Disentangling Correlated Speaker and Noise for Speech Synthesis via Data Augmentation and Adversarial Factorization

Wei-Ning Hsu[1*] **Yu Zhang**[2] **Ron J. Weiss**[2]
Yu-An Chung[1*] **Yuxuan Wang**[2] **Yonghui Wu**[2] **James Glass**[1]
[1]Massachusetts Institute of Technology [2]Google Inc.
wnhsu@csail.mit.edu, {ngyuzh,ronw}@google.com

Abstract

To leverage crowd-sourced data to train multi-speaker text-to-speech (TTS) models that can synthesize clean speech for all speakers, it is essential to learn disentangled representations which can independently control the speaker identity and background noise in generated signals. However, learning such representations can be challenging, due to the lack of labels describing the recording conditions of each training example, and the fact that speakers and recording conditions are often correlated, e.g. since users often make many recordings using the same equipment. This paper proposes three components to address this problem by: (1) formulating a conditional generative model with factorized latent variables, (2) using data augmentation to add noise that is not correlated with speaker identity and whose label is known during training, and (3) using adversarial factorization to improve disentanglement. Experimental results demonstrate that the proposed method can disentangle speaker and noise attributes even if they are correlated in the training data, and can be used to consistently synthesize clean speech for all speakers. Ablation studies verify the importance of each proposed component.

1 Introduction

Recent development of neural end-to-end TTS models [27, 2] enables control of both labelled and unlabelled speech attributes by conditioning synthesis on both text and learned attribute representations [28, 22, 11, 1, 6, 10]. This opens the door to leveraging crowd-sourced speech recorded under various acoustic conditions [19] to train a high-quality multi-speaker TTS model that is capable of consistently producing clean speech. To achieve this, it is essential to learn disentangled representations that control speaker and acoustic conditions independently. However, this can be challenging for two reasons. First, the underlying acoustic conditions of an utterance, such as the type and level of background noise and reverberation, are difficult to annotate, and therefore such labels are often unavailable. This hinders the use of direct conditioning on the acoustic condition labels in a way similar to conditioning on one-hot speaker labels [2]. Second, speaker identity can have strong correlations with recording conditions, since a speaker might make most of their recordings in the same location using the same device. This makes it difficult to learn a disentangled representation by assuming statistical independence [7].

We address this scenario by introducing three components: a conditional generative model with factorized latent variables to control different attributes, data augmentation by adding background noise to training utterances in order to counteract the inherent speaker-noise correlation and to create ground truth noisy acoustic condition labels, and adversarial training based on the generated labels to encourage disentanglement between latent variables. We utilize the VCTK speech synthesis dataset [24], and background noise signals from the CHiME-4 challenge [25] to synthesize a dataset

[*]Work performed while interning at Google.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

containing correlated speaker and background noise conditions for controlled experiments. We extensively evaluate disentanglement performance on the learned latent representations as well as the synthesized samples. Experimental results identify the contribution of each component, and demonstrate the ability of the proposed model to disentangle noise from speakers and consistently synthesize clean speech for all speakers, despite the strong correlation in the training data.

2 Proposed Method

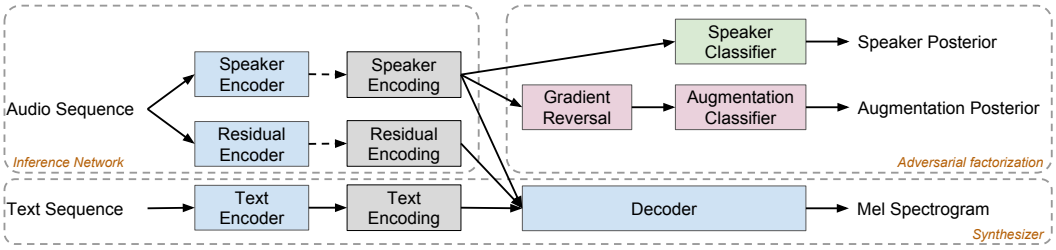


Figure 1: Overview of the components of the proposed model. Dashed lines denote sampling via reparameterization [15].

