Disentangling Speech Representations with Mutual Information Estimators for Expressive Synthesis

Thomas Kassiotis
Department of Computer Science
University of Crete
Heraklion, Greece
tkassiotis@csd.uoc.gr

Yannis Pantazis
Institute of Applied and Computational Mathematics
Foundation for Research and Technology - Hellas
Heraklion, Greece
pantazis@iacm.forth.gr

***Abstract*—Disentangled speech representation allows for precise control over individual speech attributes, such as content, speaker identity, and style, enabling more flexible and natural voice synthesis engines. This study advances speech synthesis by developing innovative disentangled speech representation algorithms. Techniques grounded in Information Theory such as recently-proposed regularized variational mutual information estimators supplemented with gradient reversal layer were integrated to refine the representation of independent speech attributes. Using the Expresso dataset within the FastSpeech 2 framework, this work demonstrates significant improvements in the controllability and quality of synthetic speech. Objective metrics including cosine similarity matrices, perceptual evaluation of speech quality (PESQ), and short-term objective intelligence (STOI), complemented by subjective assessment of speech quality, were evaluated. The results show that the proposed methods outperform existing approaches, evidenced by superior A/B testing outcomes, improved inter-cluster distance metrics, and enhanced PESQ and STOI scores, highlighting the advancements of the developed systems in intelligibility, naturalness, and overall speech quality.**

***Index Terms*—Disentangled Speech Representation, Deep Learning, Speech Synthesis, Information Theory**

I. Introduction

Speech synthesis technologies have become integral to various applications, from enabling real-time communication between users and virtual assistants to providing voiceover for multimedia content. The effective representation of speech signals is central to the advancement of these technologies. Traditional speech synthesis systems, such as concatenative [1], [2] and formant synthesis [3], provided foundational techniques for controlling speech attributes, but lacked flexibility and scalability. With the advent of deep learning, models like Tacotron [4] and WaveNet [5] introduced end-to-end architectures that significantly improved the quality of synthesized speech as well as towards few-shot or even zero-shot speech generation.

The disentanglement of speech attributes, separating content, speaker identity and style, is critical too enhancing the clarity, naturalness, and personalization of synthesized speech. Recent studies have used various architectures and frameworks to enhance disentanglement capabilities. Variational autoencoders [6], [7] and Generative Adversarial Networks [8], [9] have been particularly effective in learning more interpretable and isolated representations of speech features. Despite these advancements, current speech synthesis systems frequently face challenges in effectively separating key speech attributes, leading to synthesized speech that lacks naturalness and is difficult to control in terms of speaker characteristics and emotional tone. A key challenge in this field is content leakage, where content unintentionally influences style embeddings, and style leakage, where speaker traits affect style embeddings. This limitation stems largely from deficiencies in existing speech representation methods, which fail to isolate these attributes effectively.

Incorporating Information Theory tools into disentangled representation learning has opened new avenues for improving the precision of attribute manipulation. Variational mutual information (VMI) estimators have been particularly influential, as seen in the work of Belghazi et al. [10], which leverages these estimators by imposing independence between latent representations. However, existing estimators such as Mutual Information Neural Estimation (MINE) [10], Information Noise-Contrastive Estimation. (InfoNCE) [11], and Contrastive Log-ratio Upper Bound (CLUB) [12], suffer from high variance. This variance leads to fluctuating estimates depending on the sample or batch of data used during training, which can result in erroneous estimates of mutual information. This instability reduces the accuracy of capturing true dependencies between variables, affects model convergence, and degrades performance, resulting in less robust estimators and slower training. Reducing this variance is crucial for improving the reliability and efficiency of MI-based disentanglement methods.

This study addresses the issue of high variance in Variational Mutual Information (VMI) estimators by introducing two regularized estimators. Specifically, two recently-proposed probabilistic divergences [13], namely the Worst Case Regret (WCR) and Convex Conjugate Rényi Divergence (CCR), are employed as low-variance VMI estimators. These regularized divergences, extending the approach in [14], use Lipschitz continuous functions with bounded derivatives, providing smoother regularization compared to previous estimators that relied solely on bounded but potentially abruptly changing functions, such as step functions. By constraining function gradients and preventing sharp variations, the WCR and CCR approaches yield statistically stable VMI estimates with

