Disentangling Factor Of Variation Using Cyclic Consistency In Speech

Aayush Kumar Tyagi
IIIT
New Delhi
aayush16081@iiitd.ac.in

Dr. Saket Anand
IIIT
New Delhi
anands@iiitd.ac.in

Abstract

With recent improvements in generative models, several attempts have been made to extract the potential of disentanglement. We can use learned disentangled factor of variation to generate new data by simply tweaking the latent vector or it can be useful for data augmentation task. Also learned disentangled representation can be used for efficient classification and transferring attributes for one data(whether it is in form of speech or image) to other data point.

Although we could not find out many attempts of disentangling in speech as speech is a much harder problem as compared to the image. Taking forward what speech community has achieved, we made a sincere attempt to learn disentanglement of gender-specific information in speech. In order to disentangle gender-specific information, we introduced cyclic consistency in already existing Factorized hierarchical variational autoencoder[5]. We used weak supervision as a pair of labels were provided during training. For training purpose model is trained on the TIMIT dataset. For evaluation purpose, we performed qualitative and quantitative training gender classifier and reconstructing a speech waveform by swapping latent factors respectively.

1. Introduction

Unsupervised learning has vast potential due to the abundance of un-annotated data present around which can be learned and used in supervised and unsupervised fashion. Good amount of research has been focused on unsupervised learning domain, as a result, we have seen recent advancement in generative models like Variational Autoencoder[6], Adversarial autoencoders[7], Generative adversarial autoencoder [3].Once generative models capable enough to generate good quality speech, next step is to get hold of hidden parameters present in speech or in images. If we have hold of hidden parameters, we can simply tweak parameters to generate new data or transfer attributes from one speech waveform to another.

Our work revolves around disentangling gender-specific information from speech waveform. Sequential data such as speech waveform contains multi-scalar information. For example information about the speaker, channel, accent and linguistic content is encoded into the various level of hierarchy like utterance level or segment level. In order to use the hierarchical structure of speech for disentanglement, we used Factorized hierarchical variational autoencoder[5]. Fhvae differentiates information present in speech into sequence level attributes and segment level attributes. Sequence-level attributes consist of effects which are consistent over the long range such are fundamental frequency(f0) and volume whereas segment level attributes vary dynamically along with speech waveform and examples include phonetic content. We aim to disentangle the pitch information which highly co-relates with the gender of a person. It is observed that female(100Hz - 526Hz) have the higher pitch as compared to males(65Hz - 260Hz). For evaluation purpose, we performed quantitative analysis by training gender classifier and for qualitative analysis, we synthesizing test waveforms while switching the pitch information between two waveforms.

2. Related Work

It is interesting to note that a good amount of research is going on in the generative model, learning data representation and learning disentanglement as these problems pose a great potential for solving many other problems.

2.1. Autoencoder

2.1.1 Autoencoder

Autoencoder([2] ,[4] , [1] learns to represent data (image , speech , or text)) from higher dimensional subspace to lower dimensional subspace often called as latent representation or latent vector . Autoencoder consists of two subparts namely encoder and decoder. Encoder maps the input data into lower dimensional sub-space. The decoder uses this lower dimensional learned feature representation to reconstruct original data. We minimize mean square error between reconstructed output and original input, in order to check whether the encoder is able to map important information into latent space and decoder is able to reconstruct data using that latent feature representation. Autoencoder finds its application in various fields mainly in dimensionality reduction, initialising neural network and denoising autoencoder.

2.1.2 Variational Autoencoder

Variational Autoencoder [6] tries to learn latent space representation of data by learning the data distribution. Unlike autoencoders, latent attributes of input data are represented as probability distribution with a mean and variance. When random samples are feed to decoder, values which are nearby should correspond with similar reconstructions. Variational autoencoder use a generative model ; P(x , z) = P(x/z)P(z), where x is observed data (images or speech) , z are latent variable.The encoder estimates the parameters of the posterior ,Enc(x) = P(z / x) , decoder estimates the conditional likelihood , Dec(z) = P(x/z) . Variational autoencoder tries to closest estimate the distribution of data.

