. In the C-DSVAE architecture, for posterior speaker encoder E_{sq}, posterior content encoder E_{cq} and prior content E_{cp}, there is one dense layer used to predict the mean and one other dense layer used to predict the standard deviation. A linear classifier will be added at the end of E_{cp} if masked prediction training is involved. Both speaker and content dimension are set to 64. The decoder D consists of a prenet D_{pre} and a postnet D_{post}, which were introduced in [12], [13], [26]. We use HiFi-GAN [15], [39] pretrained with LibriTTS as the universial vocoder[2] for voice conversion and text-to-speech synthesis. We use the same mask configuration as [10] for all mask prediction experiments.

2) Training Experiments of Each Components: For C-DSVAE system, we first train it without content condition (i.e. DSVAE [13]). Next, we explore the C-DSVAE training with different content conditions, training strategies and datasets, as noted in the bracket. For instance, C-DSVAE(UA)/C-DSVAE(FA) indicates that the training is conditioned on UA/FA, and C-DSVAE(UA/FA+MP) represents the mask prediction training along with either C-DSVAE(UA) or C-DSVAE(FA). By default C-DSVAE is trained on VCTK corpus unless otherwise specified. For example, C-DSVAE(UA/FA+MP+LibriTTS) shows that C-DSVAE is pretrained on LibriTTS inherited with the C-DSVAE(UA/FA+MP) setting and is then fine-tuned on VCTK under the same configuration.

In our experiments, the speaker-aware duration predictor and the FA2UA module are trained with LibriTTS for two major reasons: 1) LibriTTS serves as a big external dataset introducing more speaker and phonetic varibilities to the UTTS AMs; 2) we justify that using out of domain speech data (instead of paired VCTK speech and textual data), UTTS can build robust text-to-speech transformations. And the following experimental results demonstrate more details about the systems we developed.

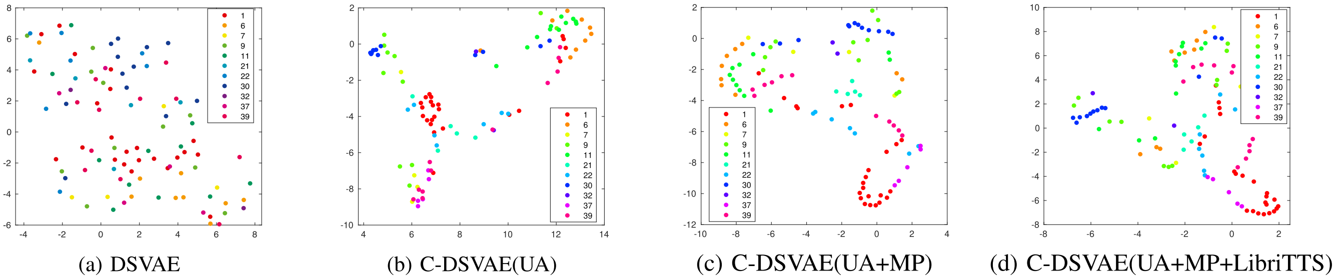
3) UTTS Inference: We use the real-time factor (RTF) to denote the model efficiency in TTS inference. More training hyperparameters can be found in Section III-C3. The training and inference latency tests are conducted on a server with 48 Intel Xeon CPUs and 4 NVIDIA V100 GPU. We set the batchsize to 1 for inference. The RTF for UTTS is 2.7e-2. Our model can be viewed as an unsupervised AM version of Fastspeech 2, with an additional module that maps text to unsupervised alignment as the input to UTTS AM.

[2] [Online]. Available: https://github.com/ming024/FastSpeech2/tree/master/hifigan

TABLE I
UTTS NETWORK ARCHITECTURE IN DETAIL. FOR CONV1D, THE CONFIGURATION IS (FILTER SIZE, KERNEL SIZE, PADDING, STRIDE)