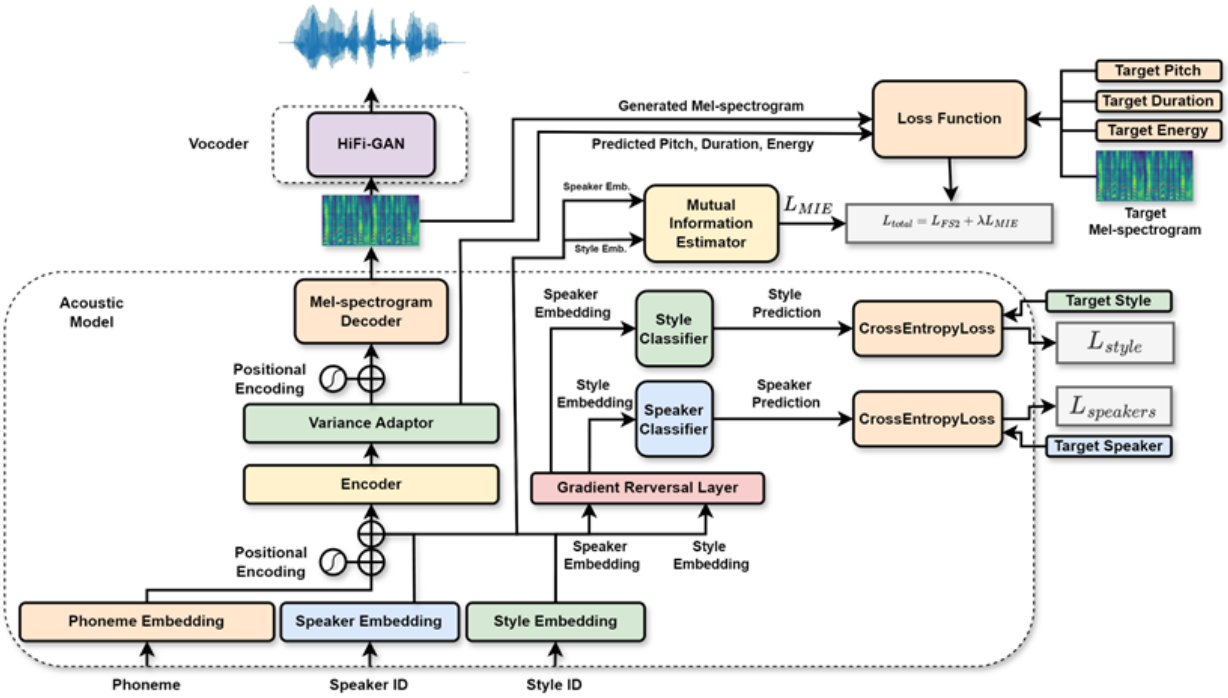

Fig. 1. Extension of FastSpeech 2 architecture with new embedding layers for speaker (blue box) and style (green box), and advanced disentanglement techniques for robust speech synthesis. VMI estimator is shown in the yellow box while GRL is shown with in the red box.

considerably reduced fluctuations. The integration of these regularized VMI estimators, along with Gradient Reversal Layers (GRL) [15], into the FastSpeech 2 model [16] resulted in significantly enhanced capability for effective separation and precise control of various speech attributes.

II. Materials, Models, and Methods

A. Dataset

The Expresso Dataset [17], provided by Meta AI, serves as the primary data source for this study. This dataset consists of high-quality expressive speech waveforms recorded at 48 kHz. It is specifically designed for speech synthesis and analysis tasks, featuring annotated audio recordings that facilitate research in these areas. The dataset includes speech samples from four speakers –two female and two male– and offers a variety of speech styles to enhance the robustness and generalization of speech models. Indeed, the Expresso Dataset supports a comprehensive array of expressive speech samples from seven distinctive speech styles. In comparison, while datasets such as VCTK Corpus and LibriTTS include recordings from multiple speakers, they lack samples with expressive variations, such as emotions or intonations. In contrast, the Expresso Dataset supports both a multi-speaker configuration and a comprehensive array of expressive speech samples from seven distinctive speech styles.

B. Model Overview

The proposed model extends FastSpeech 2 by incorporating speaker identity and style embeddings along with disentanglement techniques to improve expressive speech synthesis as shown in Figure 1. Two additional embedding layers are introduced: a speaker embedding layer (4×256) representing four distinct speakers and a style embedding layer (7×256) capturing seven speech styles (e.g., confused, enunciated, happy, sad). These conditional embeddings control the synthesis process, enabling variation in speaker and style attributes.