2.2. Generative Adversarial Network

Generative Adversarial network [3] is two network arrangement pitched against each other, each one trying to fool other network. The generator takes a random sample and produces an image, Discriminator is provided with both a fake image and original image from the dataset. Discriminators job is to provide probability between 0 and 1, that generated image is from real distribution or generated from the random sample. Generator tries to fool discriminator by generating images close to real one which discriminator cannot discriminate ie. D(G(x)) $\sim$ 1. Discriminator tries best to discriminate between generated fake image and real image from data distribution ie D(G(x)) $\sim$ 0. After successful training of the network, the discriminator is not able to tell real from generated images.

2.3. Disentangling factor of variation

Work proposed by [8] disentangle the image information into style and class latent space. Style space is represented by z and class latent space is represented as s. For disentangling the style and class content, the pair of labels X_1 and X_1' are provided to an encoder for mapping into latent space, we use the l2 loss on reconstructed output to check whether the decoder is reconstructing information correctly. In the second iteration, we swap the style information while keeping the class information constant. We use the l2 loss in order to check the reconstruction is correct. There are chances that the style information is passing through class latent space.The decoder could learn to ignore s, and the approximate posterior could map belonging to the same class to different regions of latent space, this problem is called a degenerate solution. In order to check there is no such flow of information takes place Adversarial training procedure is used. We provide X_2 of different class and style. Use encoder to map image into latent space into style and class latent vectors. In order to check style, information is not passing through space subspace we provide the class label of X_2 and style of X_1 and use a discriminator to differentiate between the class identity of X_2 and identity of X_1.In order to see class information remain intact, we provide class latent vector from X_2 and sample style latent vector from Normal distribution and train discriminator to differentiate between generated class and original class. A similar approach can be seen in challenges in disentangling independent factor of variation[9].

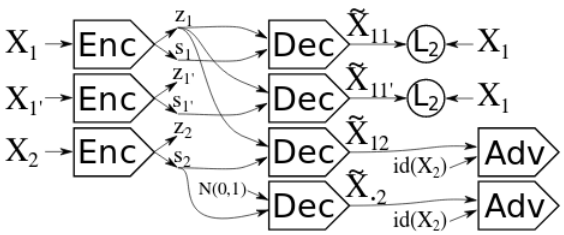

Figure 1. Training architecture . The input $x1$ and $x1'$ are two different samples with same label , whereas $x2$ can have any label.Here z represent style latent vector and s represent style latent vector.

3. Methodology

3.1. Factorised Hierarchical Variational Autoencoder

Factorized hierarchical variational autoencoder, is presented by [5], learns disentangled and interpretable representation from sequential data in an unsupervised manner by explicitly modelling the multi-scalar information with factorized hierarchical graphical model. Generation of sequential data .such as speech, often involves multiple independent factors operating at different time scales. For instance, the speaker identity affects fundamental frequency (F0) and volume at the sequence level, while phonetic content only affects spectral contour and duration of formants at segment level.

The distinction between segment level and sequence level attributes come from the fact that F0 and volume, tend to have a smaller amount of variation within utterances, compared to between utterances; while other attributes, such as phonetic content, tend to have a similar amount of variation within and between utterances.

Here the first type of attributes are referred as sequence level attribute, and the other as segment level attributes.

In this work, in order to achieve disentanglement and interpretability by encoding the type of attributes into latent sequence variable and latent segment variable respectively, where the former is regularized by the sequence-dependent prior and the latter by a sequence-independent prior.

