Neural Networks 143 (2021) 250–260

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Effective and direct control of neural TTS prosody by removing interactions between different attributes

Xiaochun An [a], Frank K. Soong [b], Shan Yang [a], Lei Xie [a,*]

[a] Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[b] Microsoft China

ARTICLE INFO

Article history:
Received 29 October 2020
Received in revised form 24 May 2021
Accepted 3 June 2021
Available online 11 June 2021

Keywords:
Neural TTS
Prosody control
Variational autoencoder (VAE)
Prosodic features
Mutual information minimization

ABSTRACT

End-to-end TTS advancement has shown that synthesized speech prosody can be controlled by conditioning the decoder with speech prosody attribute labels. However, to annotate quantitatively the prosody patterns of a large set of training data is both time consuming and expensive. To use unannotated data, variational autoencoder (VAE) has been proposed to model individual prosody attribute as a random variable in the latent space. The VAE is an unsupervised approach and the corresponding latent variables are in general correlated with each other. For more effective and direct control of speech prosody along each attribute dimension, it is highly desirable to disentangle the correlated latent variables. Additionally, being able to interpret the disentangled attributes as speech perceptual cues is useful for designing more efficient prosody control of TTS. In this paper, we propose two attribute separation schemes: (1) using 3 separate VAEs to model the real-valued, different prosodic features, i.e., F_0, energy and duration; (2) minimizing mutual information between different prosody attributes to remove their mutual correlations, for facilitating more direct prosody control. Experimental results confirm that the two proposed schemes can indeed make individual prosody attributes more interpretable and direct TTS prosody control more effective. The improvements are measured objectively by F_0 Frame Error (FFE) and subjectively with MOS and A/B comparison listening tests, respectively. The scatter diagrams of t-SNE also demonstrate the correlations between prosody attributes, which are well disentangled by minimizing their mutual information. Synthesized TTS samples can be found at https://xiaochunan.github.io/prosody/index.html.

1. Introduction

Recent advancement of neural TTS has demonstrated that it can synthesize very natural, human-like speech (Arik, et al., 2017; Ping, et al., 2018; Shen, et al., 2018; Sotelo, et al., 2017; Sun, et al., 2020; Wang, et al., 2017; Yang, et al., 2020; Yang, Yang, Zhu, Yan, & Xie, 2019). The trained neural TTS models usually consist of an encoder–decoder neural network (Bahdanau, Cho, & Bengio, 2015; Sutskever, Vinyals, & Le, 2014) which can map a text sequence to a sequence of speech frames. Extensions of these models have shown that various speech attributes, such as speaker identity and emotional expression, can also be controlled by conditioning the decoder upon additional attribute labels (Park, Zhao, Peng, & Ping, 2019; Wu, et al., 2019). Speech prosody attributes, including: fundamental frequency (F_0), energy and duration, are the key features for representing TTS prosody patterns (Wagner & Watson, 2010) which convey the linguistic, semantic, and emotional components embedded in speech. Explicit prosody control is highly desirable for producing natural, expressive speech for more engaging TTS applications, e.g., human–machine dialogue. Manual annotating a large set of speech data for training a neural TTS is usually expensive and time consuming. Recently, neural TTS model with a reference encoder has received interests for prosody control (Skerry-Ryan, et al., 2018; Stanton, Wang, & Skerry-Ryan, 2018). The idea of a reference encoder based neural TTS is to learn the latent representations via a reference encoder to render the target prosody in the synthesized speech. In Skerry-Ryan, et al. (2018), the authors augment neural TTS model with a reference encoder that extracts a learned representation of prosody from acoustic input which allows a prosody transfer. Global style tokens (GST) based model (Wang, et al., 2018) introduces a multi-head attention (Vaswani, et al., 2017) based style token layer to learn a small but rich prosody information in multiple dimensions. In this way, GST can control the prosody of TTS speech by providing style tokens with the learned prosody.

Though the GST-based model (Wang, et al., 2018) outperforms the reference encoder based structure (Skerry-Ryan, et al., 2018)

* Corresponding author.
E-mail addresses: xiaochunan@npu-aslp.org (X. An), frankkps@microsoft.com (F.K. Soong), syang@nwpu-aslp.org (S. Yang), lxie@nwpu-aslp.org, lxie@nwpu.edu.cn (L. Xie).

https://doi.org/10.1016/j.neunet.2021.06.006
0893-6080/

in prosody control, independent modeling of prosody attributes is still inadequate because of the covariant nature of attributes in the multi-dimensional prosodic space. As a result, in GST, a single learned token is insufficient in general for controlling individual prosody attributes. The major issue of the token-based approaches is that a mixture of prosody attributes needs to be explicitly factorized for controlling the prosody independently.

To alleviate the above difficulty, the first author introduces a hierarchical GST (H-GST) architecture (An, Wang, Yang, Ma, & Xie, 2019) to learn multi-level, disentangled representations to control synthesized speech prosody in coarse-to-fine granularities. Conceptually, H-GST model is an end-to-end training to decompose the reference embeddings into a set of basis vectors or "soft clusters" (style tokens) to facilitate appropriate prosody control. However, even in H-GST, there is still no straightforward way to sample utterances with varying prosody. The variational autoencoder (VAE) (Kingma & Welling, 2014) based disentangled representation framework (Hsu, et al., 2019, 2019; Zhang, Pan, He, & Ling, 2019) has been proposed to model each variable by discovering features and extracting the latent prosody attributes from observed mixed attributes (i.e., speaker and the corresponding prosody style). VAE, like GST or H-GST, is still an unsupervised and non-identifiable approach (Locatello, et al., 2019), which makes it difficult to disentangle and to interpret individual latent variables for more explicit control. Instead, usually a rather complex post-processing is needed to subjectively perceive the resultant prosodic effects in each attribute dimension.

Since the individual attributes are entangled together, manipulating one attribute may affect other dimensions. Additionally, the appropriate dimension of embedded latent space needs to be set in a trial-and-error manner which can be time consuming and database dependent. In such prosody control models, the variability of prosody attributes mostly depends on the individual attribute latent representations. As a result, such dispersion may undervalue the importance of prosody attribute separation, resulting in its inability to independently control individual prosody attributes. Therefore, the interpretable and independent control over the individual prosody attributes in neural TTS needs to be improved.

In this paper, we introduce two approaches for direct, independent control of prosody attributes in neural TTS. In addition to independent prosody control, it can also make each individual attribute more interpretable, both physically and perceptually. Firstly we analyze the VAE-based disentangled representation architecture, and show the importance of prosody attribute separation. We then introduce the real-valued, prosodic feature approach to enhance the interpretability and reliability of individual prosody attribute control. Thus, the dimensions controlling the individual attributes are determinate and aware, which does not require complex post-processing to manually evaluate their corresponding meaning. This also avoids dummy dimensions and improves the modeling capacity for individual attributes.

Besides, we find experimentally that there still exist representational entanglements in the individual attribute dimensions, which can prevent independent control along individual prosody attributes, despite that individual prosody contours are generated independently. We further propose to add a mutual information minimization estimation to enhance prosody attribute separation. The mutual information minimization can disentangle individual prosody attribute representations to facilitate independent, direct prosody control. In this way, the independent representations of individual prosody attributes are effectively generated to control individual attributes. Experimental results show that the mutual information minimization approach is more robust than the real-valued, prosodic feature method in independent control over individual prosody attributes.

The main contributions of this paper are summarized as follows:

- To facilitate independent control of prosody attributes in VAE-based, disentangled representations, we propose two approaches to separate and decorrelate the prosody attributes: (1) replacing the traditional Mel-spectrogram features with more explicit and perceivably relevant prosodic features, i.e., F_0, energy and duration, via separated VAE modules; and (2) proposing a mutual information minimization procedure to further disentangle the covariant prosody attributes. With the proposed approaches, time consuming and expensive effort in prosody annotations of the training data, discovery of the latent prosodic features and complex post-processing for identifying the relevant attributes in VAE-based, disentangled representations are no longer needed.
- The proposed attribute separation scheme provides an interpretable prosody control in disentangled dimensions where independent control of pitch, volume and speech rate can be performed effectively.
- Our approach outperforms the traditional VAE-based, baseline systems and the improvement is confirmed in both subjective and objective tests. The resultant attributes are interpretable and the performance is more robust than the counterpart in the prior art. The TTS prosody can be controlled better along separated prosody attributes.

The rest of the paper is organized as follows. In Section 2 we introduce the related prior works. In Section 3 we present the TTS model architecture. In Section 4 we review the basic theory of VAE for learning disentanglement of the attributes. In Section 5 we introduce the two proposed approaches to enhance the independence between different attributes for direct prosody control. In Section 6 we present our experiments and the corresponding results. We conclude the paper in Section 7.