To promote disentangled representations, VMI estimation methods are optimized to minimize dependencies between speaker and style embeddings. Both existing VMI estimators – MINE, InfoNCE, and CLUB– as well as novel VMI estimation techniques –CCR and WCR– are implemented and tested. A GRL is also introduced leveraging adversarial training enforcing stronger separation between speaker and style representations. The GRL requires the training of two classifiers: a speaker classifier and a style classifier which are trained offline. The model is trained on the Expresso dataset using a customized data pipeline, ensuring efficient preprocessing, speaker-style conditioning, and large-scale batch processing.

C. Variational Mutual Information Estimators

Mutual information is defined as the Kullback-Leibler divergence between the joint distribution of a pair of variables, denoted by $P_{XY}(X, Y)$, and the product of marginals, denoted by $P_X(X)P_Y(Y)$. MI value equals to 0 implies that the random variables X and Y are independent. Thus, minimizing the MI between two embeddings results in an unsupervised approach to disentangle them. In our case, speaker (X) and style (Y) embeddings from the same audio sample are concatenated to form a sample from the joint distribution. In contrast, a sample from the product of marginal distribution

is generated by concatenating the speaker embedding with a randomly permuted style embedding. This approach disrupts the dependence between the original speaker and style embeddings leading to a sample from the product of marginal distribution.

The estimation of MI from samples is a challenging problem especially in the high dimensional setting. Therefore, variational formulas for the Kullback-Leibler divergence or directly for MI have been proposed. As already mentioned, those VMI estimators such as MINE, InfoNCE and CLUB, have high variance. Instead of Kullback-Leibler divergence, we propose to use a different family of regularized divergences which also admit variational representation formulas for the estimation of MI. The CCR divergence reformulates traditional Rényi divergence by eliminating risk-sensitive terms, thereby improving the stability of MI estimation for disentanglement. The definition of CCR divergence is given by

$$D_\alpha(P_{XY}\|P_XP_Y) = \sup_{g\in \text{Lip}^1: g>0} \left\{ \frac{1}{\alpha-1}\log \int\int |g(x,y)|^{\frac{\alpha-1}{\alpha}} P_{XY}(x,y)\,dx\,dy - \int\int g(x,y)P_X(x)P_Y(y)\,dx\,dy \right\} + \frac{1}{\alpha}(\log a + 1) \tag{1}$$

where Lip^1 is the Lipschitz continuous function space with Lipschitz constant equal to 1 while α is a scalar controlling the focus put by the CCR divergence to the tails of the two distributions. Taking the limit as $\alpha \to \infty$, the WCR divergence is obtained. Its variational formula is given by

$$D_\infty(P_{XY} \parallel P_XP_Y) = \sup_{g\in \text{Lip}^1: g>0} \left\{ \log \int\int |g(x,y)|P_{XY}(x,y)\,dx\,dy - \int\int g(x,y)P_X(x)P_Y(y)\,dx\,dy \right\} + 1 \tag{2}$$

Both CCR and WCR divergences extend the traditional MI estimation techniques by focusing on the maximum possible discrimination between distributions, ensuring enhanced robustness in disentanglement processes. WCR divergence in particular is designed to address worst-case scenarios in the variability of speaker-style attributes, by measuring how significantly two distributions can diverge from one another under the least probable regimes.

The estimation of the above formulas requires the following two approximations. Firstly, the function g is parameterized by a neural network comprising four linear layers with ReLU activations. The constraint that $g \in \text{Lip}^1$, is enforced via a standard gradient penalty to bound excessive variations of the derivatives [18]. Secondly, the integrals are approximated by statistical averages using the available samples from both the joint and the product of marginal distributions. The last step for the VMI estimation of the divergences is to optimize over the parameters of the neural network which is done by incorporating it as a regularization term in the FastSpeech 2 loss, controlled by a scaling parameter λ_{VMI}:

$$L_{total} = L_{FastSpeech2} + \lambda_{VMI} L_{VMI} \tag{3}$$

CCR is defined for any α, but $\alpha = 1$ and $\lambda_{VMI} = 0.1$ were used in all experiments to reduce variance and ensure stable training. Here, L_{VMI} denotes the estimated mutual information loss based on the selected divergence (e.g., CCR or WCR given by (1) and (2), respectively).