2.1 Conditional factorized variational autoencoder for TTS

We base our TTS model on Tacotron 2 [21], which takes a text sequence as input, and outputs a sequence of mel spectrogram frames. To control speech attributes other than text, two additional latent variables, $\mathbf{z}_s$ and $\mathbf{z}_r$, are introduced to condition the generative process, where the former models speaker identity, and the latter models residual unlabelled attributes (e.g. acoustic conditions). Prior distributions for both variables are defined to be isotropic Gaussian. The full TTS model can be written as a conditional generative model with two latent variables: $p(speech \mid \mathbf{z}_s, \mathbf{z}_r, text)$.

Two variational distributions are introduced: $q(\mathbf{z}_s \mid speech)$ and $q(\mathbf{z}_r \mid speech)$, to approximate the intractable posteriors of the latent variables, following the variational autoencoder (VAE) framework [15]. Each distribution is defined to be diagonal-covariance Gaussian, whose mean and variance are parameterized by a neural network encoder. Note that $\mathbf{z}_s$, $\mathbf{z}_r$, and $text$ are assumed to be conditionally independent given $speech$, in order to simplify inference. In contrast to learning an embedding for each speaker, learning an inference model for $\mathbf{z}_s$ can be used to infer speaker attributes for previously unseen speakers.

To factorize speaker and residual information, an auxiliary speaker classifier that takes $\mathbf{z}_s$ as input is trained jointly with the TTS model. This encourages information that is discriminative between speakers to be encoded in $\mathbf{z}_s$, and leaves residual information to $\mathbf{z}_r$. A simple fully-connected network is used for the speaker classifier.

2.2 Speaker invariant data augmentation

When acoustic conditions are correlated with speakers, information about e.g. background noise level can be used to discriminate between speakers, and therefore can be encoded into $\mathbf{z}_s$. To counteract such behavior, one can decorrelate these factors by leveraging prior knowledge that adding noise should not affect speaker identity.

We propose to augment the original training set with a noisy copy that mixes each utterance with a randomly selected piece of background noise at a randomly sampled signal-to-noise ratio (SNR), but reuses the same transcript and speaker label as the original utterance. This operation can be seen as flattening the SNR distribution of each speaker, in order to make SNRs less discriminative about speakers.

2.3 Augmentation-adversarial training

To increase the degree of disentanglement, it is also useful to proactively discourage $\mathbf{z}_s$ from encoding acoustic condition information. If the ground truth acoustic condition labels are available, domain

adversarial training [4] can be applied directly to encourage $\mathbf{z}_s$ not to be informative about the acoustic condition. Nevertheless, such labels are often unavailable in crowdsourced datasets such as [19].

In order to utilize adversarial training in such a scenario, we propose to use the augmentation label (original/augmented) to replace the acoustic condition label (clean/noisy). This augmentation label can be seen as a noisy acoustic condition label: *an augmented utterance must be noisy, but an original one can be either*. If $\mathbf{z}_s$ is invariant to acoustic conditions, then it is also invariant to augmentation labels, implying that the latter is a necessary condition for the former.

Following [4], invariance of $\mathbf{z}_s$ to augmentation is measured using the empirical $\mathcal{H}$-divergence between the $\mathbf{z}_s$ distribution of the augmented data and that of the original data, given a hypothesis class $\mathcal{H}$ that is a set of binary classifiers. The empirical $\mathcal{H}$-divergence measures how well the best classifier in the hypothesis class can distinguish between samples drawn from different distributions. However, it is generally hard to compute the empirical $\mathcal{H}$-divergence. Following [3, 4], we approximate it with the *Proxy $\mathcal{A}$-distance*: $2(1-2\epsilon)$, where ϵ is a generalization error of an augmentation classifier trained to predict if $\mathbf{z}_s$ is inferred from an augmented utterance. A simple fully-connected network is used for the augmentation classifier.

2.4 Model and training objective

The complete model is illustrated in Figure 1, composed of three modules: a synthesizer, $p(speech \mid \mathbf{z}_s, \mathbf{z}_r, text)$, an inference network with two encoders, $q(\mathbf{z}_s \mid speech)$ and $q(\mathbf{z}_r \mid speech)$, and an adversarial factorization module with speaker and augmentation classifiers, $p(\mathbf{y}_s \mid \mathbf{z}_s)$ and $p(\mathbf{y}_a \mid \mathbf{z}_r)$, where $\mathbf{y}_s$ and $\mathbf{y}_a$ denotes speaker and augmentation labels. The parameters of the synthesizer, the two encoders, the speaker classifier, and the augmentation classifiers are accordingly denoted as θ, ϕ_s, ϕ_r, ψ_s, and ψ_a, respectively.

