INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

An extension of disentanglement metrics and its application to voice

Olivier Zhang[1,2], *Olivier Le Blouch*[1], *Nicolas Gengembre*[1], *Damien Lolive*[2]

[1]Orange Innovation, France
[2]Univ Rennes, CNRS, IRISA, Lannion, France

{olivier.zhang, olivier.leblouch, nicolas.gengembre}@orange.com, damien.lolive@irisa.fr

Abstract

In representation learning, the promise of disentanglement methods is to decompose an input signal into a set of independent and interpretable attributes. Some metrics, such as the DCI or MIG scores, have been proposed to evaluate how much this goal is reached. They analyse the relationship between the representation components and the desirable attributes. This paper shows that, even when applied to synthetic datasets generated from a closed list of generative factors, these metrics can be too optimistic. In particular, it reports that a generative factor can be recovered from an altered disentangled representation from which it has been supposedly removed, according to the metrics. Based on this observation, a new criterion called latent decimation is proposed to evaluate disentanglement through the accuracy of factors prediction from subsets of latents. A new metric called MIDCI is defined, and its relevance is demonstrated on voice data.

Index Terms: voice analysis, disentanglement, representation learning

1. Introduction

In a world where Transformer-based technologies [1] can create images from text prompts and answer our questions, one could imagine modeling a voice from scratch by explicitly defining its characteristics, such as states (e.g., affect, intoxication, sleepiness), traits (e.g., age, gender, accent, vocal persona [2]) and timber attributes (e.g., roughness, breathiness, nasality). Whereas we know how a human produces speech, and despite protocols to assess voice quality [3, 4] and perception [5], defining a set of explicit voice attributes is still an open issue, as there is neither an exhaustive taxonomy nor methods to disentangle or control them in voice generation. Being able to decompose a voice signal into a closed and exhaustive set of attributes can be seen as the holy grail for many applications in the fields of voice analysis, conversion or synthesis.

Within speech research fields, *disentanglement* still remains an ambiguous term. In text-to-speech (TTS) or voice conversion (VC) literature, *speech disentanglement* mainly refers to the discrimination of two sides of a speech signal: "what is said" (speech units, phonemes, words) versus "who is speaking", to control speech synthesis. Speaker embeddings [6] are widely used to control speaker identity [7]. The speaker side may be further split to handle other attributes, as prosody, style or emotion [8, 9, 10]. Zero-shot VC or TTS have also been impressively improved, with very natural speaker, prosody and style transfer [11]. Despite the very promising results, they still do not allow a fine-grained control over voice attributes. For style transfer, Global Style Tokens (GST) [12] were a first step in this direction, as styles are controllable by tokens automatically learned in an unsupervised way.

Often misregarded by speech literature, disentanglement is a full-fledged research field, where disentangled representation learning aims to align salient factors of variation within data to individual components of representations [13]. Such representations also provide interpretability and controllable data generation when generative models are involved. Unsupervised learning of disentangled representations goes further, by letting models discover independent variations directly from the raw data. In the literature, a commonly used definition of a disentangled representation states that each of its components (also called latents or codes) should be sensitive to changes of only one of the factors of variation [14].

Popular models for unsupervised disentanglement learning are usually based on Variational Autoencoders (VAE) [15]. Many contributions have tricked and extended VAE's loss in order to enforce disentanglement [16, 17, 18, 19]. Following this, a more subtle control of speech generation has been proposed via conditional generative models based on VAE and Tacotron [20, 21]. But there is still no real interpretability of explicit voice attributes except for prosody. Beyond phonemes and prosody, disentangled speaker embeddings conveying explicit speaker traits, states and timbral attributes would upgrade voice generation to a next level. And to do so, we have to discover what are the *generative factors* deeply hidden in speech data. Disentanglement learning is a promising paradigm to learn representations which can automatically capture new voice attributes, so far hard to annotate or even define, thus inaccessible for conventional approaches. Another challenge arising from this goal is how to characterize which attribute is captured, and if so by which latent. The described purposes are hopefully leading to speech synthesis systems with effective "control knobs" to tune voices at will, to mimic existing voices or to create fresh new ones.

Measuring disentanglement is still an open problem. Depending on the adopted definition of "what is a disentangled representation", a wide spectrum of metrics have been proposed and challenged [22, 23]. Furthermore, most related studies handle image disentanglement, and conveniently use synthetic image datasets [24, 25] , as true factors of variations are known, which is required to assess disentanglement. Concerning speech, the toy dataset *diSpeech* [26] is for now the only available analogous dataset. It synthesizes fake vowels from fundamental frequency and formants values. Our study is basically tied to such a synthetic corpus, as using a realistic voice dataset implies relying on annotated attributes (and not factors of variations) which may not be exhaustive or well-balanced enough to assess metrics reliability.