2. Related works

As mentioned above, to prepare a large and well annotated speech dataset for training is both tedious and expensive. To use an unlabeled dataset, a reference encoder-based prosody representation (Skerry-Ryan, et al., 2018) has been proposed. It aims at learning an embedding space of prosody from speech data directly in the training process. The learned embedding space, when used as a condition in synthesis, can generate speech similar to the prosody of the reference signal, even when the reference and target speakers are different. However, the reference encoder-based TTS cannot control the prosody in a flexible manner. The augments of the reference encoder are then proposed to perform prosody control aside from prosody transfer, such as GST (Wang, et al., 2018) and H-GST (An et al., 2019).

In GST model (Wang, et al., 2018) a multi-head attention (Vaswani, et al., 2017) based style token layer is introduced to enable the reference encoder to extract a small but rich varieties of prosodic information in the training data. Given style tokens can then be used by the model to control TTS prosody. But single-token conditioning reveals that it is difficult to show how much and what prosody embedding can be learned from a given reference token. For example, one token may learn the speech rate only while other tokens may learn a mixture of prosody attributes that reflects the stylistic co-occurrence in an embedded form. We have also found that a low-pitched reference token may encode a slower speech rate in addition to its intrinsic pitch variation.

In an H-GST architecture (An et al., 2019), by augmenting the GST model with cascaded multiple style token layers with residuals, multiple-level, disentangled information can be learned. Prosody control can then be performed in different granularities by conditioning upon individual tokens in hierarchical style token

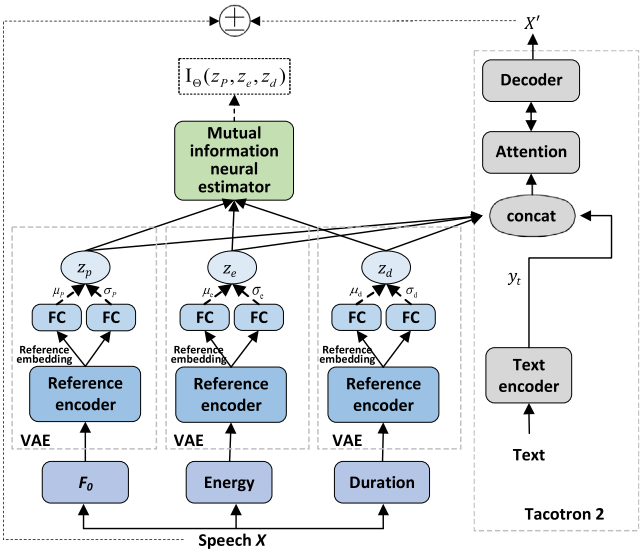

Fig. 1. System overview. The thick dashed lines denote sampling via reparameterization.

layers. By increasing the number of layers, a coarse-to-fine style decomposition can then be obtained. Despite that the model can achieve prosody control hierarchically, there is still no easy way to sample token utterances of varying prosody.

Attempts have also been made to combine probabilistic latent variable models trained with the stochastic gradient variational Bayes (SGVB) (Akuzawa, Iwasawa, & Matsuo, 2018; Kenter, Wan, Chan, Clark, & Vit, 2019; Rezende, Mohamed, & Wierstra, 2014). These models use a fully unsupervised and non-identifiable approach, which also makes it difficult to disentangle or to interpret their latent variables for control. In Hsu, et al. (2019), the authors attempt to overcome this problem by using a Gaussian mixture as the latent prior and to perform clustering in the latent space. And in Battenberg, et al. (2019), a hierarchical latent variable model is introduced to separate the modeling of style from prosody. However, all of these VAE-based disentangled representation methods are fully unsupervised and the results can be hard to interpret in the latent space or require complex post-processing. In addition, the latent variables only capture the salient attributes one would like to control, leading to poorly independent control over individual prosody attributes.

Recently, a semi-supervised generative model (Habib, et al., 2020), which has been applied to the end-to-end based TTS, is related to this paper. In Habib, et al. (2020), by providing partial supervision to some of the latent variables, it is possible to force them to be more consistent and interpretable. But this approach still relies on the labeled attributes of speech, such as affection and speech rate, and the labeled data-set is proprietary. Besides, its research objective is to evaluate/demonstrate the efficacy of semi-supervision rather than to control a particular prosody attribute. Motivated by the recent works, in this paper, we firstly analyze the control process of VAE-based disentangled representation system and demonstrate the importance of independent and direct control of prosody attributes. We then propose two interpretable prosody attribute separation approaches for independent prosody control in an interpretable and reliable manner. A series of experiments are designed to systematically analyze the effectiveness of these methods.

3. Model architecture

Fig. 1 illustrates the architecture of the proposed approach which incorporates the same Tacotron 2 network as in Hsu, et al. (2019, 2019) and Zhang et al. (2019), individual VAE modules and mutual information neural estimator are adopted for separating prosody attributes. As shown in the figure, it contains 3 VAE modules followed by a mutual information neural estimator, and Tacotron 2 (Shen, et al., 2018) which includes a text encoder, a decoder and a location-sensitive attention network (Chorowski, Bahdanau, Serdyuk, Cho, & Bengio, 2015). The output of text encoder is simply concatenated with the outputs of the 3 VAE modules and fed them into the attention network. The attention network then converts encoded sequence to a context vector for each decoder output step. For consistency, we set all VAE modules to be of the same structure. Each VAE module consists of a reference encoder followed by two separated FC layers. Individual prosody embeddings are separately generated by using their VAE modules. The mutual information neural estimator is constructed as a deep neural network (DNN) followed with two separated FC layers. Through the mutual information minimization estimation, we further make the three prosody attribute representations independent of each other. Section 5.2 will introduce the detail description of the mutual information neural estimator.

Following Fig. 3 in Section 5.1, durations of the phonemes are obtained using the HTK state level forced alignment (Young, et al., 2002) from the speech files. Then the total duration of the phoneme is obtained by summing all state durations of the corresponding phoneme. Frame number information is obtained by dividing the duration of the phoneme by the frame shift. The F_0 and energy are respectively extracted by using the WORLD vocoder (Morise, Yokomori, & Ozawa, 2016) from the aligned wav files, which are then averaged over the phoneme duration to make them piece-wise constant for learning a better distribution of each phoneme in model training. The duration of each phoneme is one value, which is represented by using the corresponding number of frames of each phoneme in our training. This makes sure that F_0, energy and duration of each phoneme are all aligned at the frame level (i.e., each frame corresponds to its F_0 and energy values). The generated frame-level prosodic features, F_0, energy and duration, are then fed into the 3 separate VAEs to learn their corresponding latent representations, respectively. With this arrangement, additional annotation of prosody, discovery of new features and clustering in the latent space are no longer necessary. The VAEs can generate separated, individual prosody attributes with enhanced representational capacity. The outputs of 3 VAEs are simply concatenated with the text encoder output and then consumed by a location-sensitive attention mechanism, which converts the encoded sequence to a context vector for each decoder step, as shown in Fig. 1. In addition, the outputs of 3 VAEs are first passed through an FC layer to make sure the dimension equal to the text encoder output before concatenation. The output of our system is the Mel-spectrogram of the target speech. We use Parallel WaveNet (Den Oord, et al., 2018) neural vocoder conditioned on the Mel-spectrogram to reconstruct the final speech waveform. The details of each individual component will be described in Section 6.2.

4. VAE for disentangled representation learning

4.1. Reference encoder

As discussed above, the VAE architecture contains a reference encoder and two separated FC layers. The reference encoder encodes the prosodic features of a variable-length audio signal into a fixed-length reference embedding. The architecture of reference encoder consists of a stack of six 2-D convolutional layers cascaded with one unidirectional 128-unit LSTM layer. In our experiments, different from Skerry-Ryan, et al. (2018), all LSTM

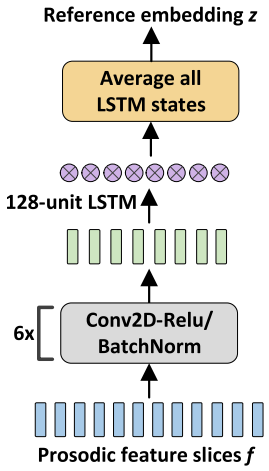

Fig. 2. An example of internal modules in reference encoder.

states are used as output, and then are averaged to serve as the reference embedding, which is better and more stable than that in Skerry-Ryan, et al. (2018), where only the last state is used as the reference embedding. The reference encoder is shown in Fig. 2 where the reference embedding output is:

$$z = \text{avg}(\text{LSTM}(\text{stack}(\text{Conv2D}(f)))) \tag{1}$$

where f is the input to the reference encoder, and z, the output of the reference encoder. The output z is then fed into two separated FC layers with 128 tanh hidden units to generate the mean and standard deviation of the latent variables.

