DISPEECH: A SYNTHETIC TOY DATASET FOR SPEECH DISENTANGLING

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

[1]Orange Innovation, France
[2]Université de Rennes 1, CNRS, IRISA, Lannion, France
{olivier.zhang, olivier.leblouch, nicolas.gengembre}@orange.com,
damien.lolive@irisa.fr

ABSTRACT

Recently, a growing interest in unsupervised learning of disentangled representations has been observed, with successful applications to both synthetic and real data. In speech processing, such methods have been able to disentangle speakers' attributes from verbal content. To have a better understanding of disentanglement, synthetic data is necessary, as it provides a controllable framework to train models and evaluate disentanglement. Thus, we introduce *diSpeech*, a corpus of speech synthesized with the Klatt synthesizer. Its first version is constrained to vowels synthesized with 5 generative factors relying on pitch and formants. Experiments show the ability of variational autoencoders to disentangle these generative factors and assess the reliability of disentanglement metrics. In addition to provide a support to benchmark speech disentanglement methods, *diSpeech* also enables the objective evaluation of disentanglement on real speech, which is to our knowledge unprecedented. To illustrate this methodology, we apply it to TIMIT's isolated vowels.

Index Terms— speech disentanglement, synthetic dataset, unsupervised learning

1. INTRODUCTION

Speech is known to carry a large variety of aspects which can be either correlated (e.g. speaker identity and gender) or independent (e.g. emotion and accent). Deep learning approaches allowed unprecedented performance on a lot of speech processing applications such as Automatic Speech Recognition or Speaker Recognition. These applications are generally focusing on a single aspect of speech, and might be disturbed and misleaded by others. Hence, being able to split the different perceptual aspects of speech should be useful to help underlying tasks to focus only on relevant information.

Unlike regular representation learning, disentangled representation learning aims to align salient factors of variation within data to individual components of representations. It provides interpretable representations, and controllable data generation when manipulating generative models. Unsupervised learning of disentangled representations goes further, by letting representations components automatically capture factors. While there is still no consensus on a clear definition of disentanglement, most studies agree that disentangled representation should have each latent dimension sensitive to changes of only one of the mutually independent factors of variation.

A popular model for unsupervised disentanglement learning is β-VAE [1], a deviation of the Variational Auto-Encoder (VAE) [2]. A bunch of β-VAE's deviations have been proposed, as Annealed-VAE [3], FactorVAE [4], or β-TCVAE [5]. We focus on β-VAE, as the purpose of this study is to explore speech disentanglement and not to find the best architecture for this problem.

Most related studies handle image disentanglement [1, 4, 6], which is already a non trivial problem. Synthetic datasets are used to make generative factors known and controllable. We noticed the lack of synthetic speech dataset, homologous to those available for image processing, which we believe to be helpful to have a deeper comprehension of speech disentanglement, and would lead to more interpretable and controllable representations of real speech.

However, going from image to speech disentanglement is not straightforward. Dependencies between factors at same granularity level (e.g. value dependencies between formants, F0 affects harmonics) or hierarchical relations between different granularity level (e.g. type of vowel influences formants, gender impacts F0), time dependencies, factors sharing the same physical (or virtual) scale space (e.g. formants), and preprocessing steps are additional difficulties which make speech disentanglement specific to handle. Interdependences between factors even go against the acknowledged definition of disentanglement. Truly independent speech factors of variations are not intuitive to specify, and it is often not trivial to identify which factors are disentangled, especially in an unsupervised approach.

In this paper, we attempt to address the mentioned challenges, by introducing *diSpeech*, a corpus of synthesized vowels, with 5 generative factors: the first 3 formants (F1, F2, F3) to control the vowel identity, F0 to control vowel's pitch, and F0's fade rate to control the final value taken by F0. We also describe experiments, intended to show *diSpeech*'s interests towards speech disentanglement. Hence, we ensure disentangling feasability on our corpus, assess the reliability of disentanglement metrics for model selection, and show that *diSpeech* allows us to quantify disentanglement of real data, which is an open issue in disentanglement studies.

The remaining discussion will follow the following outline: Section 2 presents the related work about disentanglement methods, metrics and datasets. Section 3 describes the proposed dataset and section 4 reports the experiments description and results. We discuss the results in section 5. Perspectives are then depicted in section 6.

