INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Applying the Information Bottleneck Principle to Prosodic Representation Learning

Guangyan Zhang[1], *Ying Qin*[2], *Daxin Tan*[1], *Tan Lee*[1]

[1] Department of Electronic Engineering, The Chinese University of Hong Kong, China
[2] Institute of Information Science, Beijing Jiaotong University, China
{gyzhang, daxintan}@link.cuhk.edu.hk, yingqin@bjtu.edu.cn, tanlee@cuhk.edu.hk

Abstract

This paper describes a novel design of a neural network-based speech generation model for learning prosodic representation. The problem of representation learning is formulated according to the information bottleneck (IB) principle. A modified VQ-VAE quantized layer is incorporated in the speech generation model to control the IB capacity and adjust the balance between reconstruction power and disentangle capability of the learned representation. The proposed model is able to learn word-level prosodic representations from speech data. With an optimized IB capacity, the learned representations not only are adequate to reconstruct the original speech but also can be used to transfer the prosody onto different textual content. Extensive results of the objective and subjective evaluation are presented to demonstrate the effect of IB capacity control, the effectiveness, and potential usage of the learned prosodic representation in controllable neural speech generation.

Index Terms: speech prosody, prosodic representation, information bottleneck principle

1. Introduction

Prosody is an essential component embedded in human speech. Carrying a full array of linguistic and paralinguistic functions, prosody manifests tone, stress, intonation, and rhythm of speech and contributes to naturalness, style, attitude, and emotion of speech [1, 2]. Conventionally speech prosody has been studied through the analysis of pitch, intensity, and duration features, focusing on one or more specific functions of prosody. This approach does not provide a holistic representation of prosody for speech generation purposes. In recent years, neural network-based speech synthesis systems [3–5] show superior performance in terms of speech quality and naturalness and offer an effective approach to learning speech representations. The problem of prosodic representation learning has attracted particular attention in view of its potential use in prosody transfer and style control [6–8], prosody control [9, 10].

Various approaches to unsupervised learning of prosodic representation have been proposed [6–9, 11–14]. In these studies, a neural network model is used to disentangle contributing factors of input speech and subsequently perform speech reconstruction from the disentangled factors. The prosodic representation is obtained as one of the learned factors, parallel with non-prosodic factors that correspond to content, speaker, channel, etc. In [12, 15–17], adversarial learning was applied to address the problem that the learned prosodic representation might contain substantial information related to non-prosodic factors. The use of an adversarial classifier requires the availability of the labels for one of the disentangled non-prosodic factors. The design of the adversarial classifier is specific to only one non-prosodic factor and can not be applied to other non-prosodic factors. Furthermore, the non-prosodic factors(e.g., speaker) might be related to prosody [15], while disentangling with an adversarial classifier might also result in low prosody information in the prosodic representation.

In the present study, prosodic representations learning is tackled from the perspective of information bottleneck (IB), by which a good representation is determined with the trade-off between its predictive/reconstructive power and compact representation [18]. We propose to define a good prosodic representation in three different aspects. First, it should have a good capability of capturing prosody-related information. It can reconstruct the reference speech conditioned on the other necessary factors (good reconstructive power). Second, the representation is expected to include as little as possible information about non-prosodic factors (compact representation). Third, the predicted prosodic representation should contribute to generating natural speech. The prosodic representations extracted from reference speech could be used to train a prosody predictor, which predicts the prosodic representation from the text. Specific speech generation applications (e.g., TTS) require appropriate and expressive prosody predicted from the text [10]. The predicted prosodic representation is expected to help improve the naturalness or expressiveness of the generated speech.

The contributions of this study are as follows: (1) the prosodic representation learning problem is formulated based on the IB principle; (2) a prosodic representation learning system with controllable IB capacity is developed; (3) subjective and objective evaluation results show that learned prosodic representations have good reconstructive power and can concisely capture prosody-related information by choosing appropriate IB capacity; (4) a machine-translation-based prosody predictor is proposed to realize text-to-prosody prediction. The predictor can generate prosodic representations for improving the naturalness of synthesized speech.