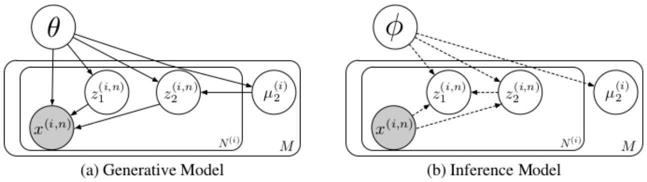

Figure 2. Graphical illustration of proposed generative model and inference model.Grey nodes denote the observed variable and white node represent hidden variable

For formulation purpose of factorized Hierarchical Variational Autoencoder(FHVAE) , dataset D consist of M i.i.d sequence ,where each sequence consist of N segments.For generative process sequence X is generated from Z_1 , Z_2 and μ_2. Here Z_1 corresponds to sequence level attributes and Z_2 corresponds to segment level attributes.μ corresponds to s-vector based on speaker attributes.Z_2 is drawn form sequence dependent prior distribution $p_\theta(z_2|\mu_2)$ and a sequence independent prior distribution $p_\theta(z_1)$..θ is used to denote set of parameters in the generative model,ϕ is used to denote the parameters of Inference model. We maximize the sequence variational lower bound $\mathcal{L}(\theta,\phi;X)$ which can be decomposed into sum of $\mathcal{L}(\theta,\phi;x^{(n)}|\mu_2)$, the conditional segment variational lower bounds, over segment plus the log prior probability of μ_2 and a constant.

$$\mathcal{L}(\theta,\phi;x^{(n)}) = \mathcal{L}(\theta,\phi;x^{(n)}|\mu_2) + \frac{1}{N}\log p_\theta(\mu_2) + const \tag{1}$$

Discriminative objective

The idea of having the sequence-specific priors for each sequence is to encourage the model to encoder the sequence-level attributes and segment level attributes into the different set of latent variables. However, when μ_2= 0 for all sequences, this result in trivial s-vector μ_2, and therefore z_1 and z_2 would not be factorized to encode sequence and segment attributes respectively.

$$\mathcal{L}^{dis} = \mathcal{L}(\theta,\phi;x^{(i,n)}) + \alpha\log p(i|z_2^{(i,n)}) \tag{2}$$

We use neural network for functional mapping of input data X to μ and σ for both $Z1$ and $Z2$. We generate the reconstructed waveform using these estimated parameters.

As temporal structure within the segment is important, LSTM module is used. Network consists of three Long short-term memory(LSTM) layers, each LSTM followed by a fully connected layer.Model have one hidden layer with 16 dimensional for both $Z1$ and $Z2$, and trained for $\alpha = 0, 10$.

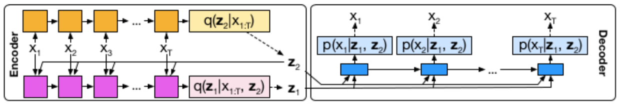
Encoder
$q(\mathbf{z}_2|x_{1:T})$
x_1
x_2
x_3
x_T
$\mathbf{z}_2$
$q(\mathbf{z}_1|x_{1:T}, \mathbf{z}_2)$
$\mathbf{z}_1$
x_1
x_2
x_T
$p(x_1|\mathbf{z}_1, \mathbf{z}_2)$
$p(x_2|\mathbf{z}_1, \mathbf{z}_2)$
$p(x_T|\mathbf{z}_1, \mathbf{z}_2)$
Decoder

Figure 3. Sequence to sequence factorized hierarchical variational autoencoder . Diashed lines indicate the sampling process using the reparameterization trick . The encoder for $z1$ and $z2$ are pink and amber , respectively ,while the decoder of x is blue . Darker colors denote the recurrent neural networks , while lighter colors denote the fully connected layers predicting the mean and log varience.

