VAE-SPACE: DEEP GENERATIVE MODEL OF VOICE FUNDAMENTAL FREQUENCY CONTOURS

Kou Tanaka[1]*, Hirokazu Kameoka*[1]*, Kazuho Morikawa*[2]

[1]NTT Communication Science Laboratories, NTT Corporation, Japan,
[2]Graduate School of informatics, Nagoya University, Japan

{tanaka.ko, kameoka.hirokazu}@lab.ntt.co.jp, morikawa.kazuho@g.sp.m.is.nagoya-u.ac.jp

ABSTRACT

Modeling the speech generation process can provide flexible and interpretable ways to generate intended synthetic speech. In this paper, we present a deep generative model of fundamental frequency (F_0) contours of normal speech and singing voices. The generative model we propose in this paper 1) is able to accurately decompose an F_0 contour into the sum of phrase and accent components of the Fujisaki model, a mathematical model describing the control mechanism of vocal fold vibration, without an iterative algorithm, and 2) can represent/generate F_0 contours of both normal speech and singing voices reasonably well.

***Index Terms*—** Deep generative model, voice F_0 contour, singing voice, variational autoencoder, gated convolutional network

1. INTRODUCTION

The fundamental frequency (F_0) contours in normal speech contain linguistic and para/non-linguistic information. For example, they are usually used to convert a regular phrase to a question. They also indicate intonation in pitch accent languages. Furthermore, they play the role of adding extra flavor to speech such as the identity, intention, attitude, and mood of the speaker to convey para/non-linguistic information to the listener. In singing voice, they are used to express the melody of the song and the singing style of the singer. If we can build a physically or musically interpretable generative model, it may provide flexible ways to synthesize expressive speech or singing voices. This paper is concerned with developing a generative model of voice F_0 contours, which allows us to generate natural sounding F_0 contours conditioned on a contextual input such as a phrase/accent command sequence and a musical score.

Conventionally, several attempts have been made to model F_0 contours of speaking and singing voices. One well-known model is called the Fujisaki model [1], which describes the control mechanism of vocal fold vibration in a physically interpretable way. This model assumes that the F_0 contour on a logarithmic scale is the superposition of a phrase component, an accent component, and a base value. The phrase and accent components are considered to be associated with mutually independent types of movement of the thyroid cartilage with different degrees of freedom and muscular reaction times. Another example is F_0 control models of singing voices [2, 3]. Similar to the Fujisaki model, these models assume that a singing voice F_0 contour is described as a mixture of several types of F_0 fluctuations such as overshoot, vibrato, and preparation. The Fujisaki model and the singing voice F_0 control models share in common that there is a need to solve an inverse problem to obtain the underlying parameters. Although the models reported in [2, 3, 4] have provided a tractable way of estimating the underlying parameters by using statistical inference techniques, shortcomings of these models are that parameter estimation algorithms typically require many iterations, which can be computationally demanding, and parameter estimation accuracy is limited due to the inherent difficulties in the inverse problem.

Recently, several types of generative model described by a neural network have been proposed, such as a variational autoencoder (VAE) [5, 6]. As the name implies, VAEs are a stochastic counterpart of autoencoders, consisting of encoder and decoder networks. The encoder network generates a set of parameters of the conditional distribution $Q(\boldsymbol{z}|\boldsymbol{x})$ of a latent space variable $\boldsymbol{z}$ given an input data vector $\boldsymbol{x}$ whereas the decoder network generates a set of parameters of the conditional distribution $P(\boldsymbol{x}|\boldsymbol{z})$ of the data vector x given the latent space variable $\boldsymbol{z}$. Given a training data set $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^N$, VAEs learn the entire network parameters so that the encoder distribution $Q(\boldsymbol{z}|\boldsymbol{x})$ becomes consistent with the posterior $P(\boldsymbol{z}|\boldsymbol{x}) \propto P(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$. If we can associate the latent space variables with a set of interpretable parameters governing the data of interest, the decoder can be seen as a generative model (like the Fujisaki model) that relates the underlying parameters to observed data and the encoder can be seen as an inverse problem solver. While the Fujisaki model, for example, is a hand-crafted or manually designed model, an interesting point of view would be that through training of VAE, we would be able to discover the structure of a generating process model in a data-driven manner as well as an inverse process of estimating the underlying parameters. Furthermore, since VAEs provide a principled and convenient way of handling semi-supervised learning tasks [6, 7], they can be very useful especially when it takes a lot of time and effort to collect a large amount of labeled data. For our task, while collecting a complete pair of F_0 contours and the underlying parameters can be a demanding process, we can easily collect a large amount of unlabeled F_0 contours. Indeed, VAEs have been applied to various supervised/semi-supervised tasks with notable success [8, 9, 10].