2. Problem Formulation

The IB principle aims to learn robust representation from data for a specific task, e.g., classification, auto-encoding [19]. Let Z denote the representation to be learned from data X, and Y denote the task target[1]. The goal of learning is to maximize the mutual information between Z and Y, and meanwhile, to discard irrelevant information about Y that might be present in the data X. Let $I(Z;Y)$ and $I(X;Z)$ be the mutual information between Z and Y and that between X and Z respectively. That is, the following objective function is to be maxi-

[1] For auto-encoding, the task target Y is the same as X.

http://dx.doi.org/10.21437/Interspeech.2021-1049

mized [18, 20, 21],

$$I(Z;Y) - \beta I(X;Z),^2 \tag{1}$$

where β is a Lagrange multiplier.

Recent studies [6–9, 11–14] share similar strategies for learning prosodic representation Z from speech data X. The learning model encodes input X into prosodic representation Z, and reconstructs X from Z in conjunction with the representations of non-prosodic factors F_i, e.g., speaker, channel, speech content. Following the information bottleneck framework, prosodic representation learning is achieved by maximizing

$$I(Z;X|F_1, F_2, \cdots, F_i, \cdots, F_N) - \beta I(X;Z). \tag{2}$$

With the first term in Equation 2, Z is expected to have the capability of reconstructing X, given F_i. The second term in Equation 2 is introduced as a constraint to prevent Z from being a direct copy of X. The present study is carried out with a single-speaker corpus, and it is assumed that content is the only non-prosodic factor needed.

The main challenge of applying the IB principle is on the computation of mutual information. Variational inference is regarded as a practical way to approximate the calculation [19]. In this study, variational inference is adopted to establish a lower bound on the IB objective function [19, 22] in Equation 2. Let x, z and t be the speech, prosodic representation, and content factor, respectively. $q_\phi(z|x)$ is the variational approximation of the true posterior $p_\theta(z|x)$, which is implemented by a reference encoder as described in subsection 3.1. The conditional probability $p_\theta(x|z,t)$ is modeled by a conditional decoder which aims to reconstruct speech (subsection 3.2). $p(z)$ is the variational approximation to the marginal distribution of z. The approximation of Equation 2 gives the objective function,

$$E_{q_\phi(z|x)}[\log p_\theta(x|z,t)] - \beta D_{KL}(q_\phi(z|x)||p(z)). \tag{3}$$

The first term in Equation 3 represents the reconstruction task. In the second term of Equation 3, $D_{KL}(q_\phi(z|x)||p(z))$ indicates the KL divergence between $q_\phi(z|x)$ and $p(z)$. The KL divergence $D_{KL}(q_\phi(z|x)||p(z))$, known as the IB capacity, is the upper bound of the mutual information $I(X;Z)$ in Equation 2 [23, 24]. Control of information transmitted from speech to prosodic representation can be achieved by adjusting the IB capacity. With limited IB capacity, the prosodic representation is expected to contain mostly prosody information. As the IB capacity increases, information of other factors would be absorbed in the prosodic representation.

3. Model Design

The proposed system consists of a reference encoder and a conditional decoder, as shown in Figure 1. In the present study, the prosodic representation is learned at the word level. The reference encoder extracts a sequence of word-level prosodic representations with controlled IB capacity from an input mel-spectrogram. The conditional decoder reconstructs the mel-spectrogram from the prosodic representation.

3.1. Reference Encoder

The reference encoder is made up of a feature extractor and a bottleneck layer. The feature extractor takes mel-spectrogram as input and generates word-level acoustic features. These features are processed by the bottleneck layer and encoded into a sequence of prosodic representations.