3.2. Cyclic consistency

Cyclic consistency is yet another way to prevent degenerate solution[8], which prevents information to leak from another latent subspace. Cyclic consistency is performed in two cycles. First forward cycle and second backward cycle. Cyclic consistency work on the principle that the forward and backward cycle composited together in any order should correspond to the identity function.For example mapping G : X → Y and F : Y → X , F(G(X)) ∼ X

Forward cycle:

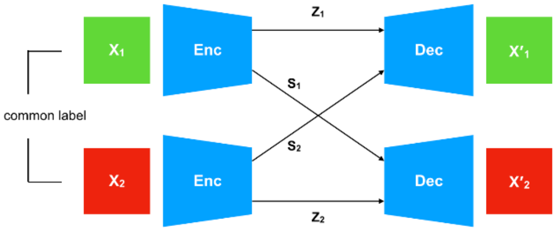

Figure 4. Here $x1$ and $x1'$ represent different images sharing same labels .In forward pass we swap the style information while reconstructing image using Variational Autoencoder

Forward cycle is provided by common labels.Encoder split the input information into class(s) and style(z) latent space .Encoder learns mapping X_1 and X_2 to (s_1,z_1) and (s_2,z_2) respectively. We interchange the style information while providing it to decoder for reconstruction.Decoder reconstructs the image with class s_1 , style z_2 and class s_2 ,style z_1. But it is possible that class information is passing through style space ie. decoder learns to neglect the class information and learns to reconstruct only through style subspace. The concern here is there is no check on the leak of information from some latent sub-space. It is possible that class information may pass through style latent sub-space and decoder learns to simply ignore z latent vector and reconstruct using style latent vector s. Backward cycle:

In order to protect that class, information is not passing through style space or vice-versa. We provide different class images to the encoder. Encoder maps the image into style and class latent space. We sample a random point

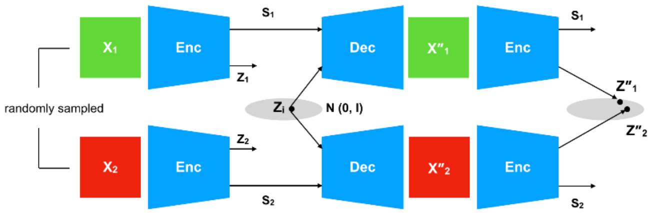

Figure 5. In backward pass we sample style information from gaussian distribution ,which is used for reconstructing image of class s .This reconstructed image is provided to encoder and style sample points are mapped close by.

from Gaussian distribution and use this information as style latent space.Decoder use class and style information to reconstruct the image. We provide the reconstructed image to the encoder. Encoder again maps image information into style and class latent space. For cyclic consistency our assumption is the as we have sampled style latent vector from Gaussian distribution and used same style vector for both the image reconstruction, the encoder should be able to map both the styles to the same point(If not same but at least closeby). Hence we can say that $|z_1 - z_1'|_2 \to 0$, that means style latent variable does not contain any information about the class.

3.3. Combining FHVAE and Cyclic Consistency

Both FHVAE and Cyclic consistency are independently great ideas for disentanglement. Although results obtained by both of them are not so promising. Waveform generated by FHVAE still sound mechanical when the latent variables are swapped between two waveforms.

For further improvement we combined both FHVAE and Cyclic consistency concept. So total loss is combination of fhvae loss and cyclic loss.

$$\mathcal{L}^{total} = \mathcal{L}^{FHVAE} + \mathcal{L}^{cyclicloss} \tag{3}$$

For training purpose, the pair of labels is created. For the forward cycle, common labels are used where both the waveform sample are of the same gender though can be of the different speaker. For the backward cycle, different labels are used where both the waveform samples are of different gender.