4.2. VAE-based disentangled representation

In training, the generative models in VAE (Hsu, et al., 2019) are trained to learn the effects of different attributes by factorizing the latent prosodic variables. To achieve this, two latent variables $\mathbf{y}_l$ and $\mathbf{z}_l$ are introduced in addition to the two observed variables, speech $\mathbf{X}$ and text $\mathbf{y}_t$. $\mathbf{y}_l$ is a K-way categorical discrete variable (e.g., speaker identity), named latent attribute class, and $\mathbf{z}_l$ is a D-dimensional continuous variable of the latent attribute representation. Specifically, it is assumed that the prior, $p(\mathbf{y}_l) = K^{-1}$, to be a non-informative prior to encourage every component to be included, and its conditional distribution is a diagonal-covariance Gaussian with learnable means and variances $p(\mathbf{z}_l|\mathbf{y}_l) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{y}_l}, diag(\boldsymbol{\sigma}_{\mathbf{y}_l}))$. As a result, the marginal prior of $\mathbf{z}_l$ becomes a Gaussian mixture model (GMM) with diagonal covariances and equal mixture weights.

With this formulation, speech is now generated by conditioning upon the text $\mathbf{y}_t$, and a sample $\mathbf{z}_l$ drawn from $p(\mathbf{z}_l|\mathbf{y}_l)$, which is $p(\mathbf{X}|\mathbf{y}_t, \mathbf{z}_l)$. Following the VAE framework (Kingma & Welling, 2014), a variational distribution $q(\mathbf{y}_l|\mathbf{X})q(\mathbf{z}_l|\mathbf{X})$ is used to approximate the intractable true posterior $p(\mathbf{y}_l, \mathbf{z}_l|\mathbf{X}, \mathbf{y}_t)$, which assumes that the posterior of unseen attributes is independent of the observed attributes. The mean and standard deviation of $q(\mathbf{z}_l|\mathbf{X})$ are parameterized by the above reference encoder network. $q(\mathbf{y}_l|\mathbf{X})$ is configured to be an approximation of $p(\mathbf{y}_l|\mathbf{X})$ that reuses $q(\mathbf{z}_l|\mathbf{X})$. The model is trained by maximizing its evidence lower bound (ELBO) as follows:

$$\begin{aligned}\mathcal{L}(p, q; \mathbf{X}, \mathbf{y}_t) &= \mathbb{E}_{q(\mathbf{z}_l|\mathbf{X})}[\log p(\mathbf{X}|\mathbf{y}_t, \mathbf{z}_l)] \\ &- \mathbb{E}_{q(\mathbf{y}_l|\mathbf{X})}[D_{KL}(q(\mathbf{z}_l|\mathbf{X}) \parallel p(\mathbf{z}_l|\mathbf{y}_l))] - D_{KL}(q(\mathbf{y}_l|\mathbf{X}) \parallel p(\mathbf{y}_l))\end{aligned} \tag{2}$$

where $q(\mathbf{z}_l|\mathbf{X})$ is estimated via Monte Carlo sampling approach (Kingma & Welling, 2014), and all components are differentiable by reparameterization. D_{KL} is the Kullback–Leibler (KL) divergence (Bowman, et al., 2016).

Therefore, latent attributes are modeled as a mixture distribution, which facilitates automatic discovery of the latent attribute clusters. The mixture parameters can be analyzed to find which latent component corresponds to what prosody attributes as in GST (Wang, et al., 2018).

4.3. Inexplicability of latent variables

Note that the approach is unsupervised and it is hard to disentangle or interpret the latent variables for more precise prosody control without extensive subjective listening to the synthesized speech. For the sake of direct, interpretable prosody attributes to be controlled in the TTS effectively, a semi-supervised generative model (Habib, et al., 2020) can be used for injecting partial supervision to certain latent variables. However, the method still relies upon how to process many intractable speech attribute labels.

The intercorrelated nature of prosody attributes, e.g., pitch, volume and speech rate, can make quantitative labeling or disentanglement of the components difficult.

5. Interpretable attribute separation for prosody control

Although the VAE-based disentangled representation model (Hsu, et al., 2019) has shown its capability to disentangle speech attributes, it is an unsupervised scheme and the latent attributes are hard to interpret. As a result, only salient features can be identified for prosody control because of the covariant nature of the individual attributes. In speech prosody control, interpretable speech attribute separation is critical for a more precise prosody control. Our experiments in Section 6.3 will examine this phenomenon in detail. In this paper, we aim at finding perceptually interpretable latent embeddings to facilitate independent and direct control of individual prosody attributes. If, like the VAE-based representation model (Hsu, et al., 2019), we use a single VAE module to encode all three, F_0, energy and duration, attributes in a 3-D latent vector space via the reparameterization trick (Kingma & Welling, 2014), the 3-D latent variables will be intrinsically entangled with one another. As a result, it will be difficult to prevent any cross-attribute control effects from happening. In addition, the perceptual association between the latent variables with the original prosody attributes will be hard to define. In other words, very likely one latent variable will be associated with more than one original prosody attribute. Then, the approach will become non-identifiable and no explicit interpretation of the latent embeddings can be obtained to guide the direct control of the original 3 prosody attributes. Hence, here we construct 3 separate VAEs to model the corresponding 3 prosodic features explicitly. Then, we minimize the mutual information between any paired VAEs to disentangle the cross-attribute correlations among any paired VAEs. The prosody control mechanism along each attribute is then explicit and direct. No more complex perceptual, post-processing is needed to identify the corresponding perceptual effects of the latent variables when the mutual information is minimized between any paired VAEs. An independent and direct prosody control along each targeted attribute is then feasible. In this section, we will introduce these two methods for further separation of interpretable prosody attributes.

5.1. Real-valued, prosodic feature based representation

In a VAE-based disentangled representation model, the latent variables cannot be easily separated, particularly when the perceived speech attributes may be highly intercorrelated in the

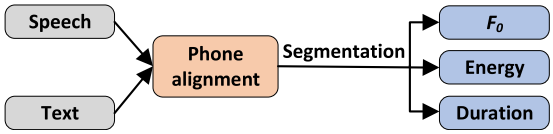

Fig. 3. A process of generating separated prosodic features.

attribute dimensions. However, the attribute disentangled representations usually play a critical role in prosody control. Inspired by Sun, et al. (2020), we propose to use the separated prosodic features, F_0, energy and duration, to keep the individual attributes and their easy interpretability. Fig. 3 shows an example of generating separated prosodic features, whose details will be described in Section 6.1. In Sun, et al. (2020), prosodic features are observed to be extracted energy-duration-F_0 order guided by scheduled training across latent dimensions. But we have found it needs to seek the representations of individual attributes by following the order. This ignores the correlations between prosody attributes, and leaks partial attribute information to other attribute dimensions, resulting in low interpretability and poor independent control. So we use real-valued, prosodic features, which are the easiest conditional signal (relative cheap) and can be better extracted from speech, to enhance the individual attribute representation and interpretability.

To control prosody attributes in addition to the text $\mathbf{y}_t$, three attribute variables, $\mathbf{z}_p$, $\mathbf{z}_e$ and $\mathbf{z}_d$, are introduced to condition the generative process, where $\mathbf{z}_p$ models pitch attribute, $\mathbf{z}_e$ models volume attribute, and $\mathbf{z}_d$ models speech rate attribute. Prior distributions for these variables are defined to be standard normal, $\mathcal{N}(\mathbf{0}, \mathbf{I})$. We modify the conditional generative model with three attribute variables as $p(\mathbf{X}|\mathbf{y}_t, \mathbf{z}_p, \mathbf{z}_e, \mathbf{z}_d)$, where $\mathbf{X}$ denotes speech.

Three variational distributions, $q(\mathbf{z}_p|\mathbf{X})$, $q(\mathbf{z}_e|\mathbf{X})$ and $q(\mathbf{z}_d|\mathbf{X})$, are introduced to approximate the intractable posteriors of the attribute variables, following the VAE framework (Kingma & Welling, 2014). Each distribution is defined to be a Gaussian with diagonal-covariance, whose mean and variance are parameterized by the above recurrent reference encoder. We modify Eq. (2) to model individual prosody attributes:

$$\begin{aligned}\mathcal{L}(p, q; \mathbf{X}, \mathbf{y}_t) &= \mathbb{E}_{q(\mathbf{z}_p|\mathbf{X})q(\mathbf{z}_e|\mathbf{X})q(\mathbf{z}_d|\mathbf{X})}[\log p(\mathbf{X}|\mathbf{y}_t, \mathbf{z}_p, \mathbf{z}_e, \mathbf{z}_d)] \\ &- D_{KL}(q(\mathbf{z}_p|\mathbf{X}) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I})) - D_{KL}(q(\mathbf{z}_e|\mathbf{X}) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I})) \\ &- D_{KL}(q(\mathbf{z}_d|\mathbf{X}) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I}))\end{aligned} \tag{3}$$

where the negative expectation term is referred to as the reconstruction loss.

The real-valued, prosodic feature based attribute representation does not need complex listening to identify the perceptual meaning of each dimension, also does not need to choose a suitable dimensionality for each variable. The dimensions for explicit control of the individual prosody attributes are deterministic and interpretable. It is also better to avoid the dummy dimensions for improving the modeling capacity of the individual attributes.