UTTS Model Architecture		
C-DSVAE	E_{Share}	(Conv1D(256, 5, 2, 1)→ InstanceNorm2D→ ReLU)×3
	E_{sq}	BiLSTM(512, 2)→ Average Pooling → (Dense(64)⇒ mean, Dense(64)⇒ std)
	E_{sp}	Identity Mapping
	E_{cq}	BiLSTM(512, 2)→RNN(512, 1)→ (Dense(64)⇒ mean, Dense(64)⇒ std)
	E_{cp}	BiLSTM(512, 2)→ (Dense(64)⇒ mean, Dense(64)⇒ std)(→ Linear Classifier)
	D_{pre}	(InstanceNorm2D→ Conv1D(512, 5, 2, 1)→ ReLU)×3
		LSTM(512, 1) → LSTM(1024, 2) → Dense(80)
	D_{post}	(Conv1D(512, 5, 2, 1)→ tanh→ InstanceNorm2D)×4
Duration Predictor		nn.Embedding→MHA(256, 128, 128, 2)→ Conv1D(256, 3, 2, 1) → Dense(1)
FA2UA		nn.Embedding→BiLSTM(256, 3)→ Linear Classifier

For Multi-Head-Attention (MHA), the configuration is (model dimension, key dimension, value dimension, #heads). For LSTM/BILTSM/RNN, the configuration is (hidden DIM, layers). For dense layer, the configuration is (output DIM).



(a) DSVAE (b) C-DSVAE(UA) (c) C-DSVAE(UA+MP) (d) C-DSVAE(UA+MP+LibriTTS)

Fig. 4. Visualizations of learned content embeddings.

C. Evaluations on C-DSVAEs

1) C-DSVAE Content Embeddings and Speaker Embeddings: We evaluate the disentanglement performance of C-DSVAE system via both the t-SNE [40] analysis and objective analysis on content and speaker embeddings.

a) t-SNE Analysis: Given pre-trained models, the same utterance that is randomly selected from VCTK test set is taken as input and the t-SNE visualization of the content embeddings is made, where the labels are the forced alignment. We assume that better phonetic discriminativity indicates better disentanglement. As shown in Fig. 4, C-DSVAE(UA) outperforms DSVAE significantly in terms of phoneme discriminativity, which was also reported in [12]. It is reasonable that C-DSVAE(UA+MP) and C-DSVAE(UA+MP+LibriTTS) delivers better phonetic structure than C-DSVAE(UA) because masked prediction training does help encode the contextualized and structured phonetic information into the content embeddings. It is also observable that C-DSVAE(UA+MP) delivers better inter-class but worse intra-class phoneme discriminativity, which is assumed to be caused by involving out-of-domain LibriTTS corpus during the training.

b) Objective Analysis: We perform speaker verification (SV) on both speaker and content embeddings given pre-trained C-DSVAEs. The experiments are conducted on the VCTK SV trials mentioned in Sec.V-A, equal error rate (EER) is the metric. Typically, lower EER on speaker embeddings and higher EER on content embeddings indicate better disentanglement. Additionally, the pre-trained C-DSVAE model is fine-tuned with forced alignment on VCTK training set and the phoneme classification accuracy (Top1) for test set is reported. As observed in Table II, C-DSVAEs outperform DSVAE consistently in all metrics, which is consistent with [12]. The C-DSVAE with FA condition outperforms that with UA condition, which is reasonable because FA is expected to capture better phonetic structure then unsupervised labels. However, such difference is small and UA-based C-DSVAE is capable of encoding disentangled speaker and content representation which is also semantically meaningful. This provides the prerequisite for the success of the UTTS systems.

TABLE II
OBJECTIVE ANALYSIS OF CONTENT EMBEDDINGS AND SPEAKER EMBEDDINGS

	content embd		spk embd
model	Phn ACC (%)	SV EER (%)	SV EER (%)
DSVAE	30.5	40.5	4.8
C-DSVAE(UA)	53.6	43.2	3.9
C-DSVAE(UA+MP)	68.2	41.6	4.1
C-DSVAE(UA+MP+LibriTTS)	72.1	44.5	3.6
C-DSVAE(FA)	54.6	43.4	3.8
C-DSVAE(FA+MP)	70.5	42.3	3.6
C-DSVAE(FA+MP+LibriTTS)	73.3	45.6	3.3

2) Voice Conversion: The other way to evaluate the disentanglement of C-DSVAE system is to perform voice conversion, as mentioned in Section III-D.