3.1.1. Feature Extractor

The feature extractor contains four Feed-Forward Transformer (FFT) [25] blocks. Each FFT block consists of a self-attention [26] and a 1D convolutional network. The input mel-spectrogram is first added with sinusoid positional encoding, then passed through the feature extractor. The feature extractor outputs frame-level acoustic features. The frame-level features are aggregated to phoneme-level features by average pooling. In the same manner, the syllable and word-level acoustic features are obtained.

3.1.2. Bottleneck Layer

The bottleneck layer takes word-level acoustic features as input and outputs prosodic representations. It aims to remove non-prosodic information (e.g., speaker, content, and channel). In [24], the bottleneck layer was implemented with the standard Variational AutoEncoder (VAE) [27]. However, this leads a notorious issue when the bottleneck layer is trained on sequential data (e.g., speech or text): the output representation forgets all information regarding the data and has the identical distribution as the uninformative marginal prior, yielding the so-called KL vanishing problem [28]. In this paper, the bottleneck layer is implemented by a modified Vector Quantized VAE (VQ-VAE) quantized layer [29]. The VQ-VAE quantized layer has a deterministic latent representation and does not have the problem of KL vanishing. VQ-VAE realizes a bottleneck that quantizes input feature, directly limiting the amount of information embedded in learned representation. It learns a dictionary (codebook) $\mathbf{E} \in \mathbb{R}^{K \times D}$, where K is the dictionary size, and D is the dimension of each code $\mathbf{e}_i$. The continuous feature $\mathbf{x}$ can be encoded as the discrete code z using the nearest neighbor lookup. The variational posterior $q_\phi(z|\mathbf{x})$ is deterministic. The marginal distribution $p(z)$ based on variational approximation is assumed to be a simple uniform distribution. The KL divergence between $q_\phi(z|\mathbf{x})$ and $p(z)$ is equal to $\log K$.

The IB capacity of the VQ-VAE quantized layer (i.e., $\log K$) can be controlled by changing the dictionary size K. The larger the size K, the higher the IB capacity. However, increasing the dictionary size would cause the out-of-memory problem. The coding strategy in [30] was used to mitigate the mode collapse problem in VQ-VAE. It can increase the bottleneck layer IB capacity without taking up extra memory,but the IB capacity cannot be controlled and the coding strategy has not been used for learning prosodic representations. In this study, a novel prosodic representation learning system with controllable IB capacity is proposed.

The word-level prosodic representation can be encoded with multiple entries in the dictionary $\mathbf{E}$. The acoustic feature $\mathbf{x} \in \mathbb{R}^D$ is first divided equally into G groups and arranged as a matrix $\mathbf{X}' = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_j, \cdots \mathbf{x}_G] \in \mathbb{R}^{(D/G) \times G}$. Each column of $\mathbf{X}'$ can be encoded by an integer index independently as the prosodic representation $\mathbf{z}' = [z_1, z_2, \cdots, z_j, \cdots, z_G] \in \mathbb{Z}^{1 \times G}$. The variational posterior distribution $q_\phi(\mathbf{z}'|\mathbf{X}')$ and marginal distribution $p(\mathbf{z}')$ can be represented as,

$$\begin{aligned} q_\phi(\mathbf{z}'|\mathbf{X}') &= q_\phi(z_1, z_2, \cdots, z_j, \cdots, z_G|\mathbf{X}') \\ &= \prod_{j=1}^{G} q_\phi(z_j = k|\mathbf{x}_j), \end{aligned} \tag{4}$$

[2] In this paper, X, Y, Z, F_i are random variables and x, z, t are instances of random variables.