To this end, we believe that the existing disentanglement metrics remain too high-level to truly disclose hidden disen-

10.21437/Interspeech.2023-383

tanglement related behaviors. Therefore, we propose a deeper analysis, based on DCI [27], to boost our interpretation of latent / factor relations. We also propose a *latent decimation* process for disentanglement analysis. Applied to *diSpeech*, it reveals misleading outcomes of the existing metrics in some situations. These observations finally bring us to propose an alternative way to compute DCI, also inspired by MIG [19], ending up to Mutual Information based DCI (MIDCI). Experiments are only presented on diSpeech, but it is worth to note that same results were obtained and assessed on the various synthetic datasets and with the a wide range of models implemented in `disentanglement_lib`.

This paper describes the main existing disentanglement metrics in Section 2. Their skills and weaknesses are compared in Section 3, and the *latent decimation* process is explained and illustrated. Section 4 describes the proposed extended metric, and Section 5 concludes.

2. Related works

Hereafter we provide an overview of the related works about voice attributes factorization and disentanglement learning.

2.1. Voice attributes factorization

Speech disentanglement mainly refers to voice attributes *factorization* in distinct parts of a model. AutoVC [7] disentangles speaker information from the remaining information for voice conversion. SpeechSplit2.0 [9] automatically decomposes speech into content, rhythm, pitch and timbre. FastSpeech2 [8] enables a control of speech generation via pitch, intensity, and phoneme duration. However, voice cloning is usually focused on the source and target speaker signatures, whereas the goal of speech disentanglement is to structure the whole speakers acoustic space.

In this direction, variational approaches such as GMVAE [21], FHVAE [28] or Capacitron [20] propose implicit embedding decomposition and conditional dependencies to improve naturalness of voice cloning or style transfer. TacoSpawn [29] is probably the closest related works to our long-term goal: it learns a distribution over a speaker embedding space and enables sampling of fresh new speakers.

More recently, VALL-E [11] significantly outperformed the state-of-the-art zero-shot TTS system in terms of naturalness, speaker similarity, emotion and acoustic environment preservation in synthesis. Although very accurate and impressive, it does not provide the freedom to explicitly control voice attributes. Even with 60k hours of data for training, it still cannot cover everyone's voice, especially accents.

2.2. Disentanglement learning

Emerging from the postulate that learned representations are still not able to properly organize discriminative information from data [13], disentanglement learning aims to extract representations which identify and separate the underlying explanatory factors of observed data in a latent space. This space is continuous (i.e. close data points should be close in representation space) and complete (i.e. interpolation between representations and sampled latents should stay informative and consistent). VAE fulfills this desiderata, as it approximates data distribution, and enforces independence between individual components. Numerous VAE extensions have been shown to effectively disentangle factors of variation in synthetic image datasets. β-VAE [16] has first emphasized the reconstruction / disentanglement trade-off, adjusted with the weight β applied to the Kullback-Leibler divergence (D_{KL}). CCI-VAE [18] highlights the information bottleneck upper-bounded by the D_{KL}, and regulates the capacity of the latent space with a constant C. Factor-VAE [30] and β-TCVAE [19] decompose the D_{KL} term to point out and penalize the total correlation between latents.

Disentanglement learning studies require tailored datasets in which generative factors of variations to disentangle are salient and known. dSprites [24] and Cars3D [25] are among the most commonly used playground to test models. diSpeech [26] is a dataset of synthetic vowels, with 5 factors of variation: the first 3 formants, F1, F2, F3, the pitch and the pitch fade rate. We will rely on this corpus, with the same settings as in [26] (15 values per factor), to study speech disentanglement and metrics.

As there is still no consensual definition of a disentangled representation, a large amount of metrics have been proposed. In Carobonneau et al.'s work [22], they are reviewed and classified in 3 categories. Intervention-based metrics set a subset of factors and sample the remaining ones, to evaluate the correlations of the chosen factors with latents (e.g., Z-diff [16]). Predictor-based metrics (e.g., DCI [27]) train regressors or classifiers to predict factor values from latents. DCI is actually composed of 3 scores: *Disentanglement* i.e. how much each latent is important to predict one factor, *Completeness* i.e. how much each factor is predicted by only one latent, *Informativeness* i.e. the factor predictions accuracy. Finally, information-based metrics rely on information theory to estimate the localization and amount of factor information in the latent space. For instance, MIG [19] computes the mutual information (MI) between each pair latent / factor and relies on gaps to assess if for each factor, information is concentrated in one latent.