We conduct a MOS test to evaluate the proposed C-DSVAE models for zero-shot non-parallel voice conversion. The evaluation setup is similar to [12] and we split the evaluation into to two sets. The first set is a collection of seen speakers sampled from training set and the second set is a collection of unseen speakers sampled from test set. Details on train/test splition of VCTK is introduced in V-A. We randomly draw 30 pairs from each set, with 10 pairs for male to male conversion, 10 pairs for female to female conversion and 10 pairs for cross-gender conversion. The listener needs to give a score for each sample in a test case according to the criterion: 1 = Bad; 2 = Poor;

TABLE III
MOS (95% CI) TEST ON VOICE CONVERSION SAMPLES WITH DIFFERENT C-DSVAE MODELS

model	seen to seen		unseen to unseen	
	naturalness	similarity	naturalness	similarity
DSVAE(WaveNet) [13]	3.38±0.08	3.46±0.07	3.36±0.10	3.56±0.08
DSVAE(HiFi-GAN) [12]	3.56±0.07	3.60±0.08	3.52±0.08	3.62±0.11
C-DSVAE(UA)	3.78±0.07	3.86±0.06	3.75±0.07	3.82±0.10
C-DSVAE(UA+MP)	**3.85±0.06**	**3.92±0.08**	**3.86±0.06**	**3.88±0.09**
C-DSVAE(UA+MP+LibriTTS)	3.80±0.07	3.78±0.09	3.82±0.09	3.86±0.08

3 = Fair; 4 = Good; 5 = Excellent. The final score for each model is calculated by averaging the collected results from 18 subjects. As introduced in the article, we employ HiFi-GAN [39] as the neural vocoder to convert the mel spectrogram to the waveform. We utilized the samples from two DSVAE based systems for a fair comparison [12]. Table III shows the MOS results of different models. Since we have different subjects for MOS test and there is always inevitable randomness in subjective evaluation. That is why we observe a difference in C-DSVAE performance between Table III and [12]. However, the trends that revealed from Table III are still in line with [12].

As illustrated in the Table III, comparing with the WaveNet, HiFi-GAN produces consistently better MOS scores given the same acoustic features. C-DSVAE with unsupervised alignment (UA) achieves 0.24 performance boost in contrast with the best DSVAE voice conversion system. Of all the systems, the C-DSVAE(UA+MP) produced the best objective scores in terms of naturalness and similarity under both seen to seen and unseen to unseen scenarios. And C-DSVAE(UA+MP+LibriTTS) with additional LibriTTS training data does not outperform the VCTK only C-DSVAE(UA+MP) system, which indicates that the domain mismatch exists in the self-supervised representation learning frameworks. It is noted that, the developed voice conversion systems are not sensitive to seen and unseen speech utterances, which is a good attribute from system generalization perspective. However, the relatively larger variance values in the unseen condition shows that unseen cases are still more challenging than the seen cases.

D. Evaluations on UTTSs

1) MOS Test: We conduct MOS tests in terms of naturalness and speaker similarity on VCTK speakers. 18 subjects are asked to make judgement about the synthesized speech samples. Since our UTTS system is a zero-shot voice cloning format, we also ask the subjects to evaluate the similarity between the UTTS voice and the target voice. We use a standard 5-scale rate to label the speech.As shown from Table IV, compared with the vanilla C-DSVAE(UA) AM, the masked unit prediction multi-task training in the content prior model of C-DSVAE does helps the overall speech quality. Having LibriTTS for pretraining helps but not so significantly. If the dual prior encoder training is applied to the C-DSVAE(UA+MP) system, both naturalness and similarity are improved clearly, showing the importance of prior modeling in this VAE based non-autoregressive TTS model. At the same time, the gap between synthesised speech and the ground truth in naturalness indicates that some form of prosody are currently missing in the proposed UTTS systems.

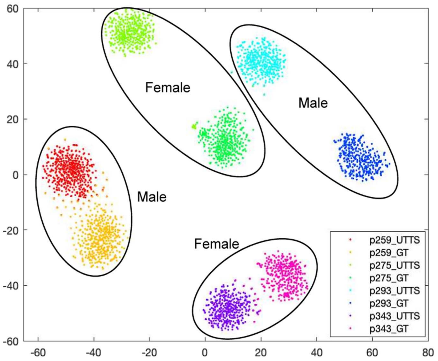


Fig. 5. T-SNE visualization of utterances of 4 VCTK speakers.