5.2. Mutual information minimization for attribute separation

Although the above real-valued, prosodic features can enhance the representational and interpretable contributions of individual prosody attributes, there are still two shortcomings. Firstly, it separately generates individual prosody attribute representations to enhance interpretability of individual prosody attributes. There exists an infinite number of bijective mappings (Locatello, et al., 2019) from the learned attribute representation space to another space with the same marginal distribution, but the two representation spaces are fully entangled, so the disentangled

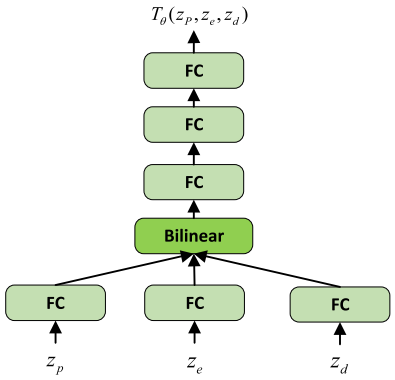

Fig. 4. An example of internal modules in DNN network.

prosody features are not enough to make their individual attribute representations independent. Secondly, our experiments in Section 6.4.1 also show that there are representational entanglements in the individual attribute dimensions. They hinder the resultant model from performing independent control over individual prosody attributes. To achieve independent representations of individual prosody attributes, we further utilize mutual information minimization estimation as additive bias in individual attribute variables, which makes individual attribute variables solely focusing on their own attributes being captured.

Compared to the real-valued, prosodic feature based approach, we can make the individual prosody attributes independent by minimizing the mutual information in their attribute representations, $\mathbf{z}_p$, $\mathbf{z}_e$ and $\mathbf{z}_d$, so that individual attributes do not contain information from other prosody attribute dimensions. However, it is not obvious how to compute and minimize the mutual information between individual prosody attribute vectors. At first, we briefly describe a recently proposed method (Belghazi, Baratin, Rajeswar, Ozair, & Bengio, 2018) to estimate the mutual information, then we present our novel application to minimize it jointly with the reconstruction loss. Following Belghazi et al. (2018), the mutual information is a Shannon entropy-based measure of dependence between random variables. The mutual information, $I(\mathbf{z}_p, \mathbf{z}_e, \mathbf{z}_d)$, of prosody attribute variables, $\mathbf{z}_p$, $\mathbf{z}_e$ and $\mathbf{z}_d$, is equivalent to the KL divergence (Kullback & Leibler, 1951) between their joint distribution, $P_{\mathbf{z}_p,\mathbf{z}_e,\mathbf{z}_d}$, and the product of marginals, $P_{\mathbf{z}_p} * P_{\mathbf{z}_e} * P_{\mathbf{z}_d}$, which can be written as: $I(\mathbf{z}_p, \mathbf{z}_e, \mathbf{z}_d) = D_{KL}(P_{\mathbf{z}_p,\mathbf{z}_e,\mathbf{z}_d} \parallel P_{\mathbf{z}_p} * P_{\mathbf{z}_e} * P_{\mathbf{z}_d})$. The intuitive meaning is clear: the larger the divergence between the joint and the product of the marginals, the stronger the dependence between individual prosody attribute variables. This divergence, hence the mutual information, vanishes for fully independent variables.

Using this fact, the mutual information neural estimator method constructs a lower bound of mutual information based on Donsker–Varadhan representation of KL divergence (Donsker & Varadhan, 1983):

$$\begin{aligned}&I(\mathbf{z}_p, \mathbf{z}_e, \mathbf{z}_d) \geq I_\Theta(\mathbf{z}_p, \mathbf{z}_e, \mathbf{z}_d) \\ &= \sup_{\theta \in \Theta} \mathbb{E}_{P_{\mathbf{z}_p,\mathbf{z}_e,\mathbf{z}_d}}[T_\theta] - \log(\mathbb{E}_{P_{\mathbf{z}_p} * P_{\mathbf{z}_e} * P_{\mathbf{z}_d}}[e^{T_\theta}])\end{aligned} \tag{4}$$

where $I_\Theta(\mathbf{z}_p, \mathbf{z}_e, \mathbf{z}_d)$ is the lower bound of mutual information, T can be any function that makes the two expectations in the above equation finite. Here, we follow Belghazi et al. (2018) to use a DNN network for T with parameter $\theta \in \Theta$, which allows us to estimate the mutual information between three prosody attribute variables by maximizing this lower bound with respect to T through gradient descent. Fig. 4 shows the internal modules in the DNN network, which consists of an FC layer, a bilinear layer and three FC layers. To avoid the possible problem of gradient

vanishing, we do not use the tanh activation function in DNN network. Instead, we apply ReLU activation function to all layers in DNN network. First, we randomly shuffle the order of sentences in a batch, where the z_p, z_e and z_d from a sentence are all paired with each other, to make the shuffled ones as input to the DNN network, respectively, to avoid possible over- or underfitting in DNN network training. In other words, z_p from sentence i is paired with z_e and z_d from sentence i instead of z_e and z_d from sentence j. In the experiments, the batch size is set to 32. To concatenate the z_p, z_e and z_d with the text encoder output, as shown in Section 3, the z_p, z_e and z_d are first passed through an FC layer to make sure their dimension equal to the text encoder output. As described above, actually the mutual information between random variables is equivalent to the KL divergence between their joint distribution and the product of marginal distributions. Since it is intractable to calculate the mutual information between the three prosody attribute variables, we compute the mutual information between any two of the three prosody attribute variables via the DNN network. And the KL divergence is computed by using their corresponding means and standard deviations in a closed form. Hence, after the DNN network, we adopt two separate FC layers, each with 128 tanh hidden units, to generate their corresponding means and standard deviations, respectively. Note that all layers in the DNN network are multiplexed and shared by the z_p, z_e and z_d. In this way, we can estimate the mutual information between every pair of the three prosody attribute variables.

We minimize the reconstruction loss and the estimated mutual information along the prosody attribute variables. Since the mutual information is always non-negative, we clip the estimated mutual information to zero if it is negative. The clipped value is not only a better estimate of the mutual information than the non-clipped one (because the true mutual information is always non-negative), it also avoids minimizing a function that is unbounded from below. Thus, the overall objective function is a min–max problem where we maximize the lower-bound of mutual information, I_Θ, and minimize the mutual information and the reconstruction loss: $\min_{\mathbf{z}_p,\mathbf{z}_e,\mathbf{z}_d,\mathbf{y}_t} \max_\Theta \{\|D(\mathbf{z}_p, \mathbf{z}_e, \mathbf{z}_d, \mathbf{y}_t) - X\|_1 + \lambda * max(0, I_\Theta(\mathbf{z}_p, \mathbf{z}_e, \mathbf{z}_d))\}$, where λ is a preset hyper-parameter for balancing the two loss functions. In our experiments, we find the optimization results are insensitive to the choice of λ, hence we set $\lambda = 0.1$. Similar to common GAN training (Goodfellow, et al., 2014), we update the speech synthesis model and the mutual information estimator function, T_θ, alternatively in each step of the training process. By optimizing them, we can jointly ensure the quality of speech reconstruction, and make individual attribute representations independent of each other. Details on the implementation of the proposed training method are provided in Algorithm 1.

Algorithm 1 Pseudocode for the proposed training method

Input: Pairs of prosodic features and text $(x_i^p, x_i^d, x_i^e, c_i)$;
x_i^p, x_i^d, x_i^e: prosodic features (F_0, duration, energy) obtained from speech X_i;
c_i: phoneme sequences.
Output: E_C, D, E_P, E_D, E_E;
E_C: text encoder; D: decoder;
E_P: pitch VAE module; E_D:duration VAE module; E_E: energy VAE module.
$E_C, D, E_P, E_D, E_E \leftarrow \arg\min_{E_C,D,E_P,E_D,E_E} \sum_i \| D(E_C(ci), E_P(x_i^p), E_D(x_i^d), E_E(x_i^e)) - X_i \|_1$
$D, E_P, E_D, E_E, T \leftarrow$ initialization with random weights
while D, E_P, E_D, E_E, T not converged **do**
  Sample and randomly shuffle a mini-batch of $(x_i^p, x_i^d, x_i^e, c_i)$; $i = 1, 2..., b$
  $\{\mathbf{y}_i\} \leftarrow \{E_C(c_i)|i = 1, 2..., b\}$
  $\{\mathbf{z}_i^p\} \leftarrow \{E_P(x_i^p)|i = 1, 2..., b\}$
  $\{\mathbf{z}_i^d\} \leftarrow \{E_D(x_i^d)|i = 1, 2..., b\}$
  $\{\mathbf{z}_i^e\} \leftarrow \{E_E(x_i^e)|i = 1, 2..., b\}$
  $L_{MI} = \frac{1}{b}\sum_{i=1}^b T(\mathbf{z}_i^p, \mathbf{z}_i^d, \mathbf{z}_i^e) - log(\frac{1}{b}\sum_{i=1}^b e^{T(\mathbf{z}_i^p, \mathbf{z}_i^d, \mathbf{z}_i^e)})$
  $L = \frac{1}{b}\sum_{i=1}^b \| D(\mathbf{y}_i, \mathbf{z}_i^p, \mathbf{z}_i^d, \mathbf{z}_i^e) - X_i \|_1 + \lambda * max(0, L_{MI})$
  $D = D - \epsilon \nabla_D L$
  $E_P = E_P - \epsilon \nabla_{E_P} L$
  $E_D = E_D - \epsilon \nabla_{E_D} L$
  $E_E = E_E - \epsilon \nabla_{E_E} L$
  $T = T + \epsilon \nabla_T L_{MI}$
end while