Nevertheless, we believe that these metrics are too high-level for our needs: they focus on grading each model with a unique and global score, which prevents a detailed per-factor analysis. Hence, we propose an in-depth interpretation of the DCI score, which yields a better comprehension of the relations between latents and factors, through the visualization of the importance matrix (as [27, Fig 3]) together with the *factor-wise* completeness and *latent-wise* disentanglement and informativeness, which are pragmatic indicators of these relations.

Furthermore, it turns out that in some cases, predicting factors when removing the most informative latents with respect to the importance matrix, while factor's completeness is actually high, is still achievable with a high accuracy. This routine, which we called *latent decimation*, brings us to question the DCI reliability, similarly to Locatello et al. [23]. Based on similar observations, Eastwood et al. [31] have proposed DCI-ES to decorrelate the DCI scores from the prediction algorithm they depend on. The proposed *latent decimation* can be seen as a complementary sanity check of DCI's indications, whatever the prediction algorithm.

3. Metrics analysis

The review of disentanglement metrics proposed by Carbonneau et al. [22] lists a range of metrics based on different approaches and assumptions. Even for synthetic datasets with a limited number of generative factors and latents, the disentanglement measure may vary significantly from one metric to another, as shown by Locatello et al. [23, Fig 2], Carbonneau et al. [22, Fig 3] and in Figure 1a. Obviously, it makes it difficult to choose an appropriate metric, and it thus appears useful to compare their skills and the approximations they rely on, so as to emphasize their advantages and drawbacks. This is de-

veloped in Subsection 3.1, with a focus on the DCI, MIG and Z-diff metrics. We then describe in Subsection 3.2 disentanglement evaluation on diSpeech augmented with in-depth analysis of metrics. Subsection 3.3 illustrates how the metrics sometimes fail to properly evaluate the disentanglement.

3.1. Comparison of the existing metrics

One major advantage of the DCI metrics is that it provides three indicators, that measure 3 different aspects of the disentanglement (see Section 2). Also, the DCI metrics are computed thanks to an importance matrix in which each component represents the relationship between the latents and the generative factors (see Figure 1c). This is useful as it allows a per-factor analysis instead of a global score (see Figure 1b). Indeed, as a general rule, some factors can be well disentangled while others are not, due to the structure of the data, the nature of the factor, or its impact on the data generation.

On the other hand, the components of the importance matrix (the importance weights), are deduced from the parameters of a regressor (or classifier when categorical factors are concerned) trained to predict the factors knowing the latents. Although they are clearly influenced by the information about each factor contained in each latent, which is relevant for the metrics, these amounts can be altered by the kind of regressor used, the implementation, the assumed relationship between latents and factors (is it linear or not?) and so on.

The information-based approaches such as the MIG score do not suffer these drawbacks, as they rely on the computation of the MI between factors and latents, which is often used as a generalized correlation coefficient [32]. But still, there are algorithmic parameters to be chosen. In addition, correlations between latents and between factors are ignored. Also, the metric relies on gaps between the most and second most important MI for each factor, favoring information to be located in a single latent for each factor, and disadvantaging cases where a factor might needs 2 latents to be perfectly captured. In addition, it totally misses out the Disentanglement part of DCI as the latents capturing multiple factors are not penalized [22].

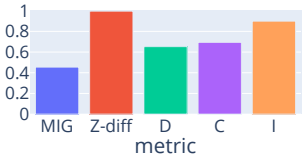
The Z-diff metric and its variants also uses a prediction algorithm to provide its outcome, but through a low-complexity linear classifier, by design. Thus the score is less dependent on tunable parameters. Nevertheless, its principle consists in finding the most correlated latent to a given factor, ignoring possible correlations to other latents, what often makes its disentanglement evaluation too optimistic.

Following [22, Tab 2], DCI is the metric that covers the most characteristics. Pragmatically, it is indeed convenient to have a precise idea of latent / factor relationships, factor-wise Completeness and latent-wise Disentanglement. MIG has the advantage to not be influenced by predictor intricacies, but has a too restrictive assumption of disentanglement by using MI gaps.