2. RELATED WORK

Deep learning has shown to be able to learn unsupervisedly disentangled representations. Such methods are introduced in subsection 2.1. Specifically, speech disentanglement is discussed in subsection 2.2. In order to objectively measure disentanglement, subsection 2.3 describes metrics proposed by the literature. Datasets designed or used in disentanglement studies are depicted in subsection 2.4. Finally, subsection 2.5 references the third-party libraries and tools used for this work.

2.1. Disentanglement models

Most models for unsupervised learning of disentangled representations are based on Variational Auto-Encoder (VAE) [2], mainly thanks to two properties of learned latent space: (1) continuity, continuous traversal in the latent space results in a continuous perceptual variation in reconstructed output, and (2) completeness, every point of the latent space has a perceptually coherent reconstructed output. Those properties are ensured by VAE's objective function, denoted as $\mathcal{L}$, and composed of two terms: a reconstructed input likelihood $\mathcal{L}_R$ (which reflects the "inversed" tendency of reconstruction error), and a Kullback-Leibler divergence D_{KL} constraining latent space's distribution to match a multidimensional centered reduced normal distribution, with a diagonal covariance matrix.

In a β-VAE [1], the Kullback-Leibler divergence is further penalized with an additional hyperparameter β. Hence, latent space's axes are more encouraged to be conditionally independent, and hopefully, they might align with independent factors of variation within the input data. The general formulation of a β-VAE is schematically expressed by:

$$\mathcal{L} = \mathcal{L}_R - \beta D_{KL}, \tag{1}$$

and corresponds to the standard VAE objective function when $\beta = 1$. Mathematical details can be found in [1, 2].

By changing the Kullback-Leibler divergence term, a lot of models were proposed, to further encourage disentanglement and gain a better trade-off with reconstruction error. Among them, we can mention AnnealedVAE [3], FactorVAE [4] and β-TCVAE [5]. Furthermore, the VQ-VAE [7] approach extends the formalism to discrete latent dimensions, to deal with discrete factors disentanglement. As an alternative to VAEs, InfoGAN [8] tackles the problem by means of generative adversarial networks with an additional regularization term in the loss function, in order to maximise the mutual information between the latent components and the generator's output.

2.2. Speech disentanglement

Interest in speech disentanglement is driven by a bunch of promising applications aiming at analyzing, modifying or canceling some of disentangled attributes. To name a few, speech synthesis with controlled styles or emotions is a key element in affective computing (e.g. for virtual assistants), and voice conversion or de-identification can be a solution for applications that require to store audio signals while preserving privacy. Speech disentanglement has been addressed in some recent papers with different methods, mostly based on VAEs variants or adversarial networks, for speaker / content separation [9], style or emotion control [10, 11].

2.3. Disentanglement metrics

Disentanglement quantification is a non trivial problem. Until now, almost all metrics need the knowledge and control on factors of variation, and consequently can only be applied to synthetic datasets.

Based on [12], there is a large choice of metrics with different approaches to measure multiple disentanglement properties. We have tested a large amount of metrics presented in [12], such as SAP score [13], Z-diff (or β-VAE score) [1], FactorVAE [4], Interventional Robustness Score (IRS) [14], Modularity score [15] and MIG score [5]. Based on Zaidi *et al.*'s conclusions and our experiments, we decided to conduct our analysis mostly using DCI [6], which seems to be the most effective metric.

DCI aspires to evaluate three disentanglement properties with three metrics: *Disentanglement* i.e. how much each latent captures only one factor, *Completeness* i.e. how much each factor is captured by only one latent, and *Informativeness* i.e. the amount of information contained in latent representations. DCI uses regression (Lasso or Random Forest) to predict factors from latents. Then latents weights obtained with the regression are aggregated into an importance-matrix, used to determine the three scores.

Among the three metrics provided by DCI, we judged *Disentanglement* to be the most important to watch for model selection, even if *Completeness* exhibits similar behavior during experiments. We check the *Informativeness* to ensure that learned latent space transmits enough information about input data.

2.4. Disentangling datasets

Most studies on disentanglement were focused on image processing applications. Synthetic datasets are also widely used as they provide a controllable framework for disentanglement investigation.

Therefore, several datasets of synthetic images were proposed. Among them, dSprites [16] is a set of binary images of white shapes on a black background, where the shapes' x and y coordinates, rotation, scale and type of shape form the five generative factors. Cars3D [17], SmallNORB [18] and 3D Shapes[19] are other well-used datasets of synthetic 3D scene images where factors such as object type or view are controllable.