D. Gradient Reversal Layer

The **GRL** method is a supervised approach employed to enforce speaker and style disentanglement by adversarially training the embedding layer. Two feed-forward classifiers are trained offline to optimize the model's speech synthesis capabilities: one for speaker identity and another for speaking style. Each classifier comprises a sequence of layers starting with a Linear layer followed by a ReLU activation, repeated three times, and culminating in another Linear layer. Both classifiers are optimized using the standard cross-entropy loss, given by (4), to effectively learn to solve the classification tasks.

$$L_{\text{CE}}(y,\hat{y}) = -\sum_{c=1}^{C} y_c \log(\hat{y_c}) \tag{4}$$

The GRL enforces disentanglement by reversing gradients from the classification losses L_{speaker} and L_{style} before updating the embedding layer. This forces the model to learn representations that are invariant to speaker and style attributes. The adversarial loss is formulated as:

$$L_{\text{GRL}} = \lambda_{GRL}(L_{\text{speaker}} + L_{\text{style}}) \tag{5}$$

where λ_{GRL} controls the influence of reversed gradients. This joint optimization enhances the model's ability to generate high-quality speech while ensuring robust and disentangled speaker and style representations. In all experiments, $\lambda_{GRL} = 0.1$ was set.

E. Metrics Used

Speaker and style disentanglement effectiveness is evaluated using Cosine Similarity, Cosine Distance, and Average Inter-Cluster Distance.

Given two embedding vectors u_i and u_j, the cosine similarity is computed as:

$$\text{Cosine Similarity} = \frac{u_i \cdot u_j}{\|u_i\|\|u_j\|} \tag{6}$$

This metric measures the angular similarity between two vectors, with values ranging from -1 (opposite) to 1 (identical), where higher values indicate stronger similarity.

The Cosine Distance, used to quantify dissimilarity between embeddings, is defined as:

$$\text{Cosine Distance} = 1 - \text{Cosine Similarity} \tag{7}$$

A higher cosine distance indicates better separation between speaker or style embeddings.

Finally, the Average Inter-Cluster Distance, which quantifies the overall separation between clusters, is defined as:

$$\text{Average Inter-Cluster Distance} = \frac{1}{N}\sum_{i,j,i\neq j} \text{Cosine Distance}(u_i,u_j) \tag{8}$$

where N represents the total number of embedding pairs. Higher values indicate greater separation between clusters, reflecting improved disentanglement of speaker and style representations.

Speech synthesis evaluation was conducted using both objective and subjective methods. Objective metrics included STOI (0-1 scale for intelligibility) and PESQ (0-4.5 scale for quality), computed on 40 randomly selected training sentences. Additionally, a subjective A/B listening test with 100 participants assessed intelligibility and expressiveness across 10 paired tests using high-quality headsets, providing perceptual insights into synthesized speech.

III. Results and Discussion

A. Objective Metrics

Speaker and style disentanglement was evaluated using the Average Inter-Cluster Distance metric. As shown in Table I, the CCR&GRL model achieved the highest distances for speaker (0.9002) and style (0.7616) embeddings, indicating maximum separation of representations. These higher average distances between embeddings signify a more distinct and clear separation between different speakers and styles, underscoring the model's effectiveness in minimizing the mutual information between these disentangled features.

The results demonstrate that the combined application of CCR & GRL techniques outperforms the application of each technique individually. Specifically, the CCR technique stabilizes the disentanglement process by effectively managing the variability inherent within the training data. In contrast, the GRL technique addresses the reduction of mutual dependencies between speaker and style embeddings. It accomplishes this by introducing an adversarial component that promotes feature independence, thereby ensuring that variations in one feature (e.g., speaker identity) do not influence the encoding of another (e.g., speech style). The integration of both techniques thus leads to a more robust and comprehensive disentangled speech representation.

The cosine similarity matrices for the CCR & GRL model, presented in Figure 2, further validate this finding, reporting low values for inter-speaker and inter-style pairs, confirming their independence. Near-zero values in the cosine similarity matrices indicate that the embeddings for various speakers and styles are almost orthogonal, demonstrating clear differentiation when comparing distinct speakers and styles separately. This significant separation highlights the model's proficiency in disentangling and preserving the unique characteristics of each speaker's identity and style.

Speech intelligibility and quality were evaluated using STOI and PESQ metrics. As shown in Table II, the CCR & GRL model achieved the highest STOI (0.3011) and PESQ (1.1576) scores, indicating superior performance. This improved performance can be attributed to the effective separation and independence achieved in speaker and style embeddings within the model. Such distinct disentanglement not only enhances the clarity and naturalness of the synthesized speech but also significantly boosts its intelligibility and quality, as quantitatively validated by these evaluations.