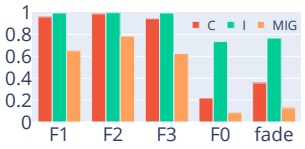
3.2. diSpeech disentanglement

Thanks to the `disentanglement_lib` [23] library, we have experimented a broad range of disentangling models on diSpeech corpus. In this article, we retained results of β-TCVAE trained with 8 latent dimensions, as it reached the best performances, but similar observations where made with other models and datasets (see Section 2). Z-diff, MIG and DCI[1] analyses are presented in Figure 1. Metrics values are reported in Figure 1a.

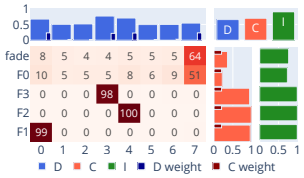
[1] implemented with XGBoost library, for faster computation



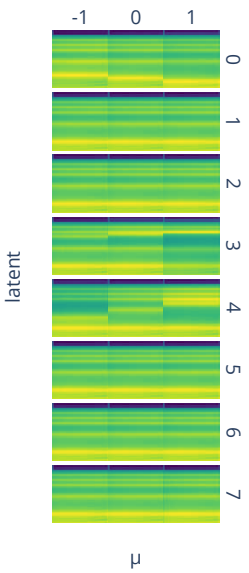
(a) *MIG, Z-diff, and DCI scores*



(b) *Factor-wise Completeness, Informativeness, and MIG*



(c) *DCI with importance matrix and in-depth measures*



(d) *diSpeech traversals*

Figure 1: *diSpeech disentanglement evaluation*

Z-diff suggests a really good disentanglement, while other metrics are more mitigated, especially MIG.

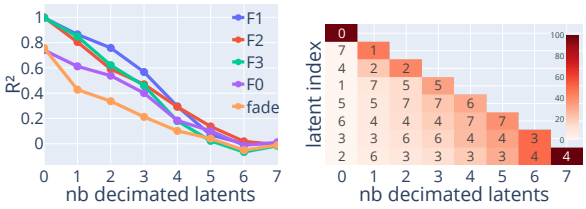
But these global measures keep hidden the disentanglement of each factor. Hence, we report in Figure 1b MIG, Completeness and Informativeness for each factor, showing that performances highly depend on the considered factor. Formants (F1,F2,F3) seem well disentangled, while pitch (F0) and fade have poor MIG, Completeness and Informativeness.

A closer look at the DCI importance matrix in Figure 1c indicates which latent disentangles each formant : F1: latent 0, F2: latent 4, and F3: latent 3. Cell values are the percentage of importance (feature importance $\times$ 100). Figure 1c also aligns the importance matrix with entropy-based factor-wise Completeness (right part) and latent-wise Disentanglement (up part), and factor-wise Informativeness. Latent and factor variable importance weights (ρs in [27]) are also reported next to their respective values (thin dark bars). Figure 1c is thus an informative yet condensed view of factor / latent relations. It is also suggested by traversals in Figure 1d, where the corresponding formants are clearly moving in dimensions 0, 4 and 3, respectively, and only in them.

3.3. Sanity check via latent decimation

In order to figure out if metric outcomes correctly reflect the disentanglement properties of a latent representation, we conducted experiments based on what we call *latent decimation*. The idea is to remove the most informative latents with respect to a given factor, and measure how much of its information has been lost. This loss is evaluated thanks to a predictor (same as DCI), trained to predict the factor from the remaining latents, and the accuracy drop is used to measure the information loss. Thus, if a factor is well disentangled, removing its most important latent should results in a drastic drop of accuracy.

The *latent decimation* performed with the model described in Subsection 3.2 is depicted in Figure 2. For each factor, the most important latent (with respect to DCI importance matrix) is removed to rerun prediction. Then, the latents importance is



(a) R^2 *scores* (b) *Latent importances for F1*

Figure 2: *diSpeech latent decimation*

deduced again, and the new most important latent is removed. This process is repeated until 1 latent is left. The R^2 scores of each iteration and factor are plotted in Figure 2a. Counter-intuitively from results in Subsection 3.2, factors are still predictable with a decent accuracy, meaning that factors' information is not only contained in the most important latents, and not that well disentangled as suggested by DCI.

At each decimation step, we can keep track of latents importance order to assess consistence along iterations. The ordered latents at each decimation for F1 are logged in Figure 2b: in each column, latent index are stacked in a importance ascending order, and the color scale reflects the importance value. It appears that importance order is not consistent: latent 7 is the second most important latent at the beginning, but is reported the most important only 5 steps further. Similar inconsistent behavior can be observed, with latent 4 and 2 for instance. These changes underline that information about factors can be spreaded in other latents, but neither used by predictors (for DCI computation) nor decoders (for transversal generation).