Training of the proposed model aims to maximize the conditional likelihood and the information $\mathbf{z}_s$ contains about speakers, while minimizing the $\mathcal{H}$-divergence between the $\mathbf{z}_s$ inferred from the original utterances and that from the augmented ones. The $\mathcal{H}$-divergence is approximated with the Proxy $\mathcal{A}$-distance obtained from the augmentation classifier. The objective function can be formulated as combining an evidence lower bound (ELBO) with a domain adversarial training [4] objective:

$$\begin{aligned}
&\mathcal{L}_1(\theta, \phi_s, \phi_r, \psi_s; speech, text, \mathbf{y}_s, \mathbf{y}_a) \\
&\quad = ELBO(\theta, \phi_s, \phi_r; speech, text) \\
&\qquad + \mathbb{E}_{q(\mathbf{z}_s \mid speech)}[\lambda_1 \log p(\mathbf{y}_s \mid \mathbf{z}_s) - \lambda_2 \log p(\mathbf{y}_a \mid \mathbf{z}_s)] \qquad (1) \\
&\mathcal{L}_2(\psi_a; speech, \mathbf{y}_a) = \mathbb{E}_{q(\mathbf{z}_s \mid speech)}[\log p(\mathbf{y}_a \mid \mathbf{z}_s)], \qquad (2)
\end{aligned}$$

where $\lambda_1, \lambda_2 > 0$ are the loss weights for the two classifiers, and $ELBO(\theta, \phi_s, \phi_r; speech, text)$ is formulated as:

$$\begin{aligned}
&\mathbb{E}_{q(\mathbf{z}_s \mid speech) q(\mathbf{z}_r \mid speech)}[\log p(speech \mid \mathbf{z}_s, \mathbf{z}_r, text)] \\
&\quad - D_{KL}\big(q(\mathbf{z}_s \mid speech) \,||\, \mathcal{N}(\mathbf{0}, \mathbf{I})\big) \\
&\quad - D_{KL}\big(q(\mathbf{z}_r \mid speech) \,||\, \mathcal{N}(\mathbf{0}, \mathbf{I})\big).
\end{aligned}$$

Note that the augmentation classifier is optimized with a different objective than the rest of the model. To train the entire model jointly, a gradient reversal layer [4] is inserted after the input to the augmentation classifier, which scales the gradient by $-\lambda_2$.

3 Related Work

Our formulation of a TTS model with latent variables are closely related to [28, 22, 1, 6, 10], which focus on modeling unlabeled speech attributes. In contrast to this work, [28, 22, 1, 6] do not address disentangling attributes to enable independent control when different attributes are highly correlated in the training data, while [10] learns to disentangle speaker attributes from the rest by encoding those with small within-speaker variance to $\mathbf{z}_s$.

The proposed augmentation-adversarial training combines data augmentation for speech [12] with domain adversarial neural networks (DANNs) [4] for disentangling correlated attributes. These two

methods have been mainly applied for training robust discriminative models [8, 23, 25, 20], and are less studied in the context of building generative models. In addition, our method provides two advantages. First, while DANNs require domain labels, our proposed method enables adversarial training even when the ground truth domain labels are unavailable. Second, domain adversarial training aims to remove domain information while preserving target attribute information; however, if domain and target attribute have very strong correlations, the two objectives conflict with each other, and one of the them will be compromised. Our proposed method alleviates such issues by using data augmentation to decorrelate the two factors.

Learning disentangled representations for deep generative models has gain much interest recently [9, 29]. Several studies also explored adversarial training for disentanglement, such as using maximum mean discrepancy [16] and generative adversarial network [18]. We particularly emphasize disentangling statistically correlated attributes, and apply $\mathcal{H}$-divergence based adversarial training on latent variables.