In this paper, we propose a generative model of F_0 contours based on a VAE with a fully convolutional architecture. In particular, we adopt a gated CNN architecture [11] to be

able to capture and reflect long- and short-term dependencies in F_0 contours. Experimental results showed that our proposed framework successfully achieved higher performance than a conventional method in terms of the subjective pairwised comparison for singing voice quality, the generation error of the underlying parameters of the F_0 contours in speaking voice, and processing time required to solve the inverse problem.

2. F_0 CONTOUR AND ITS UNDERLYING PARAMETERS

Here, we briefly review conventional work on voice F_0 contour modeling for speaking and singing voices.

2.1. Fujisaki model

The Fujisaki model [1] is one of well-known models describing the control mechanism of vocal fold vibration in a physically interpretable way. This model assumes that the F_0 contour $x[t]$ on a logarithmic scale is given as the sum of three components $x[t] = x_{\mathrm{p}}[t] + x_{\mathrm{a}}[t] + \mu_{\mathrm{b}}$ where $x_{\mathrm{p}}[t]$ and $x_{\mathrm{a}}[t]$ are a phrase component and an accent component at time frame t, and μ_{b} is a constant value, respectively. The phrase and accent components are assumed to be the outputs of different second-order critically damped filters excited with Dirac deltas (phrase commands) and rectangular pulses (accent commands), respectively. These components respectively correspond to contributions associated with the translation and rotation movements of the thyroid cartilage. The former usually contributes to phrasing, while the latter contributes to accentuation during an utterance. The magnitudes of these components correspond to how much emphasis the speaker intends to place on the associated phrase or accent. These parameters, which we call the Fujisaki model parameters, are thus physically and linguistically interpretable. If we can estimate these parameters from raw F_0 contours, we will be able to flexibly control them as desired.

2.2. Singing voice F_0 contour model

The F_0 contour of a singing voice contains the melody contour of a song and an expression contour such as overshoot, preparation and vibrato. Compared with the F_0 contours of speaking voice, those of singing voice change more rapidly, and their dynamic range is wider and so the Fujisaki model cannot be directly applied to singing voices. In [3], a singing voice version of the Fujisaki model is proposed.

3. VAE-SPACE: F_0 CONTOUR REPRESENTATION VIA DEEP GENERATIVE MODEL

3.1. Concept

In [4], we proposed formulating a stochastic counterpart of the Fujisaki model. The key idea of this model is that a phrase/accent command pair sequence, given by an impulse train and a rectangular pulse train, is modeled as an output sequence of a path-restricted hidden Markov model (HMM). Similarly, [3] proposed introducing a stochastic counterpart of a singing voice version of the Fujisaki model. With this model, two sequences, one representing a melody contour and the other representing an expression contour, are modeled as piecewise constant functions generated by a path-restricted HMM. These models have allowed us to utilize statistical inference techniques to estimate the underlying parameters. However, the parameter estimation algorithms must be run for many iterations, which can be computationally demanding. Furthermore, parameter estimation accuracy is limited due to the inherent difficulties in the ill-posed inverse problem. Another limitation of these models is the lack of flexibility needed to express a wide variety of voice F_0 contours and neither of these models can be universally applied to all possible F_0 contours. We may need to manually design different models and algorithms according to languages, speakers and types of speech (e.g., singing voices and regular/emotional speech) as long as we take a hand-engineering approach.

To overcome these limitations, we take a learning-based approach. In particular, we focus on VAEs with a gated CNN architecture for flexibly modeling voice F_0 contours. As mentioned in Sec. 1, if we can associate the latent space variables with a set of interpretable parameters (like the phrase/accent components in the Fujisaki model) governing the data of interest, the decoder can be seen as a generative model (like the Fujisaki model) that relates the underlying parameters to observed data whereas the encoder can be seen as a parameter extractor. It would be interesting if we could automatically discover the structure of a generating process model in a data-driven manner through training of VAE as well as the parameter estimation process. Furthermore, since VAEs provide a principled and convenient way of handling semi-supervised learning tasks, they can be very useful especially when it takes a lot of time and effort to collect a large amount of labeled data. For our task, even though collecting a complete pair of F_0 contours and the underlying parameters can be a painstaking process, we can easily collect a large amount of unlabeled F_0 contours. These are the main reasons we have focused on VAEs.