6. Experiments and results

6.1. Basic experimental setup

All experiments are conducted on a public, English corpus used in the Blizzard Challenge 2013.[1] It is a single-channel database of ∼19 h of speech, recorded by a single female English speaker with rich prosody. We remove long silence (> 0.1 sec) at the beginning and ending of each utterance. Log magnitude spectrogram is extracted as target speech representations with a Hanning window of 50 ms and 12.5 ms frame shift. Phoneme sequences are used as the text input. Following Fig. 3 in Section 5.1, we first align the phoneme sequence with the corresponding speech features. The duration of each phoneme is calculated along the alignment path. The F_0 and energy are respectively extracted by using the WORLD vocoder (Morise et al., 2016) from the aligned wav files, which are then averaged in all frames of a phoneme to make them piece wise continuous. We extend frame of the duration to frame rate of the F_0 and energy according to the frame number. The generated prosodic features, F_0, energy and duration, are then used as input to the corresponding VAE modules. For all different systems in our experiments, we train ∼350k steps with a single Nvidia Tesla P40 GPU, and the batch size is 32. The models on which we conduct experiments include:

- VAE-DR (baseline): We introduce a VAE into Tacotron 2 network (Shen, et al., 2018) to model latent attributes in a disentangled fashion. The VAE architecture is presented in Section 4.1, and the description of VAE for disentangled representation learning can be found in Section 4.2. We denote the baseline model as VAE-DR;
- VAE-PF: We propose to adopt three separated VAE modules for explicit, individually extracted prosodic features, i.e., F_0, energy and duration, to enhance the interpretability and reliability of individual prosody attribute control, which is denoted as VAE-PF. The detailed description of VAE-PF is in Section 5.1;
- VAE-MI: We further utilize mutual information minimization estimation as additive bias in individual attribute variables to reduce the entanglements of individual prosody attribute representations in VAE-PF approach, which is denoted as VAE-MI. In Section 5.2 we describe the VAE-MI method in detail;
- GST*: We replace the reference encoder and Tactron (Wang, et al., 2017) of original GST framework (Wang, et al., 2018) with the reference encoder discussed in Section 4.1 and Tactron 2 (Shen, et al., 2018) to build the GST-based model and make a fair comparison, which is denoted as GST*. The details on the GST* are presented in Section 6.5.

[1] The dataset is available at http://www.cstr.ed.ac.uk/projects/blizzard/2013/lessac_blizzard2013/.

Table 1
Quantitative evaluation of pitch and speech rate control of VAE-DR system.

Target dimension	Metric	$\mu_{\mathbf{z}_{l,d}} - 3\sigma_{\mathbf{z}_{l,d}}$	$\mu_{\mathbf{z}_{l,d}}$	$\mu_{\mathbf{z}_{l,d}} + 3\sigma_{\mathbf{z}_{l,d}}$
Pitch ($d = 8$)	F_0 (Hz)	176.5	185.1	202.6
	Energy (dB)	−1.06	−1.06	−1.05
	Duration (s)	3.08	2.81	2.53
Speech rate ($d = 3$)	F_0 (Hz)	198.6	189.3	179.4
	Energy (dB)	−1.06	−1.07	−1.08
	Duration (s)	2.36	2.83	3.22

To reconstruct speech with the predicted Mel-spectrogram, we train a parallel WaveNet vocoder conditioned on the ground-truth Mel-spectrogram (Den Oord, et al., 2018; Shen, et al., 2018). A parallel WaveNet trained with ground truth-aligned predicted features can usually improve the model performance (Shen, et al., 2018), we use the same parallel WaveNet (Den Oord, et al., 2018) for fair comparisons across all different systems.

We conduct Mean Opinion Score (MOS) and preference listening tests (A/B) to evaluate the performance of different experimental systems subjectively. 20 listeners take part in the subjective evaluations, where 30 randomly selected utterances in each attribute dimension are provided in each testing session.[2] The F_0 Frame Error (FFE) (Chu & Alwan, 2009) is used to quantify objectively the reconstruction distortion of different systems, where 30 randomly selected utterances in each attribute are provided in each testing session. FFE combines the voicing decision error and F_0 error metrics to capture how well F_0 information, which forms a major component of the prosody, is retained.

6.2. Model details of common components

As shown in Fig. 1 in Section 3, the framework of our interpretable prosody attribute separation model consists of 3 subsystems: three VAE modules; a mutual information neural estimator; and a Tacotron 2 network (Shen, et al., 2018). For the reference encoder sub-module in each VAE module, each convolutional layer is composed of 3 × 3 filters with a 2 × 2 stride, ReLU activation. Batch normalization (Ioffe & Szegedy, 2015) is applied to every layer. The number of filters in each layer doubles at half of the down-sampling rate: 32, 32, 64, 64, 128, 128. The output of convolution stack is fed into a unidirectional LSTM with 128 units, and all LSTM states are averaged to serve as the reference embedding. The mutual information neural estimator consists of a DNN network as shown in Fig. 4 in Section 5.2, followed by two separate FC layers with 128 tanh hidden units.

In the Tacotron 2 part, text encoder contains three 1-D convolutional layers with 512 filters and 5 × 1 shape, followed by a bidirectional LSTM layer of 256 units using zoneout (Krueger, et al., 2017) with probability 0.1. The resulting text encodings are simply concatenated by individual attribute representations and then are consumed by a location-sensitive attention mechanism (Chorowski et al., 2015) which converts encoded sequence to a fixed-length context vector for each decoder output step. The pre-net is comprised of two FC layers of 256 hidden ReLU units and the hidden size of a stack of two uni-directional LSTM layers is 1024. There is a five-layer convolutional post-net, where each post-net layer consists of 512 filters with shape 5 × 1 with batch normalization (Ioffe & Szegedy, 2015), followed by tanh activations on all but the final layer.

We built different systems to analyze the performance of the VAE-based disentangled representation model (Hsu, et al., 2019) and the proposed interpretable prosody attribute separation methods. The difference of each system will be described in the following experiments. For the parallel WaveNet vocoder (Den Oord, et al., 2018), there are totally 30 dilated layers, where each of 10 layers shares the dilatation pattern $2^0, 2^1, \ldots, 2^9$. Both the causal convolution and the 1 × 1 residual convolution have 256 channels, while the channel number of the convolutions in skip-connection is 2048.

6.3. Analysis of VAE-based disentangled representation

The VAE-based disentangled representation framework (Hsu, et al., 2019) learns a representation, which disentangles speech attributes, to make them to be controllable. Specially, latent dimensions of the framework are assumed to be conditionally independent. But individual dimensions only capture salient features and cannot provide an interpretable and completely independent control over individual attributes. We firstly build such a VAE-based disentangled representation model (i.e., VAE-DR) to analyze the details of latent dimensions. For the VAE architecture, it contains a recurrent reference encoder, followed by two FC layers to generate the mean and standard deviation of latent variables, $\mathbf{z}_l$.

We use 16-dimensional $\mathbf{z}_l$ to capture the speech attributes one would like to control, such as style and prosody attributes. We can judge whether the individual speech attributes can be controlled through adjusting single target dimension d of the $\mathbf{z}_l$ while fixing others. We randomly draw 10 samples of seed $\mathbf{z}_l$ from the prior, deterministically set the target dimension d to $\mu_{\mathbf{z}_{l,d}} - 3\sigma_{\mathbf{z}_{l,d}}$, $\mu_{\mathbf{z}_{l,d}}$ and $\mu_{\mathbf{z}_{l,d}} + 3\sigma_{\mathbf{z}_{l,d}}$ to construct modified $\mathbf{z}_l^*$, where $\mu_{\mathbf{z}_{l,d}}$ and $\sigma_{\mathbf{z}_{l,d}}$ are the mean and standard deviation of the marginal distribution of the target dimension d. We then synthesize a set of the same 20 text sequences for each of the 30 resulting values of $\mathbf{z}_l^*$. For each value of the target dimension, we compute the averaged metric over 200 synthesized utterances (10 seed $\mathbf{z}_l$ × 20 text inputs).

Since the VAE-DR model is fully unsupervised, it is not obvious to interpret the corresponding latent components without performing complex listening tests to find their perceptual attributes. In practice, we find that majority of the dimensions are interpretable after the extensive subjective tests. For example, the 3rd dimension corresponds to the speech rate attribute and the 8th dimension correlates well with the pitch attribute, but no dimension seems to address the volume attribute. In other words, the VAE-DR system cannot control the volume attribute. We also find that there are some spurious don't-care dimensions which do not affect the model output. For example, the VAE-DR model has two don't-care dimensions of $\mathbf{z}_l$, $d = 1$ and $d = 11$, which do not seem to affect the output. As a result, certain dimensions are not interpretable in their prosody control capability in the VAE-DR system and experimental trials are necessary to find an appropriate dimensionality of $\mathbf{z}_l$.