To our knowledge, the only analogous dataset conceived for audio signal processing is dMelodies [20], a set of monophonic melodies with nine independent factors as the tonic or the octave, forming a dataset homologous to dSprites but for music applications. However, no such disentanglement dataset was designed for speech. For this reason, we decided to generate our own dataset for speech disentanglement purpose, called *diSpeech*.

2.5. Implementation

In this study, trainings and disentanglement evaluations were mainly performed thanks to *disentanglement_lib* [21], which implements most of the mentioned disentanglement models and metrics, provides a set of tools to launch trainings, evaluations and visualizations. The integration of custom models, metrics and datasets is also intuitive, making the development and the use of our *diSpeech* dataset easier.

3. DISPEECH DATASET

In this section, we describe *diSpeech*[1], the first synthetic speech dataset intended for speech disentanglement experiments. As a first step, we have chosen to constrain our dataset by synthesizing only vowel-like waveforms of one second. We will refer to vowels in the remaining discussion. To do so, we used Klatt Synthesizer [22], a speech synthesizer which provides a complete set of parameters able to generate quite realistic speech. More precisely, we used *tdklatt*[2]. For our purposes, we only needed to tune the following parameters: (1) the first three formants F1, F2 and F3, to cover vowels ; (2) the fundamental frequency F0, to control vowels' pitch ; and (3) a fade factor which determines the final value taken by F0, to approximate realistic pronunciation of vowels, and to have a more complex time dependent factor. Indeed, the F0 value is not constant, and the fade parameter is a percentage ($\in [0, 99]$), defining the proportion of the initial F0 value the vowel will reach at its end. For instance, F0 can

[1] https://github.com/Orange-OpenSource/diSpeech
[2] https://github.com/guestdaniel/tdklatt

	F1	F2	F3	F0	fade
min	275	779	2579	50	0
max	830	2585	3815	200	99

Table 1: Range of values for generative factors.

be constant (fade = 0), or linearly decreasing to 50% of its initial value (fade = 50). To summarize, *diSpeech* is built upon 5 generative factors: F1, F2, F3, F0 and (F0's) fade, which permit to generate a consequent set of vowels, depending on how factors are discretized. Hence, the total number of samples used in our experiments is $15^5 = 759375$.

For our experiments, we discretized each factors into 15 equally spaced values, in predetermined ranges. Based on [23], we determine for each formant the minimum and maximum values reached. Hence, we ensure coherent synthesized vowels and suitable coverage of vocalic phoneme space. Table 1 shows and summarizes those extreme values.

Furthermore, we explicitly defined the five generative factors used for experiments, even if Klatt Synthesizer supports a large set of parameters that can be tuned to generate other vowel variations, and more generally phonemes. Hence, in corpus generation code, there is no constraint on the parameters to tune, enabling any parameter to be considered as a generative factor. It means that *diSpeech* can (and is intended to) be extended and is not limited to vowels.

In order to extract meaningful features from vowel audios, and also to integrate our dataset into *disentanglement_lib*, we processed the synthesized vowels to obtain inputs homologous to dSprites or Cars3D. We used ESPnet [24] to compute from waveforms the log mel-filter banks, with 64 mel filters, a fast Fourier transform of length 1024, and a hop length of 252, resulting in a 64x64 image, perfectly matching dSprites and Cars3D image definition. *diSpeech* is then compatible with *disentanglement_lib*, as it is designed to process images.

4. EXPERIMENTS

Now that we have a new dataset of synthetic vowels, we are able to experiment speech disentanglement. The first stage, described in subsection 4.1, ensures that a β-VAE is able to disentangle *diSpeech*'s factors, and if so, evaluates which factors are successfully disentangled and to which degree. Then, disentanglement metrics are computed to determine the optimal β value.

To go towards real speech disentanglement, we trained a β-VAE on TIMIT's vowels, and thanks to *diSpeech*, we were able to objectively evaluate disentanglement on this realistic dataset, as described in subsection 4.2.

4.1. *diSpeech* disentanglement

On a first stage, we trained a β-VAE on *diSpeech* with $\beta \in \{1, 2, 3, 5, 10, 20\}$, and a latent space of 10 dimensions. In order to select the best value of β, we launched the evaluation metrics provided by *disentanglement_lib*. Metrics depending on the value of β can be visualized in Figure 1.