2) Speech Intelligibility: Results of intelligibility evaluation using a Librispeech ASR model are also listed in Table V. We generated 1600 utterances each model for this portion of study. Basically, the objective and subjective evaluation show a similar trend. Speech from the best UTTS model (C-DSVAE(UA+MP+Dual)) achieves a 15.1% WER, the gap between ground truth and UTTS speech suggests that the proposed method can be further improved by either improving the FA2UA mapping module or the content prior modeling. In Table V, we also list the ASR based speech intelligibility test results on UTTS samples by gender. As shown from the table, the ASR system does not show much performance preference w.r.t. different genders on real speech. However, the ASR results for male and female synthesised samples are showing a big difference. The much worse CER and WER for female generated speech across all 4 UTTS systems indicates that generating female voice seems to be more difficult in the multi-speaker TTS systems. This is also observed in Table IV, where gender dependent MOS scores indicates that the current systems are generating better male voices.

3) Speaker Similarity Between the Real and UTTS Speech: Since UTTS is a zero-shot voice cloning TTS system, Table III demonstrates how human evaluates the similarity between the synthesised and target speech. In the meantime, we are interested to know how machine perceives the similarity and difference. For this purpose, we use a pretrained speaker embedding system to extract speaker embeddings from real and synthesised utterances, and project them to a 2D T-SNE [40], [41] space. Fig. 5 illustrates the speaker embedding projections of 4 speakers (P259, P275, P293 and P343). P259 and P293 are used for the UTTS AM training, while P275 and P343 are never used in the training stage.

From the clusters in Fig. 5, it seems that each generated speaker localizes at its unique speaker space. The similarity between real and UTTS speech is showing some randomness. For example, P259 (male & seen) and P343 (female & unseen) are showing much obvious speaker similarities in speaker embedding space, while P275 (female & unseen) and P293 (male & seen) demonstrates a more separated speaker attribute in the same space. All the synthesised utterances are formulating a distinct distribution that is different with the target speaker

[3] [Online]. Available: https://huggingface.co/espnet/simpleoier_librispeech_asr_train_asr_conformer7_hubert_ll60k_large_raw_en_bpe5000_sp

TABLE IV
MULTI-SPEAKER UTTS MOS TEST OF NATURALNESS AND SIMILARITY ON VCTK

	seen target		**unseen target**		**male sample**		**female sample**	
UTTS model	naturalness	similarity	naturalness	similarity	naturalness	similarity	naturalness	similarity
GT	4.45				4.52		4.39	
C-DSVAE(UA)	3.22	3.56	3.28	3.51	3.38	3.64	3.26	3.44
C-DSVAE(UA+MP)	3.52	3.72	3.56	3.64	3.65	3.76	3.45	3.62
C-DSVAE(UA+MP+Dual)	**3.76**	**3.82**	**3.72**	**3.74**	**3.82**	**3.90**	**3.67**	**3.71**
C-DSVAE(UA+MP+Dual+LibriTTS)	3.58	3.63	3.61	3.70	3.68	3.73	3.56	3.60

TABLE V
ASR INTELLIGIBILITY TEST ON GENDER DEPENDENT UTTS SAMPLES

	overall		**male**		**female**	
UTTS model	CER (%)	WER (%)	CER (%)	WER (%)	CER (%)	WER (%)
GT	2.3	7.0	2.4	7.3	2.2	6.7
C-DSVAE(UA)	11.3	19.3	8.8	16.9	13.8	21.6
C-DSVAE(UA+MP)	9.7	16.5	6.1	12.8	13.3	20.2
C-DSVAE(UA+MP+Dual)	**8.3**	**15.1**	**5.2**	**11.5**	**11.4**	**18.7**
C-DSVAE(UA+MP+Dual+LibriTTS)	10.8	18.2	7.3	14.7	14.5	21.8

distribution, which indicates that the synthetic speech from the current UTTS systems still can be detected with a normal speaker verification system.