To objectively evaluate the prosody control performance of the VAE-DR model, we use 3 metrics: the averaged F_0 in voiced frames, the averaged energy (in log scale) and the averaged speech duration. They can be used to measure the difference of pitch, volume and speech rate between two samples. Table 1 shows quantitative evaluation of pitch and speech rate control of VAE-DR system. In the pitch dimension, the measured F_0 varies from low to high, but a corresponding decrease of duration can also be observed. Similarly, in the speech rate dimension, the measured speech duration is changed along with the change of speed rate, but a change of F_0 is also observed. These observations imply that by manipulating individual dimensions of the $\mathbf{z}_l$, $d = 3$ and $d = 8$, the target attribute can be controlled, but other attributes can also be affected unintentionally. This cross-attribute effect in prosody control makes independent control of individual prosody attribute difficult. It is observed that VAE-DR model cannot control the volume attribute, the measured energy

[2] Samples can be found at https://xiaochunan.github.io/prosody/index.html.

Table 2
FFE and MOS evaluations of VAE-DR systems with different dimensionality of $\mathbf{z}_l$, and MOS evaluation with 95% confidence interval.

$\mathbf{z}_l$ dimensionality	FFE	MOS
8	63.6%	3.55 ± 0.17
16	55.3%	3.67 ± 0.14
32	**51.6%**	$\mathbf{3.83 \pm 0.11}$

Table 3
Quantitative evaluation of individual prosody attribute control of VAE-PF system, where $\mathbf{z}_*$ denotes $\mathbf{z}_p$, $\mathbf{z}_e$ and $\mathbf{z}_d$, respectively.

Prosody attribute	Metric	$\boldsymbol{\mu}_{\mathbf{z}_*} - 3\boldsymbol{\sigma}_{\mathbf{z}_*}$	$\boldsymbol{\mu}_{\mathbf{z}_*}$	$\boldsymbol{\mu}_{\mathbf{z}_*} + 3\boldsymbol{\sigma}_{\mathbf{z}_*}$
Pitch ($\mathbf{z}_p$)	F_0 (Hz)	171.8	190.4	222.7
	Energy (dB)	−1.76	−1.03	0.62
	Duration (s)	3.01	2.89	2.75
Volume ($\mathbf{z}_e$)	F_0 (Hz)	190.6	198.3	207.1
	Energy (dB)	−2.79	−1.64	0.87
	Duration (s)	3.13	2.98	2.90
Speech rate ($\mathbf{z}_d$)	F_0 (Hz)	202.5	193.0	185.5
	Energy (dB)	−0.52	−1.17	−1.84
	Duration (s)	2.25	2.86	3.50

is essentially unchanged when the other two attributes, pitch and speech rate, are under variable control. Demonstrations of pitch and speech rate attribute control are given in the demo page.

To choose an appropriate dimension in VAE-DR model, we vary the dimensionality of $\mathbf{z}_l$ to measure the controlling effect. In addition to the 16-dimensional vector, both 8-dimensional and 32-dimensional $\mathbf{z}_l$ are also used for comparison. Similarly, we also examine the control over individual attributes by varying a target dimension d of the $\mathbf{z}_l$ while fixing others. The selected samples and the method of setting target dimension are the same as those in 16-dimensional $\mathbf{z}_l$. Compared to 16-dimensional $\mathbf{z}_l$, we find that increasing the dimensionality of $\mathbf{z}_l$ from 16 to 32 increases the difficulty of interpreting each dimension, and the VAE-DR system has more "don't care" dimensions in $\mathbf{z}_l$. Similar to the 16-dimensional $\mathbf{z}_l$, we also cannot find the volume dimension in the 32-dimensional setup, but the pitch and speech rate dimensions can be found. On the other hand, reducing the dimensionality to 8 can result in insufficient modeling capacity for latent attributes, and all individual attributes are entangled together, especially the prosody related attributes. The attribute controlled by a certain dimension becomes less unique and we observed a dimension can be both the pitch and the speech rate related. Therefore, the MOS and FFE evaluations are conducted for the VAE-DR systems with 8-dimension, 16-dimension and 32-dimension $\mathbf{z}_l$, as shown in Table 2, to quantify their reconstruction performance. Here, $\boldsymbol{\sigma}_l$ is the corresponding standard deviation of the 8-dimension, 16-dimension and 32-dimension $\mathbf{z}_l$. A lower value is better for the FFE distortion while a higher score is better for MOS subjective score. We can see that the results of the 32-dimension $\mathbf{z}_l$ is the best, which indicates that increasing the dimension of $\mathbf{z}_l$ improves reconstruction quality. However, this would reduce interpretability and generalization of the latent attribute control. Reducing the dimension of $\mathbf{z}_l$ results in poor modeling of latent attributes and weakening precise control of the corresponding speech attributes. Overall, for the VAE-DR model, we find the 16-dimension $\mathbf{z}_l$ is the best compromise to capture the salient attributes for effective prosody control.

6.4. On the effects of the interpretable attribute separation

Considering the inadequacy of the VAE-based disentangled representation model (Hsu, et al., 2019), we propose to apply two interpretable prosody attribute separation methods to make the representations for individual prosody attribute independent without using any prosodic labels. Different from our previous architecture (An et al., 2019), the introduced approaches improve independent controllability of individual prosody attributes while providing the interpretability and reliability of individual attribute control. In this section, we conduct experiments to analyze the performance of the proposed interpretable prosody attribute separation methods.

6.4.1. Experiments on real-valued, prosodic feature attribute representation

To achieve the interpretability of individual prosody attribute control, we propose to use the real-valued, prosodic features to learn individual prosody attribute representations (i.e., VAE-PF). We use 3 separated VAEs for the 3 corresponding prosodic features, F_0, energy and duration, and they are separately processed by the 3 VAEs to study the effects of individual prosody attribute representations, $\mathbf{z}_p$, $\mathbf{z}_e$ and $\mathbf{z}_d$. Note that the z_p, z_e and z_d are all one-dimensional, and they are all assigned for controlling their target attributes. This arrangement improves the VAE modeling capacity for the assigned prosody attributes. If we set z_p, z_e and z_d to have a dimension greater than 1, we will not be able to associate perceptual relevance to the coordinate components in the vector z_p, z_e and z_d, and it will result in a low interpretability problem and difficulty for direct control. Complex and tedious subjective listening tests need to be performed to identify the perceptual meaning of each coordinate in z_p, z_e and z_d, and they are particularly difficult in a fine-grained scale.

We investigate the controlling performance of individual prosody attributes in the VAE-PF model and the results are shown in Table 3. The selected samples of seed $\mathbf{z}_p$, $\mathbf{z}_e$ and $\mathbf{z}_d$ are from the prior, the method of setting the target dimension and inference text sequences are the same as those in VAE-DR system. Table 3 clearly indicates that by adopting separated prosodic features, we can significantly improve the interpretability of individual attribute control. Unlike the VAE-DR system, all the VAE-PF systems can interpret the corresponding prosody attributes. We also calculate 3 metrics, the averaged F_0 in voiced frames, the averaged energy (in log scale) and the averaged speech duration, to evaluate the ability of the VAE-PF model for controlling prosody attributes individually. When adjusting the pitch attribute, $\mathbf{z}_p$, following the $\boldsymbol{\mu}_{\mathbf{z}_p} - 3\boldsymbol{\sigma}_{\mathbf{z}_p}$, $\boldsymbol{\mu}_{\mathbf{z}_p}$ and $\boldsymbol{\mu}_{\mathbf{z}_p} + 3\boldsymbol{\sigma}_{\mathbf{z}_p}$, we find that the measured F_0 changes from low to high, while a minor cross-attribute effect is observed in both the volume and speech rate attributes, where the measured energy (in log scale) increases and the measured duration decreases with an increasing pitch. Similarly, when the volume control varies from low to high, measured energy varies from low to high, but both pitch and duration change, where F_0 increases and duration decreases slightly. Similar cross-attribute results are observed in speech rate control. These results are an indication that by manipulating individual embedding components, $\mathbf{z}_p$, $\mathbf{z}_e$ and $\mathbf{z}_d$, we can control the corresponding attribute, but the proposed VAE-PF system provides better interpretability and reliability of individual prosody control than the VAE-DR baseline. Demonstrations of control over individual prosody attributes can be found at our demo page.

We then conduct FFE and MOS evaluations of individual prosody attribute control to quantify the reconstruction performance of the VAE-PF system. Here, $\boldsymbol{\sigma}$ is the standard deviation of individual prosody attributes, $\mathbf{z}_p$, $\mathbf{z}_d$ and $\mathbf{z}_e$. The results are shown in Table 4 where the VAE-PF model achieves better FFE and MOS in each attribute than the VAE-DR with 32-dimensional $\mathbf{z}_l$. These results indicate that using the separated prosodic features can enhance the individual prosody attribute representations and improve the reconstruction quality.