At first sight, metrics seem not to agree on the disentanglement quality. SAP score and especially MIG badly rate disentanglement, whereas Explicitness and Modularity are giving good scores. Nevertheless, the overall variations of the metrics indicate that for most of them, higher values are reached for $\beta = 2$. Figure 1 also shows that focusing on DCI seem sufficient to analyse results and take a decision on β: *Disentanglement* indicates $\beta = 2$, and *Completeness* and

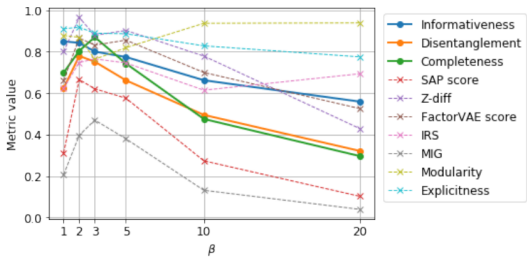


Fig. 1: Metrics on *diSpeech* depending on β.

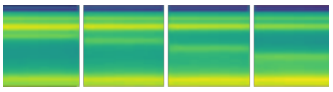


Fig. 2: Four log mel-filter banks showing latent 6 traversal. We can observe F2 moving downwards

informativeness both exhibit higher scores when $\beta \in [1, 5]$. Overall, DCI suggests better disentanglement for small βs and falling performances when β is increasing.

Then, *disentanglement_lib* provides the necessary to easily visualize reconstructions of traversals of learned models' latent space, allowing us to investigate concretely how *diSpeech* is disentangled, specifically with $\beta = 2$. For instance, Figure 2 shows the traversal of the latent 6 in the latent space i.e by traversing the latent space along only one component, we can see the impact of the considered latent on reconstructions. We can clearly see the second formant F2 moving from the top to the bottom of the log mel-filter bank, indicating a good disentanglement of F2. The same can be observed with F1 and F3, in latent 8 and 1 respectively.

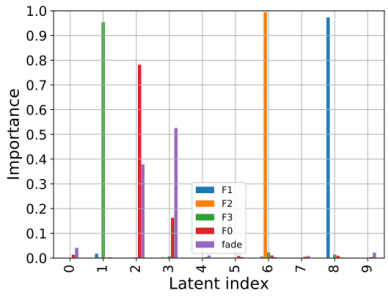
In addition, thanks to the importance matrix computed with DCI, we can directly visualize the relative importance of each latent to predict each factor, which hopefully represents the disentanglement degree. Hence, we can see in Figure 3a for $\beta = 2$ the factor-wise importance of each latent to predict *diSpeech*'s factors. We can notice the latent 6 exhibiting a high importance to predict F2, which corresponds to the traversal shown in Figure 2. It also shows that latents 8 and 1 achieve good disentanglement of F1 and F3, and that F0 and fade are learned but still entangled together in latent 2 and 3.

This first experiment shows that β-VAE successfully achieves disentangling on *diSpeech*. Formants are correctly learned and aligned each on a single axis, as indicated by single peaks for formant factors in Figure 3a. Hence, DCI allows us to objectively identify disentangled (or entangled) factors and corresponding disentangling (or entangling) latents.

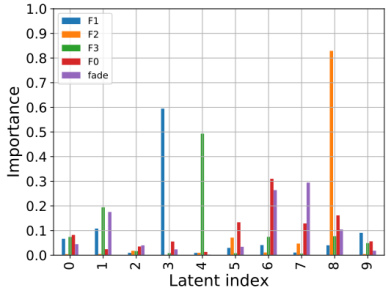
4.2. TIMIT's vowels disentanglement

One main obstacle of real speech disentanglement is the absence of knowledge of generative factors. We propose to use *diSpeech* to compute DCI on real vowels. Hence, we trained a β-VAE on TIMIT's isolated vowels (always with 10 latent dimensions), symmetric-padded to reach 1 second and preprocessed as in *diSpeech* to have equivalent inputs.