4. Experiment and results

We used TIMIT for our experimentation purpose. The corpora contain the broadband 16KHz recording of phonetically balanced read speech. A total of 6300 utterances (5.4 hours) are presented with 10 sentences for every 630 speakers, of which approximates to 70 % are male and 30% are female. All speech is represented as a sequence of 80 dimensional Mel-scale filter bank (FBank) features or 200 dimensional log-magnitude(only for audio reconstruction), computed every 10 ms. Mel-scale features are a popular auditory approximation for many speech applications. We consider a sample x to be 200 ms sub-sequence, which is on the order of length of a syllable and implies T=20 for each x.For the seq2seq FHVAE model, all the models LSTM and MLP are one layered and ADAM is used for optimization. FHVAE model is also trained with cyclic consistency loss i.e. if reconstructed output is feed to the FHVAE as input to get the new latent variables z_1' and z_2' then $d(z_1, z_1')$ and $d(z_2, z_2')$ is used as a regularizer. Since output is reconstruction of input then z_1, z_1' and z_2,z_2' should be close to each other. The $d(,)$ that is tested are $l_1(z - z')$ and $mse(z, z')$. For FHVAE training with cyclic constraints λ weighing is used which is given by:

$$\lambda = \frac{2}{1 + \exp(-\gamma * p)} - 1$$

, where p changes from 0 to 1 linearly as training progresses and γ is set to 10. Additionally, a weighting factor of 0.8 is also used and it seems to perform better than the previous one.

Hypotheses:

- z_1 / z_2 is invariant or not to speaker gender.
- Encoder is mapping input and reconstructed input to the same point in latent space.
- Gender swapped reconstructed speech contains the gender of the target reconstructed speech i.e. If segment latent variable of source audio contains the gender information also then it will confuse the gender classifier, so segment latent variable should not contain gender information.

To test how much invariance/ not the FHVAE model has to ...

4.1. Evaluation of cyclic consistency

Latent factor z_1 encode the *sequence* information and latent factor z_2 encode the *segment* information. Sequence level latent variable has a prior $N(\mu_2, {\sigma^2}_{z_1})$, μ_2 has a prior $N(0, {\sigma^2}_{\mu_2})$ and segment level latent variable has prior $N(0, \sigma_{z_2})$. In the original FHAVE implementation segment level latent variable prior is sampled as $\mu_2 + sample(N(0, 1))$ and segment level latent variable prior is sampled as sample from$N(0, 1)$. In the first set of experiments, cyclic consistency of FHVAE encoder is tested. Lets consider z_1 and z_2 are latent variables from original audio and z_1' and z_2' latent variables are from reconstructed audio. In the first column, of Table 1, the error values are for original FHVAE model. In the second and third column a reconstruction loss of mse and $l1$ is applied on (z_1, z_1') and (z_2, z_2').

The reconstructed output is feed to the FHVAE as input to get the new latent variables z_1' and z_2'. Since original utterance and reconstructed utterance contains the same content, z_1' and z_2' should be close to each other given some noise in the reconstructed utterance. The error values are reported on the test set. The 5th and 6th column indicates the cyclic loss. The idea is We feed the data in pairs(Data frames from the same sequence but different frames). The pairwise data is feed to the encoder and we get pairs of z_1's and z_2's. A sample from the Sequence prior is taken and is feed to the decoder along with the original Segment pair. The reconstructed audio is fed to the encoder again to get another pair of z_1's and z_2's. The l1 and l2 loss are minimized between the pairs of z_1's from the reconstructed audio.

Error measure	Original	l2 constraint	l1
Mean(abs(z1 - z1'))	0.157239	0.3644	0.5569
Mean(abs(z2 -z2'))	0.172542	0.2945	1.098951
MSE Z1	0.049682	0.2097	0.5014
MSE Z2	0.061015	0.1451	1.9781
Trace(Covariance(Error Z1))	1.575439	6.711	14.6475
Mean(diag(Covariance(Error Z1)))	0.0492324	0.209	0.457734
Trace(Covariance(Error Z2))	1.9279739	4.298	61.41
Mean(diag(Covariance(Error Z2)))	0.0602491	0.13431	1.91

4.2. Evaluate Gender content in Z_1 and Z_2

This experiment was performed in order to check whether z_1 and z_2 contains information about gender.We trained two classifiers one on z_1 and other on z_2.Dimension of z_1 and z_2 is (batch size,32).Classifier is a simple multi-layer perceptron consisting 2 hidden layers[32,25,10,2],where 32 is input layer and 2 is number of neurons in output layer.