As discussed above, except for volume attribute which is not interpretable in VAE-DR system, the system VAE-DR and VAE-PF

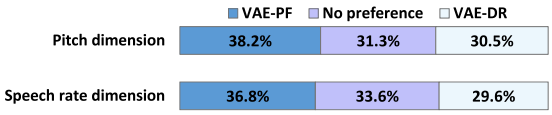

Fig. 5. Preference test results between VAE-DR and VAE-PF.

Table 4
FFE and MOS evaluations of VAE-PF systems with different prosody attributes, and MOS evaluation with 95% confidence interval.

Prosody attribute	FFE	MOS
Pitch ($\mathbf{z}_p$)	40.8%	3.92 ± 0.08
Volume ($\mathbf{z}_e$)	42.5%	3.91 ± 0.07
Speech rate ($\mathbf{z}_d$)	44.9%	3.87 ± 0.12

Table 5
Quantitative evaluation of individual prosody attribute control of VAE-MI system, where $\mathbf{z}_*$ denotes $\mathbf{z}_p$, $\mathbf{z}_e$ and $\mathbf{z}_d$, respectively.

Prosody attribute	Metric	$\boldsymbol{\mu}_{\mathbf{z}_*} - 3\boldsymbol{\sigma}_{\mathbf{z}_*}$	$\boldsymbol{\mu}_{\mathbf{z}_*}$	$\boldsymbol{\mu}_{\mathbf{z}_*} + 3\boldsymbol{\sigma}_{\mathbf{z}_*}$
Pitch ($\mathbf{z}_p$)	F_0 (Hz)	161.5	198.6	230.0
	Energy (dB)	−1.19	−1.19	−1.19
	Duration (s)	2.85	2.85	2.85
Volume ($\mathbf{z}_e$)	F_0 (Hz)	192.6	192.6	192.6
	Energy (dB)	−3.05	−1.03	1.07
	Duration (s)	2.85	2.85	2.85
Speech rate ($\mathbf{z}_d$)	F_0 (Hz)	192.6	192.6	192.6
	Energy (dB)	−1.19	−1.19	−1.19
	Duration (s)	1.90	2.86	3.65

can control both pitch and speech rate. To further confirm the benefits of the VAE-PF, we conduct A/B preference tests in pitch and speech rate dimensions between the system VAE-DR and VAE-PF. The results are shown in Fig. 5, where VAE-DR system uses 16-dimensional $\mathbf{z}_l$. The results demonstrate that with the separated prosodic features and separated 3 VAE modules, the VAE-PF model can significantly improve the synthesis quality over VAE-DR system.

However, when we adjust one prosody attribute dimension, there are still variations in the other attribute dimensions. Considering that we use separated prosodic features to learn individual prosody attribute representations, the generated individual attribute representation spaces may be correlated, which may not be desirable for independent control over individual prosody attributes. So we further analyze the prosody attribute representation spaces by visualizing the learned pitch, volume and speech rate attribute representations, $\mathbf{z}_p$, $\mathbf{z}_e$ and $\mathbf{z}_d$, using t-SNE (Der Maaten & Hinton, 2008), which is a technique for projecting high-dimensional vectors into a two-dimensional space. Results are shown in Fig. 6, where we use the same 30 utterances of each prosody attributes and each point corresponds to the projected $\mathbf{z}_p$ or $\mathbf{z}_e$ or $\mathbf{z}_d$ of a single utterance. Points are color-coded according to pitch, volume, and speech rate, respectively. In the figure, the entangled clusters of $\mathbf{z}_p$, $\mathbf{z}_e$ and $\mathbf{z}_d$ show clearly that independent control of individual prosody attributes is not successful.

6.4.2. Mutual information minimization for prosody attribute separation

With the VAE-PF, precise and direct prosody control has not been successful since the individual prosody attribute representations, $\mathbf{z}_p$, $\mathbf{z}_e$ and $\mathbf{z}_d$, remain highly entangled with each other. The fact that the information in one attribute dimension can leak into other attribute dimensions has made the independent control along individual attributes difficult. We further propose to minimize the mutual information between different prosody attributes in VAE-PF to make them more independent (i.e., VAE-MI). The t-SNE results of VAE-MI are shown in Fig. 7, where the

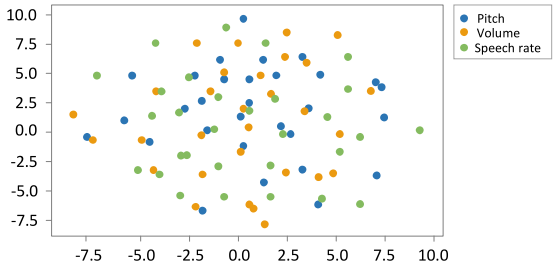

Fig. 6. Visualization of learned $\mathbf{z}_p$, $\mathbf{z}_e$ and $\mathbf{z}_d$ using two-dimensional t-SNE projected embeddings in the VAE-PF model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

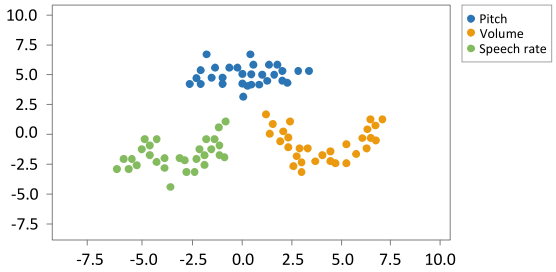

Fig. 7. Visualization of learned $\mathbf{z}_p$, $\mathbf{z}_e$ and $\mathbf{z}_d$ using two-dimensional t-SNE projected embeddings in the VAE-MI model.

Table 6
FFE and MOS evaluations of VAE-MI systems with different prosody attributes, and MOS evaluation with 95% confidence interval.

Prosody attribute	FFE	MOS
Pitch ($\mathbf{z}_p$)	35.8%	4.02 ± 0.06
Volume ($\mathbf{z}_e$)	36.7%	3.99 ± 0.07
Speech rate ($\mathbf{z}_d$)	38.5%	3.96 ± 0.04

same testing utterances tested by VAE-PF are used. The projected $\mathbf{z}_p$, $\mathbf{z}_e$ and $\mathbf{z}_d$ latent attributes are clearly separated into their own clusters to show that individual prosody attribute representations after the minimization of their mutual information are disentangled from each other.

In order to examine how effective the independent control of VAE-MI model over individual prosody attributes is, we evaluate performance along each individual attribute dimension, the results are shown in Table 5. The testing method is the same as that in VAE-PF system. When just adjusting the pitch dimension by using the $\boldsymbol{\mu}_{\mathbf{z}_p} - 3\boldsymbol{\sigma}_{\mathbf{z}_p}$, $\boldsymbol{\mu}_{\mathbf{z}_p}$ and $\boldsymbol{\mu}_{\mathbf{z}_p} + 3\boldsymbol{\sigma}_{\mathbf{z}_p}$, the measured F_0 varies from low to high, and the volume and speech rates essentially remain unchanged. Similarly, in the volume dimension, the measured energy (in log scale) varies from low to high without changing the pitch and speech rates. Similarly, when we vary the speech rate control, both pitch and volume dimensions are unchanged. These results clearly confirm that the VAE-MI can exercise independent control over individual prosody attributes in a reliable manner. Demonstrations of control over individual prosody attributes are given in the demo page.

The FFE and MOS evaluations of individual prosody attribute control are also conducted to assess the reconstruction performance of VAE-MI. Similarly, $\boldsymbol{\sigma}$ is separately the standard deviation of individual prosody attribute, $\mathbf{z}_p$, $\mathbf{z}_d$ and $\mathbf{z}_e$. The results are summarized in Table 6 where VAE-MI outperforms the VAE-PF significantly in each dimension. The performance gain is essentially contributed by the proposed mutual information minimization scheme in decorrelating the individual prosody attributes.

Besides, we conduct A/B preference tests along individual prosody attribute dimensions between the system VAE-PF and

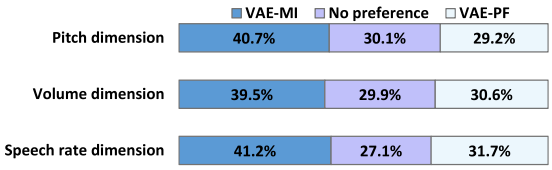

Fig. 8. Preference test results between VAE-PF and VAE-MI.

Table 7
FFE and MOS evaluations of GST* system, and MOS evaluation with 95% confidence interval.

Model	FFE	MOS
GST*	65.3%	3.48 ± 0.11

VAE-MI, the results are shown in Fig. 8, where the listeners give more preferences to the VAE-MI system, showing that adding the mutual information minimization estimation in individual prosody attribute representations improves reconstruction quality and yields more precise individual attribute control.