4 Experiments

We artificially generated a noisy speech dataset with correlated speaker and noise conditions using the VCTK corpus [24] and background noise from the CHiME-4 challenge [25]. The motivation here is to simulate real noisy data while evaluating the model under carefully controlled conditions. VCTK contains 44 hours of clean English speech from 109 speakers. We downsample the signals to 16 kHz to match the background noise sample rate, and split it into training and testing sets in a 9:1 ratio. The CHiME-4 corpus contains 8.5 hours of background noise recorded in four different locations (bus, cafe, pedestrian area, and street junction), which we split into three partitions: `train`, `test`, and `aug`. To simulate speaker-correlated noise, we randomly selected half the speakers to be noisy, and mixed all of their train and test utterances with noise sampled from `train` and `test` respectively, at SNRs ranging from 5 - 25 dB. As described in Section 2.2, we generated an augmented set by mixing every (potentially noisy) training utterance with a noise signal sampled from `aug` at SNRs ranging from 5 - 25 dB. Utterances in the augmented set are annotated with $\mathbf{y}_a = 1$, and those in the original noisy training set are annotated with $\mathbf{y}_a = 0$. We strongly encourage readers to listen to the samples on the demo page.[2]

4.1 Model and training setup

The synthesizer network use the sequence-to-sequence Tacotron 2 architecture [21], with extra input $\mathbf{z}_s$ and $\mathbf{z}_r$ concatenated and passed to the decoder at each step. If not otherwise mentioned, $\mathbf{z}_s$ is 64-dim and $\mathbf{z}_r$ is 8-dim. The generated speech is represented as a sequence of 80-dim mel-scale filterbank frames, computed from 50ms windows shifted by 12.5ms. We represent input text as a sequence of phonemes, since learning pronunciations from text is not our focus.

The speaker and the residual encoders both use the same architecture which closely follow the attribute encoder in [10]. Each encoder maps a variable length mel spectrogram to two vectors parameterizing the mean and log variance of the Gaussian posterior. Both classifiers are fully-connected networks with one 256 unit hidden layer followed by a softmax layer to predict the speaker or augmentation posterior.

The synthesizer, encoders, and speaker classifier are trained to maximize Eq (1) with $\lambda_1 = \lambda_2 = 1$, and the augmentation classifier is trained to maximize Eq (2). The entire model is trained jointly with a batch size of 256, using the Adam optimizer [14], configured with an initial learning rate of 10^{-3}, and an exponential decay that halves the learning rate every 12.5k steps, starting at 50k steps.

4.2 Latent space disentanglement

We quantify the degree of disentanglement by training speaker and noise classifiers on $\mathbf{z}_s$ and $\mathbf{z}_r$ separately. The classification accuracy on a held-out set is used to measure how much information a latent variable contains about the prediction targets. A simple linear discriminative analysis classifier is used for all four tasks. If the classifier input contains no information about the target, the best a classifier can do is to predict the highest prior probability class. Since the distributions of both

[2] `https://google.github.io/tacotron/publications/adv_tts`

speaker and acoustic conditions are close to uniform, a speaker-uninformative input should result in about 1% accuracy, and a noise-uninformative input should result in about 50%.

Results are shown in Table 1, comparing the full proposed model with two alternative models: one which removes adversarial training, denoted as "- *adv*,", and a second which further removes data augmentation, denoted as "- *adv - aug*." Without data augmentation and adversarial training, the second alternative completely fails to disentangle speaker from noise, i.e. its speaker encoding $\mathbf{z}_s$ can infer both, while its residual encoding $\mathbf{z}_s$ cannot infer either. The first alternative learns to encode acoustic condition into $\mathbf{z}_r$, reaching 96.5% accuracy on noise prediction; however, part of such information still leaks to $\mathbf{z}_s$, as indicated by the 85% noise prediction accuracy. The full proposed model achieves the highest noise prediction accuracy using $\mathbf{z}_r$, and the lowest accuracy using $\mathbf{z}_s$, implying the best allocation of acoustic information. Nevertheless, adversarial training also results in slight degradation of speaker information allocation, where the speaker prediction accuracy using $\mathbf{z}_r$ increases from 1.4% to 2.3%.

Table 1: Accuracy (%) of speaker and noise classifiers trained on $\mathbf{z}_s$ or $\mathbf{z}_r$ on a held-out set.