By taking the learned β-VAE and encoding *diSpeech*, we are now able to measure disentanglement on TIMIT's vowels, relatively to *diSpeech*'s factors. As we can see in Figure 4, *diSpeech*'s DCI reaches better values than with TIMIT, which is not surprising. There is no guarantee that TIMIT will be unsupervisedly disentangled following *diSpeech*'s factors, and observing TIMIT's latent



(a) *diSpeech*, $\beta = 2$



(b) TIMIT's vowels, $\beta = 2$

Fig. 3: Factor-wise latent importance from DCI

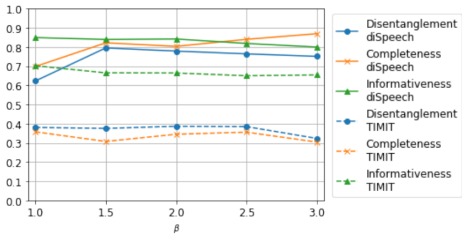


Fig. 4: DCI depending on β for *diSpeech* (solid) and TIMIT's vowels (dashed)

traversals confirms that it is not well disentangled. But the theorical *Disentanglement* score for a totally entangled latent space tends to 0, as confirmed by [12]'s experiments in subsection 5.2. Thus, the non-zero values reached by TIMIT's DCI suggest a partial disentanglement, and we actually observed latents somehow capturing the variations of single formants, mainly F2, while still being a bit entangled with other factors. Figure 3b confirms that formants, especially F2, are partly disentangled even though it is clearly below performances on *diSpeech*.

Even if β-VAE does not disentangle perfectly TIMIT's vowels, relatively to the factors we defined, we showed that *diSpeech* allows us to compute disentanglement scores for a model trained on real speech, and it even gives a coherent score, which is unprecedented. Note that all vowel phonemes of TIMIT's phonetic transcript were used as extracted samples. Hence, diphtong phonemes are also included during models training, increasing training data complexity and divergence with respect to evaluation data.

5. DISCUSSION

As pointed out in previous section, β-VAE is not able to achieve a truly efficient disentanglement on TIMIT's vowels, relatively to performances on *diSpeech* or more generally on other synthetic image datasets described in subsection 2.4.

As mentioned in section 1, speech disentanglement has specific obstacles, due to time dependencies and complex relations between generative factors. Defining a set of perfectly independent factors is already not trivial. On the other hand, assuming independence may lead to exploitable results, as does naive Bayes classifier.

As speech attributes are hard to annotate and subjective, unsupervised disentanglement is a promising approach to automatically extract relevant and interpretable features for tasks with few annotations. But there is no guarantee that models will align with expected factors, or if we don't set expectation, identify disentangled factors is not simple. On top of that, nothing ensure that latent will learn usefull features. It is also noteworthy that nothing prevents factors to be captured by more than one latent (e.g. rotation as angle or sin and cos components), discrediting *Completeness* score, adding more complexity to disentanglement models analysis.

Furthermore, experiments were performed with β-VAE, which may not have the capacity to handle time dependencies and speech related complexities previously mentioned. Leveraging more advanced models as those depicted in subsection 2.1 would lead to better results.

6. PERSPECTIVES

Our experiments show that generative factors of synthetic vowels can be partly disentangled with a β-VAE. But it also appears that transfer to evaluate TIMIT real vowels disentanglement is complicated.

Extending *diSpeech* towards more realistic content, with other phonemes (e.g. consonants), variable durations, combination of phonemes and so on, would hopefully lead to a more reliable disentanglement evaluation on real data, and a wider coverage of speech factors and a fairer approximation of speech complexities.

As the disentanglement process is unsupervised, and hence data driven, the generative factors can be well disentangled or not, depending on the way they appear in the input audio features. Choosing MFCC instead of Log-Mel-spectrograms may have an influence on the disentanglement performance. Similarly, the evaluation of the reconstruction error could be modified to better reflect the similarity between spectrograms; we believe that using MCD [25] could help.

Hierarchical dependencies issues inherent to speech could be addressed by variants of β-VAE such as VQ-VAE, AnnealedVAE, NVAE or novel strategies parallelizing "multi-βs". Note also that a well-defined latent space, by instance inspired by Poincaré embeddings [26], could be helpful.

7. CONCLUSION

In this paper, a new corpus of synthetic phonemes called *diSpeech* has been presented. It has been designed to study disentanglement of voice attributes. Its first declination relies on synthetic vowels, parameterized by the fundamental frequency, the first three formants and the fade rate. It has been used in disentanglement experiments based on a β-VAE model. It results in a clear disentanglement of the formants whereas the other two factors stay partly entangled, emphasizing the influence of the nature of the generative factor on its disentanglement. We believe that *diSpeech* paves the way towards disentanglement evaluation on real speech, as shown by experiments on TIMIT's vowels. Forthcoming studies and improvements of the corpus and the methodology have finally been proposed.