As pointed out, good disentanglement of formants deduced by DCI is compromised by *latent decimation* sanity-check. This is following Locatello et al.'s [23] conclusion on the importance of the assessment of the practical benefit of disentanglement. Biases induced by predictors lead to a misleading DCI scoring. It can be overcome by using an importance matrix based on MI.

4. Metric extension

DCI appears in the literature and the experiments in Section 3 as a useful metric to disclose factor / latent relations. We also showed that DCI assessments can be contradicted by the *latent decimation* procedure. Hence, we coined the MIDCI metric, which is detailed hereafter. Its accordance to *latent decimation* is then demonstrated.

4.1. MIDCI

In order to overcome predictor biases in DCI, we propose to compute importance matrix based on MI as done in MIG and deduce Disentanglement and Completeness as in DCI. Let $i \in \{1, \dots, F\}$ and $j \in \{1, \dots, L\}$, F the number of factors and L the number of latents. The MI matrix is defined as

$$R_{i,j} = \frac{I(f_i; l_j)}{H(f_i)}, \tag{1}$$

with $I(f; l)$ the MI between factor f and latent l. MI is divided by $H(f)$, f's entropy, so that $R_{i,j} \in [0, 1]$. Straightforwardly, Disentanglement and Completeness are defined as Eastwood et al. [27], by using entropy along latents and factors respectively.

Note that $S = \sum_{j=1}^{L} R_{i,j}$ does not necessarily equal to 1, as $R_{i,j}$ embodies f_i's rate of information captured by l_j which can be incomplete ($S \leq 1$) or redundant ($S \geq 1$), due to "cross-information" shared with other latents.

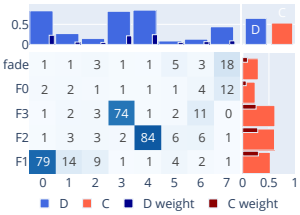


Figure 3: *MIDCI with MI matrix, Completeness and Disentanglement*

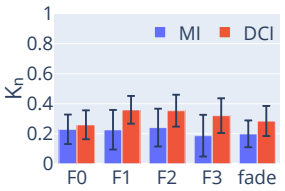


Figure 4: K_n *between MI (blue) / DCI importance (red) and latent decimation rankings*

We can also define an information-theory-based formulation of the factor-wise Informativeness Info_i, as:

$$\text{Info}_i = 1 - \frac{H(f_i|l_1, \dots, l_L)}{H(f_i)}. \tag{2}$$

Nevertheless in practice, the great number of data and a possibly important number of latents and factors result in a multivariate distribution, which makes the computation of Info_i a complex challenge. This definition of MIDCI takes benefits from both DCI and MIG: MI based importance matrix overcomes predictors biases, and latent-wise Disentanglement / factor-wise Completeness provide in-depth insights of latent / factor relationships.

4.2. diSpeech MIDCI

Coming back to diSpeech disentanglement, applying MIDCI is equivalent to replace importance matrix 1c with MI matrix, resulting in Figure 3. In conformity with *latent decimation*, Completeness appears less optimistic.

In order to assess if MIDCI is closer than DCI to *latent decimation* latents ordering, we used Normalized Kendall τ distance (K_n) [33]. Figure 4 shows that for several models (described in Subsection 2.2), with several number of latents (8, 16, 32), K_n is, for each latent, in average smaller with MIDCI than with DCI. Hence, a better accordance is achieved when using MI, demonstrating an improved reliability.

We have extended the described experiments on visual synthetic datasets, as those mentioned in Section 2. Results are overall similar to what is observed with diSpeech, but we noticed some fluctuations for complex factors (rotation, azimuth in visual datasets), where DCI has a better K_n than MIDCI.

5. Conclusion

In this study, we proposed a first step to bridge the gap between two worlds: disentanglement theory, generally associated to image analysis, and voice analysis/generation where disentanglement terminology is mainly restricted to phonetic content vs speaker identity discrimination, without making any use of standard disentanglement metrics. Based on a factor-wise approach deriving from a long-term goal of controlling voice generation via explicit voice attributes, we highlighted the behavior of dedicated metrics on a toy dataset of synthesized speech. We extracted hidden biases of the DCI and proposed an analysis grid coupled to a *latent decimation*-based sanity check and a more reliable version of DCI relying on mutual information, MIDCI. Although explicited on a single dataset of synthesized speech, the trends are also observed on standard disentanglement datasets, and we believe this study serves as a preliminary research work, which can be beneficial to further investigations in the field of voice generation.