Model	$\mathbf{z}_s$		$\mathbf{z}_r$	
	speaker	noise	noise	speaker
Proposed	97.58	**60.20**	**97.44**	2.33
- adv	97.64	85.35	96.53	1.40
- adv - aug	93.68	97.93	51.17	1.13

4.3 Visualization

We further analyze the latent space of the proposed model by visualizing the learned speaker and residual representations using t-SNE [17], which is a technique for projecting high-dimensional vectors to a two-dimensional space. Results are shown in Figure 2, where each point corresponding to a projected $\mathbf{z}_r$ (left column) or $\mathbf{z}_s$ (right column) inferred from a single utterance. Points are color-coded according to speaker, gender, and accent labels in each row.

In the left column, projected $\mathbf{z}_r$ are clearly separated by acoustic condition, but not by gender or speaker. In contrast, projected $\mathbf{z}_s$ shown in the right column forms many small clusters, with one speaker each cluster; Moreover, as shown in the middle row, clusters of speakers are further separated according to their genders. In the lower right panel, projected $\mathbf{z}_s$ of noisy utterances and clean utterances are overlaid, demonstrating that $\mathbf{z}_s$ have similar distributions conditioning on different acoustic conditions.

4.4 Evaluation of synthesized speech

To evaluate how well the two latent variables, $\mathbf{z}_s$ and $\mathbf{z}_r$, can control the synthesized speech, we sample five clean speakers and five noisy speakers, and select one testing utterance for each speaker with duration $\geq$ 3s. For each of the ten utterances, the two latent variables are inferred using the corresponding encoders. We construct an evaluation set of 100 phrases that does not overlap with the VCTK corpus, and synthesize them conditioned on each combination of $\mathbf{z}_r$ and $\mathbf{z}_s$, including those inferred from different utterances. The total 10,000 synthesized samples are divided into four groups, depending on the set of speakers (clean/noisy) $\mathbf{z}_r$ and $\mathbf{z}_s$ are inferred from.

4.4.1 Control of acoustic conditions

To quantify the ability to control noise, we use waveform amplitude distribution analysis (WADA) [13] to estimate an SNR without a clean reference signal. We compare to a baseline multi-speaker Tacotron model, which removes the residual encoder and replaces the speaker encoder with a lookup table of 64-D speaker embeddings. The upper half of Table 2 presents the estimated SNRs of synthesized speech using this baseline, conditioning on the same five clean speakers and the five noisy speakers mentioned above. The difference in SNR between clean and noisy speakers indicates that the acoustic condition is tied to speaker identity in this baseline model.