VI. CONCLUSION

In this study, we proposed a novel method to develop TTS systems without using the paired TTS training data. Specifically, we explored C-DSVAE, a self-supervised model that learned disentangled intermediate content and speaker representations with unsupervised alignment as the prior condition, and generated high quality speech with the latent speaker and content features using a decoder network. We employed C-DSVAE as the backbone module for developing the UTTS systems. For that purpose, we proposed a simple yet effective forced alignment to unsupervised alignment module to bridge the TTS front-end and the C-DSVAE. To further improve the speech quality and naturalness, we proposed to have a dual prior encoder reconstruction training strategy in our VAE based AM training, which significantly alleviated the mismatch between training and test phase of UTTS inference. Preliminary experiments demonstrated that our proposed UTTS systems can produce speech with an impressive quality. The newly developed method shows great potential for its self-supervised/unsupervised nature that can utilize general quality speech data for future needs.

At the same time, we found the UTTS generated samples still have some limitations in speech quality, speech expressiveness and speaker similarity. The the speech fidelity related issues can be addressed by introducing more advanced text encoding networks or methods [8], [42]. While the latter one can be improved with more data involved in the self-supervised or semi-supervised speaker representation training [27]. Thus, our study serves as the preliminary research work, proving the concept of unsupervised TTS AM, while leaving observed inadequacies as the future work.

REFERENCES

[1] Y. Wang et al., "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[2] C. Yu et al., "DurIAN: Duration informed attention network for speech synthesis," in *Proc. Interspeech*, 2020, pp. 2027–2031.

[3] Y. Ren et al., "Fastspeech: Fast, robust and controllable text to speech," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[4] Y. Ren et al., "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proc. Int. Conf. Learn. Representations*, 2021.

[5] C.-M. Chien, J.-H. Lin, C.-Y. Huang, P.-C. Hsu, and H.-Y. Lee, "Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 8588–8592.

[6] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 8067–8077,.

[7] J. Shen et al., "Natural TTs synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 4779–4783.

[8] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 5530–5540.

[9] D. Lim, S. Jung, and E. Kim, "JETS: Jointly training FastSpeech2 and HiFi-GAN for end to end text to speech," in *Proc. Interspeech*, 2022, pp. 21–25.

[10] 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 Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[11] S. Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.

[12] J. Lian, C. Zhang, G. K. Anumanchipalli, and D. Yu, "Towards improved zero-shot voice conversion with conditional DSVAE," in *Proc. Interspeech*, 2022, pp. 2598–2602.

[13] J. Lian, C. Zhang, and D. Yu, "Robust disentangled variational speech representation learning for zero-shot voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 6572–6576.

[14] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92), [sound]," *The Centre for Speech Technology Research*. Edinburgh, Scotland, U.K.: University of Edinburgh, 2017, doi: 10.7488/ds/2645.

[15] H. Zen et al., "Libritts: A corpus derived from librispeech for text-to-speech," in *Proc. Interspeech*, 2019, pp. 1526–1530.

[16] A. V. D. Oord et al., "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.

[17] J. Cong, S. Yang, L. Xie, and D. Su, "Glow-wavegan: Learning speech representations from GAN-based variational auto-encoder for high fidelity flow-based speech synthesis," in *Proc. Interspeech*, 2021, pp. 2182–2186.

[18] D. Lim, S. Jung, and E. Kim, "JETS: Jointly training FastSpeech2 and HiFi-GAN for end to end text to speech," in *Proc. Interspeech*, 2022, pp. 21–25.

[19] A. H. Liu, C.-I. Lai, W.-N. Hsu, M. Auli, A. Baevski, and J. Glass, "Simple and effective unsupervised speech synthesis," in *Proc. Interspeech*, 2022, pp. 843–847.

[20] J. Ni et al., "Unsupervised text-to-speech synthesis by unsupervised automatic speech recognition," in *Proc. Interspeech*, 2022, pp. 461–465.

[21] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021 vol. 34, pp. 27826–27839.

[22] A. H. Liu, W.-N. Hsu, M. Auli, and A. Baevski, "Towards end-to-end unsupervised speech recognition," 2022, *arXiv:2204.02492*.

[23] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1876–1887.

[24] L. Yingzhen and S. Mandt, "Disentangled sequential autoencoder," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5670–5679.

[25] A. Polyak et al., "Speech resynthesis from discrete disentangled self-supervised representations," in *Proc. Interspeech*, 2021, pp. 3615–3619.

[26] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5210–5219.