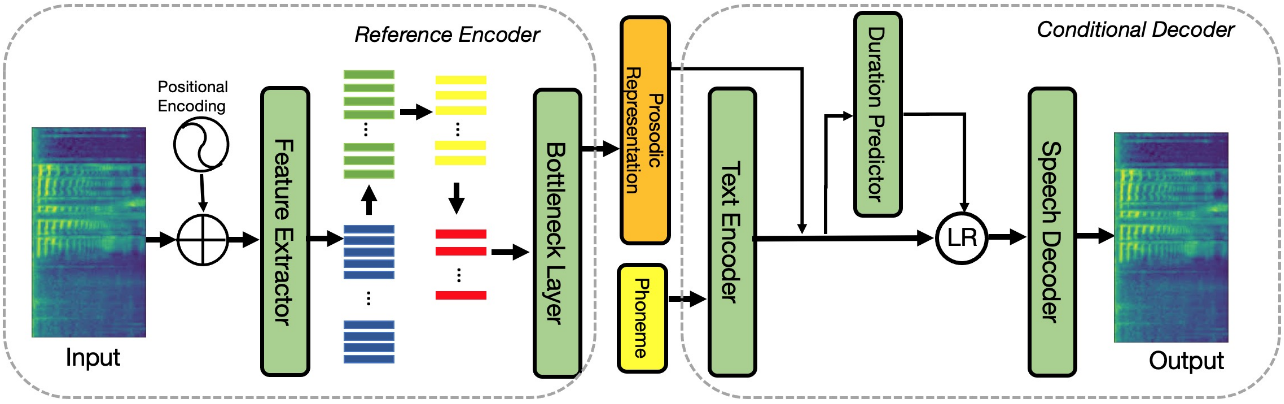

Figure 1: *The structure of prosodic representation learning system. The vectors in blue, green, yellow, and red colors represent frame, phoneme, syllable and word-level acoustic features respectively.*

$$p(\mathbf{z}') = p(z_1, z_2, \cdots, z_j, \cdots, z_G) = \prod_{j=1}^{G} p(z_j). \tag{5}$$

Then, the $D_{KL}(q_\phi(\mathbf{z}'|\mathbf{X}')||p(\mathbf{z}'))$ can be derived as,

$$\begin{aligned} D_{KL}(q_\phi(\mathbf{z}'|\mathbf{X}')||p(\mathbf{z}')) &= \sum_{\mathbf{z}'} q_\phi(\mathbf{z}'|\mathbf{X}') \log \frac{q_\phi(\mathbf{z}'|\mathbf{X}')}{p(\mathbf{z}')} \\ &= G \log K. \end{aligned} \tag{6}$$

In this paper, the group number G is set as 2.

3.2. Conditional Decoder

The design of the conditional decoder follows the Fastspeech2 system [31], which comprises a text encoder and a speech decoder. The phone sequence is passed through the text encoder to obtain phone-level features. The word-level prosodic representations are up-sampled to match the granularity of phone level and concatenated to the phone-level features. The Length Regulator (LR) module further up-samples all phone-level features to frame-level features, which are presented to the speech decoder to generate mel-spectrogram. The phone-to-frame upsampling operation requires phone duration information provided by forced alignment in the training stage. At the inference stage, the duration is predicted with a duration predictor.

4. Performance Evaluation

4.1. Experimental Setup

A series of experiments are carried out with the part of Blizzard 2013 English dataset. Phone-level time alignment and word-to-syllable conversion were performed using the Festival software toolkits [32]. The Parallel WaveGAN vocoder [33] is used to generate speech waveform from mel-spectrograms.

Different settings of IB capacities are experimented. The capacity is changed by varying the dictionary size K as described in subsubsection 3.1.2. The dictionary sizes and corresponding IB capacities are shown in Table 1.