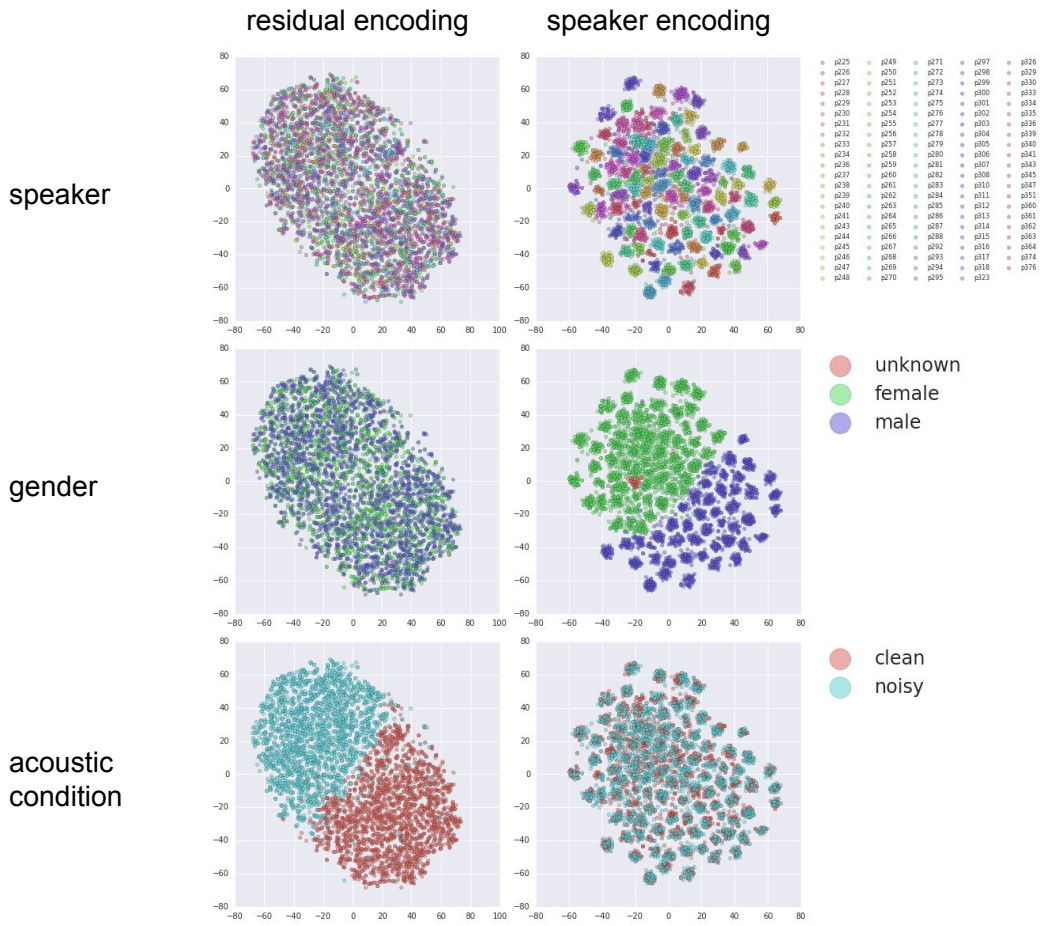
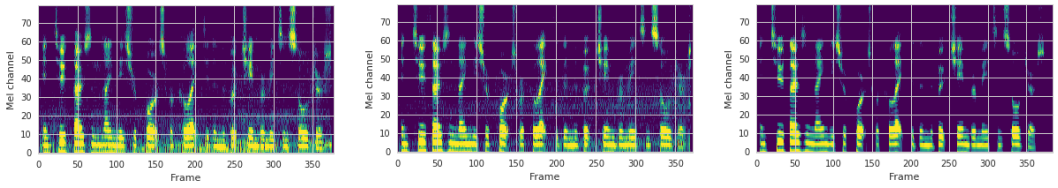


Figure 2: Visualization of learned speaker and residual encodings using two-dimensional t-SNE projected embeddings, colored coded by speaker, gender, and acoustic condition labels.



(a) Baseline (b) Proposed model with noisy $\mathbf{z}_r$ (c) Proposed model with clean $\mathbf{z}_r$

Figure 3: Synthesized utterances of a noisy speaker (p252) from the baseline model (a), and our proposed model conditioning on $\mathbf{z}_s$ inferred from a noisy utterance of p252, along with two $\mathbf{z}_r$ from different speakers, one from a noisy utterance (b), and the other from a clean utterance (c).

Results of the proposed model and the two alternatives mentioned in Section 4.2 are shown in the lower half of Table 2. By conditioning on $\mathbf{z}_r$ inferred from clean utterances, the proposed model is able to synthesize clean speech even for noisy speakers whose training utterances all had background noise. Moreover, when conditioning on the same set of $\mathbf{z}_r$, the proposed achieves the smallest discrepancy in SNR between different $\mathbf{z}_s$ sets. On the other hand, the "-adv" variant has a larger discrepancy between different $\mathbf{z}_s$ sets, indicating worse disentanglement comparing to the full model, while the "-adv-aug" variant fails to control noise through $\mathbf{z}_r$. These results are in line with the noise prediction results using $\mathbf{z}_s$ and $\mathbf{z}_r$ shown in Table 1. Figure 3 illustrates synthesized samples for a noisy speaker, comparing the baseline to our proposed model. Our model is capable of controlling

Table 2: Average WADA-SNR of synthesized samples from a multi-speaker baseline conditioning on different speaker embeddings, and the proposed model and the two alternatives conditioning on different $(\mathbf{z}_r, \mathbf{z}_s)$ combinations. "C" denotes latents inferred from clean testing utterances of the clean speaker set, and "N" denotes those inferred from noisy testing utterances of the noisy speaker set.