[27] C. Zhang and D. Yu, "C3-DINO: Joint contrastive and non-contrastive self-supervised learning for speaker verification," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1273–1283, Oct. 2022.

[28] A. Nagrani, J. S. Chung, S. Albanie, and A. Zisserman, "Disentangled speech embeddings using cross-modal self-supervision," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6829–6833.

[29] Y. Jia et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 4485–4495.

[30] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1996, vol. 1, pp. 339–341.

[31] Y. Zhu, M. R. Min, A. Kadav, and H. P. Graf, "S3VAE: Self-supervised sequential vae for representation disentanglement and data generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6538–6547.

[32] J. Han, M. R. Min, L. Han, L. E. Li, and X. Zhang, "Disentangled recurrent wasserstein autoencoder," in *Proc. Int. Conf. Learn. Representations*, 2021.

[33] J. Bai, W. Wang, and C. P. Gomes, "Contrastively disentangled sequential variational autoencoder," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 10105–10118.

[34] S. Yuan, P. Cheng, R. Zhang, W. Hao, Z. Gan, and L. Carin, "Improving zero-shot voice style transfer via disentangled representation learning," in *Proc. Int. Conf. Learn. Representations*, 2021.

[35] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Proc. Interspeech*, 2017, pp. 498–502.

[36] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An Asr Corpus Based on Public Domain Audio Books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.

[37] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2007, pp. 1027–1035.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[39] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 17022–17033.

[40] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[41] C. Zhang, K. Koishida, and J. H. Hansen, "Text-independent speaker verification based on triplet convolutional neural network embeddings," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 26, no. 9, pp. 1633–1644, Sep. 2018.

[42] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 6840–6851.

Jiachen Lian (Member, IEEE) received the B.S. degree in electrical engineering from Zhejiang University, Hangzhou, China, in 2019 and the M.S. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2021. He is currently working toward the Ph.D. degree in EECS from UC Berkeley, Berkeley, CA, USA. His research interests include deep learning and speech processing.

Chunlei Zhang (Member, IEEE) received the B.S. degree in environmental engineering and the M.S. degree in acoustics from Northwestern Polytechnical University, Xi'an, China, in 2011 and 2014, respectively, and the Ph.D. degree in electrical and computer engineering from The University of Texas at Dallas, Richardson, TX, USA, in 2018. He joined the Tencent AI Lab, Bellevue, WA, USA, in 2019, where he is currently a Senior Research Scientist. His research interests include automatic speech recognition, speaker recognition/diarization and text-to-speech synthesis. Since 2016, he has been a Member of the ISCA and IEEE.

Gopala K. Anumanchipalli (Member, IEEE) received the B.T. and M.S. degree in computer science from the International Institute of Information Technology, Hyderabad, India, in 2008 and the Ph.D. degree in language and information technologies from Carnegie Mellon University, Pittsburgh, PA,USA, and in electrical and computer engineering from IST, Lisbon, in 2013. He joined the Faculty of the Department of Electrical Engineering and Computer Sciences with UC Berkeley in Spring 2021 and continues to hold an adjunct position with the Department of Neurosurgery, UC San Francisco, San Francisco, CA, USA. He works at the intersection of Speech Processing, Neuroscience, and Artificial Intelligence with an emphasis on human-centered speech and Assistive technologies, including new paradigms for bio-inspired spoken language technologies, automated methods for early diagnosis, characterizing and rehabilitating disordered speech.

Dong Yu (Fellow, IEEE) with the Tencent AI Lab as a distinguished scientist and vice general Manager. Prior to joining Tencent in 2017, he was a Principal Researcher with Microsoft Research (Redmond), where he has been since 1998. He has authored or coauthored two monographs and more than 300 papers. His research interests include speech recognition and processing and natural language processing. His works have been widely cited and recognized by the prestigious IEEE Signal Processing Society best transaction paper award in 2013, 2016, 2020, and 2022, the 2021 NAACL best long paper award, 2022 IEEE Signal Processing Magazine best paper award, and 2022 IEEE Signal Processing Magazine best column award. Dr. Dong Yu was the Chair of the IEEE Speech and Language Processing Technical Committee during 2021–2022. He was on the editorial boards of numerous journals and magazines, as well as on the organizing and technical committees of various conferences and workshops. He is currently an ACM/IEEE/ISCA Fellow.