TABLE I
Average Inter-Cluster Distance for Speaker and Style Embeddings

Model	Speaker Embeddings	Style Embeddings
CCR&GRL	**0.9002**	**0.7616**
CCR	0.8754	0.7566
GRL	0.8749	0.7316
WCR	0.8484	0.6950
CLUB	0.8444	0.6911
MINE	0.7528	0.7200
InfoNCE	0.6576	0.6266

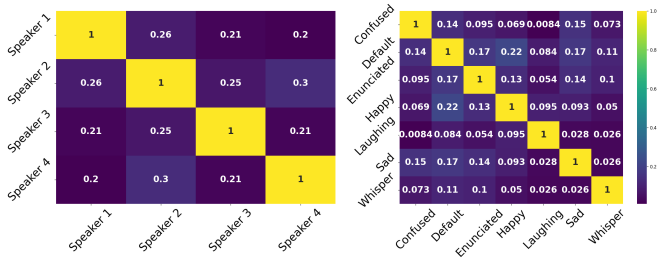

Fig. 2. Cosine similarity matrices for Speaker and Style embeddings.

TABLE II
STOI and PESQ Scores: Mean and Standard Deviation

Model	STOI		PESQ	
	Mean	Std Dev	Mean	Std Dev
CCR&GRL	**0.3011**	**0.1202**	**1.1576**	**0.2748**
CCR	0.2537	0.1771	1.1162	0.2117
WCR	0.2379	0.1661	1.0837	0.2009
GRL	0.2272	0.0967	1.0907	0.2995
MINE	0.2163	0.1207	1.0722	0.1905
InfoNCE	0.2147	0.1202	1.0795	0.1870
CLUB	0.1864	0.0924	1.1114	0.2559

B. Subjective Evaluation

Ten A/B tests evaluated on different pairs of model configurations and focused on various speech styles and intelligibility were conducted. Table III summarizes the results, showing the number of participants who preferred each sample. To ensure diverse and representative feedback, participants were selected across different age groups and genders. The gender distribution consisted of 40% female and 60% male listeners, while the age distribution was as follows: 30% aged 18–25, 40% aged 26–35, 20% aged 36–50, and 10% aged 51–60.

The CCR&GRL model consistently demonstrated superior performance, particularly in conveying intelligibility, sadness, and confused speech styles, highlighting its effectiveness in speaker-style disentanglement. The results from the A/B tests highlight the CCR&GRL model's ability to effectively adapt to nuanced emotional tones and complex linguistic patterns, crucial for realistic speech synthesis. Its strong performance in handling emotions like sadness and confusion demonstrates

the technical and emotional effectiveness of its disentanglement techniques. The significant user preference for the model trained with both CCR and GRL, as demonstrated by the tests, suggests that the CCR&GRL model could achieve greater adoption in practical applications. Audio samples generated by the proposed models can be found at https://stoma.iacm.forth.gr/eusipco2025.html.

TABLE III
A/B TESTING RESULTS FOR SPEECH SYNTHESIS EVALUATION

Test	Type	Sample A	Count	Sample B	Count
1	Intelligibility	**CCR&GRL**	**84**	CLUB	16
2	Confused	**CCR&GRL**	**87**	WCR	13
3	Laughing	MINE	16	**CCR&GRL**	**84**
4	Whisper	InfoNCE	15	**CCR&GRL**	**85**
5	Sad	**CCR&GRL**	**89**	InfoNCE	11
6	Confused	InfoNCE	12	**CCR&GRL**	**88**
7	Happy	**CCR&GRL**	**71**	CCR	29
8	Sad	**CCR&GRL**	**87**	CLUB	13
9	Intelligibility	CLUB	7	**CCR&GRL**	**93**
10	Sad	GRL	15	**CCR&GRL**	**85**

IV. CONCLUSIONS AND FUTURE WORK

This work introduced improved disentangled speech representation methods within the FastSpeech 2 framework, combining advanced VMI estimators and GRL. The model improves clarity, naturalness, and personalization of synthetic speech, as validated by objective metrics (PESQ, STOI, inter-cluster distance) and subjective A/B testing with diverse listeners.