Table 1: *Different dictionary sizes and corresponding IB capacities set for bottleneck layer.*

dictionary size	0	2	4	8	16	32	64
IB capacity(nats)	0	1.39	2.77	4.16	5.54	6.93	8.31

4.2. Reconstruction of Speech

In this task, the reference speech is reconstructed by the conditional decoder using the learned prosodic representations and the phone sequence from reference speech. The reconstruction performance reflects whether the learned prosodic representation carries enough prosody information. Four metrics that are related to prosody in the acoustic aspect, including Voicing Decision Error (VDE), Gross Pitch Error (GPE), F0 Frame Error (FFE), and Mel Cepstral distortion (MCD$_{22}$) are used to evaluate the reconstruction performance [6]. The results are shown as in Table 2, where smaller values mean better performance. As the IB capacity is increased from 0nats to 8.31 nats, the reconstruction performance improves. The values of all four metrics drop rapidly as the capacity starts to increase. This suggests that the information passed to the prosodic representation is most useful for reconstruction and highly related to speech prosody. As the capacity continues to increase, the degree of performance improvement becomes less significant. For the range of 6.93 nats to 8.31 nats, the GPE, FFE, MCD$_{22}$ fall slightly by 0.26%, 0.51%, and 0.09, respectively. It suggests that the information passed to the prosodic representation has a low correlation with prosody as the IB capacity is large. Readers are recommended to listen to demo examples [3].

Table 2: *Objective evaluation on speech reconstruction.*

IB capacity(nats)	VDE(%)	GPE(%)	FFE(%)	MCD$_{22}$
0	15.14	39.10	37.39	6.74
1.39	11.48	27.70	27.39	6.13
2.77	10.83	13.21	18.30	5.69
4.16	10.21	9.05	15.38	5.57
5.54	**9.37**	7.56	13.72	5.43
6.93	9.89	6.40	13.47	5.17
8.31	9.42	**6.14**	**12.96**	**5.08**

4.3. Mutual Information between Content and Prosodic Representation

In Section 4.2, it is noted that the IB capacity should be properly controlled to reach a balance between reconstruction performance and the efficacy of prosodic representation. Hence

[3] https://patrick-g-zhang.github.io/pt/

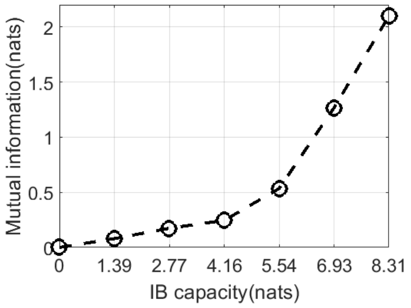

Figure 2: *Mutual information between content and prosodic representation.*

the mutual information between content and prosodic representation is evaluated. The phone embeddings in each word are aggregated to form a continuous-valued vector. The mutual information between this word-level vector and the respective prosodic representation is estimated by MINE [34]. The results of the evaluation are shown as in Figure 2. It can be seen that the mutual information increases mildly as the IB capacity is in the range of 0 to 5.54 nats, meaning that the prosodic representation does not entangle with much content information. The mutual information increases sharply as the IB capacity exceeds 5.54 nats. In this case, content information is leaked into the learned prosodic representation.

4.4. Cross-text Prosody Transfer

The efficacy of learned prosodic representation can be evaluated in cross-text prosody transfer, where speech is generated using the prosodic representation and a given phone sequence which is different from reference speech. Subjective evaluation was carried out by a listening test implemented via the Amazon Mechanical Turk platform. 20 native English speakers participated in the listening test. There were two tasks in the test. One aimed to evaluate the effectiveness of prosody transfer. The listener was asked to give a 5-point preference score on the prosody similarity between the reference speech and the generated speech, ranging from “completely different” (1) to “completely same” (5). The other task was designed to evaluate the clearness of generated speech content. Each listener was required to give a score that rates the clearness of generated speech content. The score ranges from “completely unclear” (1) to “completely clear” (5) with five levels.

Results of the subjective evaluation are shown in Table 3. In terms of prosody similarity, the score grows initially and then saturates as the IB capacity increases. The clearness of content gradually declines as the IB capacity increases and degrades significantly when the IB capacity increases from 5.54 nats to 6.93 nats. By examining the test samples of generated speech, it is found that the deterioration of clearness is mainly caused by slurred words. At the IB capacity of 5.54 nats (dictionary size of 16), the learned prosodic representation achieves the best performance in subjective evaluation. This represents a balance between prosodic and non-prosodic factors embedded in the prosodic representation.