Model	clean speakers		noisy speakers	
Baseline	18.16		11.26	
Model	(set of $\mathbf{z}_r$, set of $\mathbf{z}_s$)			
	(C, C)	(C, N)	(N, C)	(N, N)
Proposed	18.62	**18.35**	**8.89**	**8.62**
- adv	18.64	17.03	10.43	8.59
- adv - aug	**18.78**	9.49	18.80	9.50

noise using $\mathbf{z}_r$, and can generate clean speech for the noisy speaker, while the baseline output always contains background noise.

4.4.2 Control of speaker identity

We next examine if $\mathbf{z}_s$ can control the speaker identity of synthesized speech, using a text-independent speaker verification system [26] to is compute speaker discriminative embeddings, called *d-vector*s [5], from the reference and synthesized speech samples. The system is trained to optimize a generalized end-to-end speaker verification loss, so that the embeddings of two utterances are close to each other if they are from the same speaker, and far way if from different speakers.

We build a nearest-neighbor classifier, which assigns an input signal the speaker label of the reference signal whose d-vector is closest to that of the input, measured using Euclidean distance. To prevent background noise from affecting d-vector quality, we only evaluate synthesized samples conditioned on $\mathbf{z}_r$ from clean utterances. Table 3 shows that the synthesized samples closely resemble the speaker characteristics of their corresponding reference samples, regardless of $\mathbf{z}_r$ used for conditioning. The results indicate that speaker identity is controlled by $\mathbf{z}_s$, while being invariant to change in $\mathbf{z}_r$.

Table 3: Speaker classification accuracy (%) of clean synthesized samples conditioning on $\mathbf{z}_s$ inferred from clean and noisy utterances.

Model	$\mathbf{z}_s$ from clean utt	$\mathbf{z}_s$ from noisy utt
Proposed	99.92	98.36

4.4.3 Subjective naturalness evaluation

To quantify fidelity, we rely on crowd-sourced mean opinion scores (MOS), which rates the naturalness of the synthesized samples by natives speakers using headphones, with scores ranging from 1 to 5 in 0.5 increments. To quantify fidelity, we rely on crowd-sourced mean opinion scores (MOS), which rates the naturalness of the synthesized samples by natives speakers using headphones, with scores ranging from 1 to 5 in 0.5 increments. Results shown in Table 4 compares the baseline and the proposed model conditioning on $\mathbf{z}_r$ from clean utterances. When conditioning on $\mathbf{z}_r$ from clean utterances, the proposed model achieves a higher MOS score than the baseline. In contrast, the MOS drops significantly when conditioning on $\mathbf{z}_r$ inferred from noisy utterances. The results indicate that disentangling speaker and noise improves the naturalness of the generated speech, and the proposed model can synthesize more natural speech with less background noise than the baseline when conditioning on $\mathbf{z}_r$ inferred from clean signals.

4.5 Hyperparameter sensitivity

Finally, we study the sensitivity of disentanglement performance with respect to the choice of speaker encoding dimensions. As shown in the previous two sections, good latent space disentanglement

Table 4: MOS scores of the baseline and the proposed model.

Baseline	Proposed w/ clean $\mathbf{z}_r$	Proposed w/ noisy $\mathbf{z}_r$
3.22	**4.52**	1.32

translates to good performance in terms of control of speaker identity and acoustic conditions for synthesis. In this section, we only evaluate latent space disentanglement when changing the dimension of $\mathbf{z}_s$

Table 5 compares performance of the proposed model when the dimensionality of $\mathbf{z}_s$ is 32, 64, 128, and 256. Variants without data augmentation or adversarial training fail to disentangle in all configurations. When the dimension of $\mathbf{z}_s$ increases, both the proposed model and "-adv" report worse separation of information, as indicated by increased noise prediction accuracy using $\mathbf{z}_s$. Specifically, the "-adv" fails to encode noise information in $\mathbf{z}_r$ when $\mathbf{z}_s$ has 128 dimensions, which could result from a bad initialization of model parameters; however, such a behavior also indicates that when adversarial training is not applied, the disentanglement performance may rely heavily on the model initialization. On the other hand, the proposed model is least sensitive to the change of $\mathbf{z}_s$ dimensionality. It always achieves the highest noise prediction accuracy using $\mathbf{z}_r$, and the lowest noise prediction accuracy using $\mathbf{z}_s$.