Looking forward, the focus will shift towards scaling the model to handle larger and more varied datasets to ensure greater robustness against diverse linguistic and acoustic environments. This scale-up is essential for addressing more complex speech synthesis challenges. Extending towards zero-shot learning techniques will enable to effectively handle unseen data without the need for retraining. Furthermore, advancing the development and integration of VMI estimators is anticipated to enhance the precision of attribute disentanglement even further. By refining these estimators, the model will be able to better separate and control individual speech characteristics such as tone, style, and emotional content. Extending these refined methods beyond the realm of speech synthesis into multimodal systems presents another important research direction.

V. ACKNOWLEDGEMENTS

This work was partially funded by the Hellenic Foundation for Research and Innovation (HFRI) through the "Second Call for HFRI Research Projects to support Faculty Members and Researchers" under Project 4753.

REFERENCES

[1] Soumya Priyadarsini Panda and Ajit Kumar Nayak. A waveform concatenation technique for text-to-speech synthesis. International Journal of Speech Technology, 20(4):959–976, December 2017. ISSN 1572-8110. doi: 10.1007/s10772-017-9463-8. URL https://doi.org/10.1007/s10772-017-9463-8.

[2] Jerneja Zganec Gros and Mario Zganec. An Efficient Unit-selection Method for Concatenative Text-to-speech Synthesis Systems. Journal of Computing and Information Technology, 16(1):69, 2008. ISSN 1330-1136, 1846-3908. doi: 10.2498/cit.1001049. URL http://cit.srce.unizg.hr/index.php/CIT/article/view/1660.

[3] Pablo Perez Zarazaga, Zofia Malisz, Gustav Eje Henter, and Lauri Juvela. Speaker-independent neural formant synthesis, June 2023. URL http://arxiv.org/abs/2306.01957.

[4] R. J. Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, and Rif A. Saurous. Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron, March 2018. URL http://arxiv.org/abs/1803.09047.

[5] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals,Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio, September 2016. URL http://arxiv.org/abs/1609.03499.

[6] Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE, April 2018. URL http://arxiv.org/abs/1804.03599.

[7] Wei-Ning Hsu, Yu Zhang, Ron J. Weiss, Yu-An Chung, Yuxuan Wang, Yonghui Wu, and James Glass. Disentangling Correlated Speaker and Noise for Speech Synthesis via Data Augmentation and Adversarial Factorization. In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5901–5905, February 2019. doi: 10.1109/ICASSP.2019.8683561. URL https://ieeexplore.ieee.org/abstract/document/8683561.

[8] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, Pieter Abbeel. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets, June 2016. URL http://arxiv.org/abs/1606.03657.

[9] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks, March 2019. URL http://arxiv.org/abs/1812.04948. arXiv:1812.04948 [cs, stat].

[10] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual Information Neural Estimation. In Proceedings of the 35th International Conference on Machine Learning, pages 531–540. PMLR, July 2018. URL https://proceedings.mlr.press/v80/ belghazi18a.html.

[11] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding, January 2019. URL http://arxiv.org/abs/1807.03748.

[12] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information, July 2020. URL http://arxiv.org/abs/2006.12013.

[13] Jeremiah Birrell, Yannis Pantazis, Paul Dupuis, Markos A. Katsoulakis, and Luc Rey-Bellet. Function-space regularized Rényi divergences, February 2023. URL http://arxiv.org/abs/2210.04974.

[14] Jeremiah Birrell, Paul Dupuis, Markos A. Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet. (f, Γ)-Divergences: Interpolating between f-Divergences and Integral Probability Metrics. Journal of Machine Learning Research, 23(39):1-70, 2022. URL http://jmlr.org/papers/v23/21-0100.html.

[15] Yaroslav Ganin and Victor Lempitsky. Unsupervised Domain Adaptation by Backpropagation, February 2015.

[16] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech, August 2022. URL http://arxiv.org/abs/2006.04558. arXiv:2006.04558 [cs, eess] version: 8

[17] Tu Anh Nguyen, Wei-Ning Hsu, Antony D'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, Felix Kreuk, Yossi Adi, and Emmanuel Dupoux. EXPRESSO: A Benchmark and Analysis of Discrete Expressive Speech Resynthesis, August 2023. URL http://arxiv.org/abs/2308.05725. arXiv:2308.05725 [cs, eess].

[18] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved Training of Wasserstein GANs. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), pages 5769–5779, Red Hook, NY, USA, 2017. Curran Associates Inc.