6. References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] C. Noufi, L. May, and J. Berger, "The role of vocal persona in natural and synthesized speech," in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, 2023, pp. 1–4.

[3] E. San Segundo and J. A. Mompean, "A simplified vocal profile analysis protocol for the assessment of voice quality and speaker similarity," *Journal of Voice*, vol. 31, no. 5, pp. 644–e11, 2017.

[4] J. Laver, S. Wirz, J. Mackenzie, and S. Hiller, "A perceptual protocol for the analysis of vocal profiles," *Edinburgh University Department of Linguistics Work in Progress*, vol. 14, pp. 139–155, 1981.

[5] R. I. Zraick, G. B. Kempster, N. P. Connor, S. Thibeault, B. K. Klaben, Z. Bursac, C. R. Thrush, and L. E. Glaze, "Establishing validity of the consensus auditory-perceptual evaluation of voice (cape-v)," *American journal of speech-language pathology*, 2011.

[6] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 165–170.

[7] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5210–5219.

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

[9] C. H. Chan, K. Qian, Y. Zhang, and M. Hasegawa-Johnson, "Speechsplit2. 0: Unsupervised speech disentanglement for voice conversion without tuning autoencoder bottlenecks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6332–6336.

[10] X. An, F. K. Soong, and L. Xie, "Disentangling style and speaker attributes for tts style transfer," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 646–658, 2022.

[11] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[12] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[13] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[14] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner, "Towards a definition of disentangled representations," *arXiv preprint arXiv:1812.02230*, 2018.

[15] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2013.

[16] 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 *International conference on learning representations*, 2017.

[17] R. T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud, "Isolating Sources of Disentanglement in Variational Autoencoders," in *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/1ee3dfcd8a0645a25a35977997223d22-Abstract.html

[18] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-vae," *arXiv preprint arXiv:1804.03599*, 2018.

[19] R. T. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," *Advances in neural information processing systems*, vol. 31, 2018.

[20] E. Battenberg, S. Mariooryad, D. Stanton, R. Skerry-Ryan, M. Shannon, D. Kao, and T. Bagby, "Effective use of variational embedding capacity in expressive end-to-end speech synthesis," *arXiv preprint arXiv:1906.03402*, 2019.

[21] W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen *et al.*, "Hierarchical generative modeling for controllable speech synthesis," *ICLR*, 2019.

[22] M.-A. Carbonneau, J. Zaïdi, J. Boilard, and G. Gagnon, "Measuring disentanglement: A review of metrics," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2022.

[23] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *international conference on machine learning*. PMLR, 2019, pp. 4114–4124.

[24] L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner, "dsprites: Disentanglement testing sprites dataset," 2017.

[25] S. Reed, Y. Zhang, Y. Zhang, and H. Lee, "Deep visual analogy-making," in *NIPS*, 2015.

[26] O. Zhang, N. Gengembre, O. Le Blouch, and D. Lolive, "diSpeech : A Synthetic Toy Dataset for Speech Disentangling," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

[27] C. Eastwood and C. K. Williams, "A framework for the quantitative evaluation of disentangled representations," in *International Conference on Learning Representations*, 2018.

[28] W.-N. Hsu and J. Glass, "Scalable factorized hierarchical variational autoencoder training," *arXiv preprint arXiv:1804.03201*, 2018.

[29] D. Stanton, M. Shannon, S. Mariooryad, R. Skerry-Ryan, E. Battenberg, T. Bagby, and D. Kao, "Speaker generation," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7897–7901.

[30] H. Kim and A. Mnih, "Disentangling by factorising," *ArXiv*, vol. abs/1802.05983, 2018.

[31] C. Eastwood, A. L. Nicolicioiu, J. von Kügelgen, A. Kekić, F. Träuble, A. Dittadi, and B. Schölkopf, "Dci-es: An extended disentanglement framework with connections to identifiability," *arXiv preprint arXiv:2210.00364*, 2022.

[32] L. Song, P. Langfelder, and S. Horvath, "Comparison of co-expression measures: mutual information, correlation, and model based indices," *BMC bioinformatics*, vol. 13, no. 1, pp. 1–21, 2012.

[33] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, pp. 81–93, 1938.