8. REFERENCES

[1] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner, "β-vae: Learning basic visual concepts with a constrained variational framework," *ICLR 2017 - Conference Track Proceedings*, 2016.

[2] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," in *Proceedings of ICLR*, 2014.

[3] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner, "Understanding disentangling in β-vae," *arXiv preprint arXiv:1804.03599*, 2018.

[4] Hyunjik Kim and Andriy Mnih, "Disentangling by factorising," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2649–2658.

[5] Tian Qi Chen, Xuechen Li, Roger Grosse, and David Duvenaud, "Isolating sources of disentanglement in variational autoencoders," in *ICLR 2018 - Workshop Track Proceedings*, 2018.

[6] Cian Eastwood and Christopher K.I. Williams, "A framework for the quantitative evaluation of disentangled representations," in *ICLR 2018 - Conference Track Proceedings*, 2018.

[7] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu, "Neural discrete representation learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6309–6318.

[8] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel, "InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets," *Advances in Neural Information Processing Systems*, pp. 2180–2188, jun 2016.

[9] Jennifer Williams, Yi Zhao, Erica Cooper, and Junichi Yamagishi, "Learning Disentangled Phone and Speaker Representations in a Semi-Supervised VQ-VAE Paradigm," in *ICASSP*, 2021, pp. 7053–7057.

[10] Ting Yao Hu, Ashish Shrivastava, Oncel Tuzel, and Chandra Dhir, "Unsupervised style and content separation by minimizing mutual information for speech synthesis," in *ICASSP*, 2020, vol. 2020-May, pp. 3267–3271.

[11] Mohamed Elgaar, Jungbae Park, and Sang Wan Lee, "Multi-speaker and multi-domain emotional voice conversion using factorized hierarchical variational autoencoder," in *ICASSP*, 2020, vol. 2020-May, pp. 7769–7773.

[12] Julian Zaidi, Jonathan Boilard, Ghyslain Gagnon, and Marc-André Carbonneau, "Measuring disentanglement: A review of metrics," *arXiv preprint arXiv:2012.09276*, 2020.

[13] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan, "Variational inference of disentangled latent concepts from unlabeled observations," *arXiv preprint arXiv:1711.00848*, 2017.

[14] Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer, "Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6056–6065.

[15] Karl Ridgeway and Michael C Mozer, "Learning deep disentangled embeddings with the f-statistic loss," *arXiv preprint arXiv:1802.05312*, 2018.

[16] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner, "dsprites: Disentanglement testing sprites dataset," https://github.com/deepmind/dsprites-dataset/, 2017.

[17] Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee, "Deep visual analogy-making," *Advances in neural information processing systems*, vol. 28, pp. 1252–1260, 2015.

[18] Yann LeCun, Fu Jie Huang, and Leon Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in *Proceedings of CVPR 2004*. IEEE, 2004, vol. 2, pp. II–104.

[19] Chris Burgess and Hyunjik Kim, "3d shapes dataset," https://github.com/deepmind/3dshapes-dataset/, 2018.

[20] Ashis Pati, Siddharth Gururani, and Alexander Lerch, "dmelodies: A music dataset for disentanglement learning," *arXiv preprint arXiv:2007.15067*, 2020.

[21] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *International Conference on Machine Learning*, 2019, pp. 4114–4124.

[22] Dennis H. Klatt, "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America*, vol. 67, pp. 971–995, mar 1980.

[23] Laurianne Georgeton, Nikola Paillereau, Simon Landron, Jiayin Gao, and Takeki Kamiyama, "Analyse formantique des voyelles orales du français en contexte isolé: à la recherche d'une référence pour les apprenants de fle," in *Conférence conjointe JEP-TALN-RECITAL 2012*, 2012, pp. 145–152.

[24] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al., "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[25] John Kominek, Tanja Schultz, and Alan W Black, "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion," in *Spoken Languages Technologies for Under-Resourced Languages*, 2008.

[26] Maximillian Nickel and Douwe Kiela, "Poincaré embeddings for learning hierarchical representations," *Advances in neural information processing systems*, vol. 30, pp. 6338–6347, 2017.