Table 3: *Subjective evaluation scores of cross-text prosody transfer.*

IB capacity(nats)	prosody similarity	content clearness
0	1.87 ± 0.73	$\mathbf{3.61 \pm 0.88}$
1.39	2.33 ± 0.84	3.53 ± 0.94
2.77	2.96 ± 0.82	3.50 ± 0.89
4.16	3.12 ± 0.86	3.49 ± 0.90
5.54	$\mathbf{3.25 \pm 0.84}$	3.49 ± 0.94
6.93	3.24 ± 0.86	3.23 ± 0.96
8.31	$\mathbf{3.25 \pm 0.89}$	3.17 ± 0.90

4.5. Predicting Prosodic Representation from Text

A prosody predictor is proposed to predict the prosodic representation given a text sequence. The input of the prosody predictor is given by the contextual embedding extracted from the BERT [35, 36], which incorporates rich syntactic and semantic information. The output of the prosody predictor is the prosodic representation. The prosody predictor model follows the transformer-based machine translation system [37] with both text and prosodic representation being discrete. The prosody predictor is trained with prosodic representations extracted with different IB capacities settings. The predicted prosodic representations are then passed to the conditional decoder to generate speech. A MOS test was carried out to evaluate the naturalness of generated speech, which is shown as in Table 4. Compared to the case with no input to the conditional decoder (i.e., IB capacity of 0), the predicted prosodic representations with IB capacity greater than 0 could improve the naturalness of generated speech. The predicted prosodic representations with an IB capacity of 5.54 nats (dictionary size of 16) achieve the highest score among all settings. It suggests that with appropriate control of IB capacity, the predicted prosodic representation could help a TTS system improve the naturalness of output speech.

Table 4: *MOS test result.*

IB capacity(nats)	0	1.39	2.77	4.16	5.54	6.93	8.31
MOS	3.75	3.79	3.80	3.87	**3.93**	3.92	3.85

5. Conclusions

The information bottleneck (IB) principle is applied to learning word-level prosodic representation from speech data. Through proper control of the IB capacity, the learned prosodic representations, on the one hand, are adequate for speech reconstruction and, on the other hand, compactly capture prosody-related information in speech. The predicted prosodic representation from the prosody predictor has shown the potential to improve the naturalness of speech for existing TTS systems. In our future work, speaker or channel factors will be investigated in the multi-speaker and multi-channel scenarios.

6. Acknowledgement

This research is partially supported by a Tier 3 funding from ITSP (Ref: ITS/309/18) of the Hong Kong SAR Government, and a Knowledge Transfer Project Fund (Ref: KPF20QEP26) from the Chinese University of Hong Kong.

We also thank to Xu Qin, Wei Zhou, and Zhiyuan Peng for their kind help and discussions.

7. References

[1] P. A. Taylor, Text-to-speech synthesis. Cambridge University Press, 2009.

[2] M. Wagner and D. G. Watson, "Experimental and theoretical advances in prosody: A review," Language and cognitive processes, vol. 25, no. 7-9, pp. 905–945, 2010.

[3] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio et al., "Tacotron: Towards end-to-end speech synthesis," in Proc. Interspeech, Aug. 2017, pp. 4006–4010.

[4] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in Proc. ICASSP, 2018, pp. 4779–4783.

[5] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in Proc. AAAI, vol. 33, no. 01, 2019, pp. 6706–6713.

[6] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in Proc. ICML, 2018, pp. 4700–4709.

[7] V. Klimkov, S. Ronanki, J. Rohnke, and T. Drugman, "Fine-grained robust prosody transfer for single-speaker neural text-to-speech," in Proc. Interspeech, 2019, pp. 4440–4444.

[8] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in Proc. ICASSP, 2019, pp. 5911–5915.

[9] G. Zhang, Y. Qin, and T. Lee, "Learning syllable-level discrete prosodic representation for expressive speech generation," in Proc. Interspeech, 2020, pp. 3426–3430.