6.5. Comparison with GST-based models

This goal of this research is to solve the problem of interpretable and independent prosody control with separate VAEs, we like to compare our approach with GST-based model proposed previously. For a fair comparison, we replace the reference encoder and Tactron (Wang, et al., 2017) of original GST framework (Wang, et al., 2018) with the reference encoder discussed in Section 4.1 and Tactron 2 (Shen, et al., 2018) to build the GST-based model with the same training and testing data, which is denoted as GST*. During training, the log-mel spectrogram of the training target is fed to the reference encoder followed by a style token layer. The resulting style embedding is used to condition the Tacotron 2 text encoder states. During inference, we can feed an arbitrary reference signal to synthesize text with its speaking style. Alternatively, we can remove the reference encoder and directly control synthesis using the learned tokens. In the experiment, 10 style tokens are used as a condition control of the speaking style and prosody. Similar to the VAE-DR baseline, the GST* model is also a fully unsupervised system, which is difficult to disentangle or interpret the latent variables for independent control. When evaluating the meaning of individual tokens, we find that not every token can capture clean and separated attributes. While some tokens may capture just the speech rate attribute, others may represent the overall style in the training data as a mixture of multiple attributes. For example, a low F_0 token can also encode a slower speech rate attribute or a low volume attribute. Hence, the above objective metrics, the average F_0, energy and speech duration, are not used separately to evaluate the performance of the GST* model in each individual attribute sense.

We conduct the FFE and MOS evaluations to quantify the reconstruction performance of the GST* system, as shown in Table 7. From the results, we find that the GST* model is significantly inferior to the proposed VAE-MI system in their synthesized speech, both objectively and subjectively.

An A/B preference test is similarly conducted between GST* and VAE-MI. The 30 utterances are randomly selected from the two models separately and the results are shown in Fig. 9. The subjective listeners give a higher preference to the proposed approach than GST*, i.e., 46.4% to 25.7%, by a large margin.

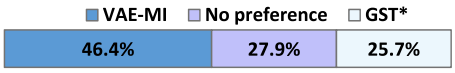

Fig. 9. Preference test results between GST* and VAE-MI.

7. Conclusion

In this paper, we investigate how to train a neural TTS for direct and independent prosody control along different attribute dimensions in an interpretable and reliable manner. To alleviate cross-attribute interactions in prosody control, we proposed two approaches: (1) adopting 3 individual VAE modules to isolate the real-valued, prosody information in F_0, energy and duration; (2) minimizing the mutual information between different prosody attributes to further reduce the cross-attribute interactions in prosody control. The scatter diagrams of t-SNE demonstrate the correlations between prosody attributes can be well disentangled by minimizing their mutual information. The prosody control performance of the proposed approaches has been quantitatively measured objectively by distortions and subjectively with MOS and A/B comparison listening tests, respectively. Both the objective and subjective evaluations show that the proposed prosody control is effective and MOS scores and A/B preference are improved progressively with the proposed approaches over various baseline systems.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by the National Key Research and Development Program of China (No. 2020AAA0108600).

References

Akuzawa, K., Iwasawa, Y., & Matsuo, Y. (2018). Expressive speech synthesis via modeling expressions with variational autoencoder. In *Proc. INTERSPEECH* (pp. 3067–3071).

An, X., Wang, Y., Yang, S., Ma, Z., & Xie, L. (2019). Learning hierarchical representations for expressive speaking style in end-to-end speech synthesis. In *Proc. ASRU* (pp. 184–191).

Arik, S. O., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., et al. (2017). Deep voice: Real-time neural text-to-speech. In *Proc. ICML* (pp. 195–204).

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proc. ICLR*.

Battenberg, E., Mariooryad, S., Stanton, D., Skerry-Ryan, R., Shannon, M., Kao, D. T. H., et al. (2019). Effective use of variational embedding capacity in expressive end-to-end speech synthesis. arXiv preprint arXiv:1906.03402.

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., & Bengio, Y. (2018). Mutual information neural estimation. In *Proc. ICML*.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., & Bengio, S. (2016). Generating sentences from a continuous space. In *Proc. CoNLL* (pp. 10–21).

Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-based models for speech recognition. In *Proc. NIPS* (pp. 577–585).

Chu, W., & Alwan, A. (2009). Reducing F0 frame error of F0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend. In *Proc. ICASSP* (pp. 3969–3972).

Den Oord, A. V., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., et al. (2018). Parallel WaveNet: Fast high-fidelity speech synthesis. In *Proc. ICML* (pp. 3915–3923).

Der Maaten, L. V., & Hinton, G. E. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579–2605.

Donsker, M. D., & Varadhan, S. S. (1983). Asymptotic evaluation of certain markov process expectations for large time. IV. *Communications on Pure and Applied Mathematics, 36*, 183–212.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial networks. arXiv preprint arXiv:1406.2661.

Habib, R., Mariooryad, S., Shannon, M., Battenberg, E., Skerry-Ryan, R., Stanton, D., et al. (2020). Semi-supervised generative modeling for controllable speech synthesis. In *Proc. ICLR*.

Hsu, W., Zhang, Y., Weiss, R. J., Chung, Y., Wang, Y., Wu, Y., et al. (2019). Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization. In *Proc. ICASSP* (pp. 5901–5905).

Hsu, W., Zhang, Y., Weiss, R. J., Zen, H., Wu, Y., Wang, Y., et al. (2019). Hierarchical generative modeling for controllable speech synthesis. In *Proc. ICLR*.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. ICML* (pp. 448–456).

Kenter, T., Wan, V., Chan, C., Clark, R. A. J., & Vit, J. (2019). CHiVE: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network. In *Proc. ICML* (pp. 3331–3340).

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In *Proc. ICLR*.

Krueger, D., Maharaj, T., Kramar, J., Pezeshki, M., Ballas, N., Ke, N. R., et al. (2017). Zoneout: Regularizing RNNs by randomly preserving hidden activations. In *Proc. ICLR*.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

Locatello, F., Bauer, S., Lucic, M., Ratsch, G., Gelly, S., Scholkopf, B., et al. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proc. ICLR*.

Morise, M., Yokomori, F., & Ozawa, K. (2016). WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, *99*(7), 1877–1884.

Park, J., Zhao, K., Peng, K., & Ping, W. (2019). Multi-speaker end-to-end speech synthesis. arXiv preprint arXiv:1907.04462.

Ping, W., Peng, K., Gibiansky, A., Arik, S. O., Kannan, A., Narang, S., et al. (2018). Deep voice 3: 2000-speaker neural text-to-speech. In *Proc. ICLR*.

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *Proc. ICML* (pp. 1278–1286).

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., et al. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *Proc. ICASSP* (pp. 4779–4783).

Skerry-Ryan, R., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., et al. (2018). Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron. In *Proc. ICML* (pp. 4693–4702).

Sotelo, J., Mehri, S., Kumar, K., Santos, J. F., Kastner, K., Courville, A., et al. (2017). Char2Wav: End-to-end speech synthesis. In *Proc. ICLR*.

Stanton, D., Wang, Y., & Skerry-Ryan, R. (2018). Predicting expressive speaking style from text in end-to-end speech synthesis. In *Proc. SLT* (pp. 595–602).

Sun, G., Zhang, Y., Weiss, R. J., Cao, Y., Zen, H., Rosenberg, A., et al. (2020). Generating diverse and natural text-to-speech samples using a quantized fine-grained VAE and autoregressive prosody prior. In *Proc. ICASSP*.

Sun, G., Zhang, Y., Weiss, R. J., Cao, Y., Zen, H., & Wu, Y. (2020). Fully-hierarchical fine-grained prosody modelling for interpretable speech synthesis. In *Proc. ICASSP*.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proc. NIPS* (pp 3104–3112).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Proc. NIPS* (pp. 5998–6008).

Wagner, M., & Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, *25*(7), 905–945.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., et al. (2017). Tacotron: Towards end-to-end speech synthesis. In *Proc. INTERSPEECH* (pp. 4006–4010.

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R., Battenberg, E., Shor, J., et al. (2018). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Proc. ICML* (pp. 5180–5189).

Wu, P., Ling, Z., Liu, L., Jiang, Y., Wu, H., & Dai, L. (2019). End-to-end emotional speech synthesis using style tokens and semi-supervised training. In *Proc. APSIPA ASC*.

Yang, S., Lu, H., Kang, S., Xue, L., Xiao, J., Su, D., et al. (2020). On the localness modeling for the self-attention based end-to-end speech synthesis. *Neural Networks*, *125*, 121–130.

Yang, F., Yang, S., Zhu, P., Yan, P., & Xie, L. (2019). Improving mandarin end-to-end speech synthesis by self-attention and learnable gaussian bias. In *Proc. ASRU* (pp. 208–213).

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., et al. (2002). *The HTK book*.

Zhang, Y., Pan, S., He, L., & Ling, Z. (2019). Learning latent representations for style control and transfer in end-to-end speech synthesis. In *Proc. ICASSP* (pp. 6945–6949).