latent var	Test Accuracy	l2	l1
Z_1	95.677	66.73	95.96
Z_2	66.73	83.029	70.63

4.3. Evaluate gender content in Reconstructed Speech

In this experiment gender classifier is trained on input speech data and it is tested on reconstructed same gender(Male to male, Female to female) and cross-gender(Male to Female, Female to male). For testing purpose, we create 200 utterances for each of sub-class ie for male-male,female-female,male-female,female-male. Here input speech data dimension is (batch size,).The architecture used for the classifier. The Classifier is a simple multi-layer perceptron consisting 4 hidden layers[4000,2000,1000,500,100,2],where 4000 is input layer and 2 is number of neurons in output layer.

Reconstruction	Test Accuracy
Input Speech	91.1
Male to Male	93.9
Female to Female	93.7
Male to Female	81.4
Female to Male	86.5

4.4. Visualization of latent space

We can verify what we have proposed visually by plotting TNSE plot of latent vectors. Figure 6 shows the TSNE plot of Z_2 latent vector which represents segment level attributes. Here red dots are male speaker waveforms and blue dots are samples from female speaker waveforms. As expected male and female samples are randomly distributed meaning that segment level attributes do not contain gender-specific information. Figure 7 shows the Z_1 latent vector which represents sequence level attributes.

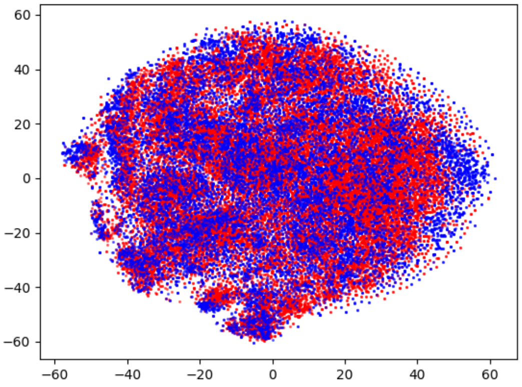

Figure 6. TSNE plot of Z_2 latent vector.Here blue represents female samples and red represent male samples.As plot is completely random ,which means that Z_2 does not contain any gender information

5. Conclusion

Although disentangling is a challenging problem, we propose another way of looking at the problem of combining frameworks of FHVAE and cyclic consistency. We tried to make Z_1 is speaker independent ie Z_2 contains all the speaker information. Results obtained by using classifier proves that Z_1 contains gender information whereas $Z2$ produced accuracy around 60 per cent which is close to a random guess.

References

[1] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on*

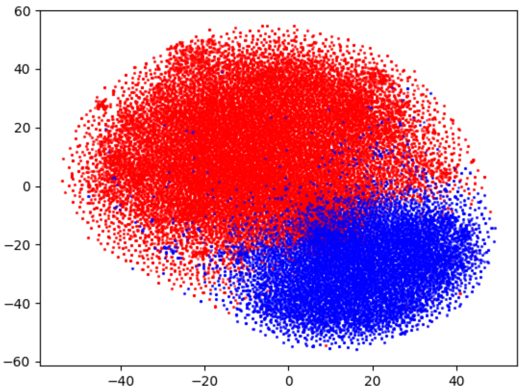

Figure 7. TSNE plot for Z_1 latent vector.

pattern analysis and machine intelligence, 35(8):1798–1828, 2013.

[2] H. Bourlard and Y. Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4-5):291–294, 1988.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[4] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[5] W.-N. Hsu, Y. Zhang, and J. Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in neural information processing systems*, pages 1878–1889, 2017.

[6] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[7] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[8] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*, pages 5040–5048, 2016.

[9] A. Szabó, Q. Hu, T. Portenier, M. Zwicker, and P. Favaro. Challenges in disentangling independent factors of variation. *arXiv preprint arXiv:1711.02245*, 2017.