[10] Z. Hodari, O. Watts, and S. King, "Using generative modelling to produce varied intonation for speech synthesis," in Proc. 10th ISCA Speech Synthesis Workshop, 2019, pp. 239–244.

[11] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, and Y. Wu, "Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis," in Proc. ICASSP, 2020, pp. 6264–6268.

[12] D. Tan and T. Lee, "Fine-grained style modeling, transfer and prediction in text-to-speech synthesis via phone-level content-style disentanglement," arXiv preprint arXiv:2011.03943, 2020.

[13] X. Wang, S. Takaki, J. Yamagishi, S. King, and K. Tokuda, "A vector quantized variational autoencoder (vq-vae) autoregressive neural f0 model for statistical parametric speech synthesis," IEEE/ACM Trans. on Audio, Speech, and Language Processing, vol. 28, pp. 157–170, 2019.

[14] T. Kenter, V. Wan, C.-A. Chan, R. Clark, and J. Vit, "Chive: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network," in Proc. ICML, 2019, pp. 3331–3340.

[15] G. Zhang, S. Qiu, Y. Qin, and T. Lee, "Estimating mutual information in prosody representation for emotionalprosody transfer in speech synthesis," in Proc. ISCSLP, 2021, pp. 1–5.

[16] W.-N. Hsu, Y. Zhang, R. J. Weiss, Y.-A. Chung, Y. Wang, Y. Wu, and J. Glass, "Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization," in Proc. ICCASP. IEEE, 2019, pp. 5901–5905.

[17] S. Karlapati, A. Moinet, A. Joly, V. Klimkov, D. Sáez-Trigueros, and T. Drugman, "CopyCat: Many-to-Many Fine-Grained Prosody Transfer for Neural Text-to-Speech," in Proc. Interspeech, 2020, pp. 4387–4391.

[18] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in Proc. Information Theory Workshop, 2015, pp. 1–5.

[19] A. Alemi, I. Fischer, J. Dillon, and K. Murphy, "Deep variational information bottleneck," in Proc. ICLR, 2017.

[20] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," arXiv preprint physics/0004057, 2000.

[21] G. Chechik, A. Globerson, N. Tishby, Y. Weiss, and P. Dayan, "Information bottleneck for gaussian variables." Journal of machine learning research, vol. 6, no. 1, 2005.

[22] A. Alemi, B. Poole, I. Fischer, J. Dillon, R. A. Saurous, and K. Murphy, "Fixing a broken elbo," in Proc. ICML, 2018, pp. 159–168.

[23] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework." in Proc. ICLR, 2017.

[24] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in beta-vae." in Proc. ICLR, 2018.

[25] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," in Proc. NIPS, 2019, pp. 3171–3180.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. NIPS, 2017, pp. 3171–3180.

[27] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," stat, vol. 1050, p. 1, 2014.

[28] H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, and L. Carin, "Cyclical annealing schedule: A simple approach to mitigating kl vanishing," in Proc. NAACL, 2019, pp. 240–250.

[29] A. Van Den Oord, O. Vinyals et al., "Neural discrete representation learning," in Proc. NIPS, 2017, pp. 6306–6315.

[30] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," Proc. ICLR, 2020.

[31] Y. Ren, C. Hu, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text-to-speech," Proc. ICLR, 2021.

[32] P. Taylor, A. W. Black, and R. Caley, "The architecture of the festival speech synthesis system," in Proc. The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis, 1998.

[33] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. ICASSP, 2020, pp. 6199–6203.

[34] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in Proc. ICML, 2018, pp. 531–540.

[35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL, 2019, pp. 4171–4186.

[36] G. Jawahar, B. Sagot, and D. Seddah, "What does bert learn about the structure of language?" in Proc. ACL, 2019, pp. 3651–3657.

[37] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in Proc. NAACL, 2019, pp. 48–53.