Table 5: Held-out set accuracy (%) of speaker and acoustic condition classifiers trained on $\mathbf{z}_s$ or $\mathbf{z}_r$ with different $\mathbf{z}_s$ dimensions.

dim($\mathbf{z}_s$)	Model	$\mathbf{z}_s$		$\mathbf{z}_r$	
		speaker	noise	noise	speaker
32	Proposed	97.58	**57.66**	**97.62**	2.80
	- adv	97.62	81.41	97.51	1.82
	- adv - aug	92.82	98.02	55.17	2.31
64	Proposed	97.58	**60.20**	**97.44**	2.33
	- adv	97.64	85.35	96.53	1.40
	- adv - aug	93.68	97.93	51.17	1.13
128	Proposed	97.67	**65.15**	**97.40**	2.09
	- adv	97.64	97.29	52.72	0.98
	- adv - aug	93.48	98.15	52.50	1.00
256	Proposed	97.53	**64.80**	**97.31**	2.80
	- adv	97.64	97.98	85.90	1.05
	- adv - aug	93.95	98.15	53.66	1.18

5 Conclusion and Future Work

We build a neural network TTS model which incorporates conditional generative modeling, data augmentation, and adversarial training to learn disentangled representations of correlated and partially unlabeled attributes, which can be used to independently control different aspects of the synthesized speech. Extensive studies on a synthetic dataset verify the effectiveness of each element of the proposed solution, and demonstrate the robustness to the choice of hyperparameters.

The proposed methods for disentangling correlated attributes is general, and can potentially be applied to other pairs of correlated factors, such as reverberation and speaker, or to other modalities, such as controllable text-to-image generation. In addition, for future work, we would also like to investigate the capability of the proposed method to disentangle pairs of attributes which are both unsupervised.

6 Acknowledgement

The authors thank Heiga Zen, Eric Battenberg, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathen Shen, and the Google Brain and Machine Perception teams for their helpful feedback and discussions.

References

[1] Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. Expressive speech synthesis via modeling expressions with variational autoencoder. In *Interspeech*, 2018.

[2] Sercan Arik, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. Deep Voice 2: Multi-speaker neural text-to-speech. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[3] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

[4] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research (JMLR)*, 17(1):2096–2030, 2016.

[5] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer. End-to-end text-dependent speaker verification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[6] Gustav Eje Henter, Jaime Lorenzo-Trueba, Xin Wang, and Junichi Yamagishi. Deep encoder-decoder models for unsupervised learning of controllable speech synthesis. *arXiv preprint arXiv:1807.11470*, 2018.

[7] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017.

[8] Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation. In *Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017.

[9] Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[10] Wei-Ning Hsu, Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathan Shen, Patrick Nguyen, and Ruoming Pang. Hierarchical generative modeling for controllable speech synthesis. *arXiv preprint arXiv:1810.07217*, 2018.

[11] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in Neural Information Processing Systems (NIPS)*, 2018. to appear.

[12] Chanwoo Kim, Ananya Misra, Kean Chin, Thad Hughes, Arun Narayanan, Tara Sainath, and Michiel Bacchiani. Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home. In *Interspeech*, 2017.

[13] Chanwoo Kim and Richard M Stern. Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. In *Interspeech*, 2008.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[15] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

[16] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. In *International Conference on Learning Representations (ICLR)*, 2016.

[17] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[18] Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[19] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[20] David Pearce and J Picone. Aurora working group: DSR front end LVCSR evaluation au/384/02. *Inst. for Signal & Inform. Process., Mississippi State Univ., Tech. Rep*, 2002.

[21] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[22] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, and Rif A. Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron. In *International Conference on Machine Learning (ICML)*, 2018.

[23] Sining Sun, Binbin Zhang, Lei Xie, and Yanning Zhang. An unsupervised deep domain adaptation approach for robust speech recognition. *Neurocomputing*, 257:79–87, 2017.

[24] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit, 2017.

[25] Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language*, 46:535–557, 2017.

[26] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[27] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis. In *Interspeech*, 2017.

[28] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A. Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning (ICML)*, 2018.

[29] Li Yingzhen and Stephan Mandt. Disentangled sequential autoencoder. In *International Conference on Machine Learning (ICML)*, 2018.