To realize the above-mentioned concept, an architecture design is a key to success. Given the fact that both the Fujisaki model and the singing voice F_0 contour model mentioned in Sec. 2 are described as a mixture of linear time-invariant systems, we believe that convolutional architectures can be a reasonable choice for our architecture design. In particular, we focus on a convolutional architecture called the gated CNN. The gated CNN has recently been shown to be powerful in modeling long-term sequential data. It was originally introduced to model word sequences for language modeling and was shown to outperform long short-term memory (LSTM) language models trained in a similar setting [11]. We previously applied a gated CNN architecture for speech sequence modeling and its effectiveness has already been confirmed [12]. With a gated CNN, the output of a hidden layer of a network is described as a linear projection modulated by an output gate. Similar to an LSTM [13] and gated recurrent unit (GRU) [14], the output gate controls what information should be propagated through the hierarchy of layers and allows capturing long-term structures.

3.2. VAE-SPACE

Let us use $\boldsymbol{z}$ to denote a sequence of parameters governing the generating process of F_0 contours. In the case of the Fujisaki

model, this corresponds to a sequence of a phrase/accent component pair. Here, we consider a "decoder" network that generates the parameters of a conditional distribution $P_\theta(\boldsymbol{x}|\boldsymbol{z})$ of an F_0 contour $\boldsymbol{x}$. The posterior distribution $P_\theta(\boldsymbol{z}|\boldsymbol{x})$ can be seen as an inverse process of generating $\boldsymbol{z}$ given $\boldsymbol{x}$. Since obtaining the exact posterior is intractable, we introduce another network, i.e., "encoder", that generates the parameters of a conditional distribution $Q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and train both the decoder and encoder networks so that $Q_\phi(\boldsymbol{z}|\boldsymbol{x})$ becomes consistent with the exact posterior $P_\theta(\boldsymbol{z}|\boldsymbol{x}) \propto P_\theta(\boldsymbol{x}|\boldsymbol{z})P(\boldsymbol{z})$. We can show that the log marginal distribution $\log P_\theta(\boldsymbol{x})$ is given as

$$\log P_\theta(\boldsymbol{x}) = \mathcal{L}(\theta, \phi; \boldsymbol{x}) + D_{\mathrm{KL}}\left[Q_\phi(\boldsymbol{z}|\boldsymbol{x})||P_\theta(\boldsymbol{z}|\boldsymbol{x})\right], \quad (1)$$

$$\mathcal{L}(\theta, \phi; \boldsymbol{x}) = \underbrace{-D_{\mathrm{KL}}\left[Q_\phi(\boldsymbol{z}|\boldsymbol{x})||P(\boldsymbol{z})\right]}_{\text{Regularization term over } \boldsymbol{z}} + \underbrace{\mathbb{E}_{Q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log P_\theta(\boldsymbol{x}|\boldsymbol{z})\right]}_{\text{Reconstruction term}} \quad (2)$$

where $D_{\mathrm{KL}}[\cdot|\cdot]$ denotes the Kullback-Leibler (KL) divergence. This implies we can minimize the KL divergence between $P_\theta(\boldsymbol{z}|\boldsymbol{x})$ and $Q_\phi(\boldsymbol{z}|\boldsymbol{x})$ by maximizing $\mathcal{L}(\theta, \phi; \boldsymbol{x})$ with respect to θ and ϕ. One typical way of modeling $Q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and $P_\theta(\boldsymbol{x}|\boldsymbol{z})$ is to assume normal distributions. As for the prior distribution $P(\boldsymbol{z})$, we can design its specific form according to the assumption we would like to make about $\boldsymbol{z}$. For example, if we associate $\boldsymbol{z}$ with a phrase/accent command pair sequence, we can employ the path-restricted HMM with Gaussian emission densities proposed in [4]. In this case, by using s to denote the state sequence of the HMM, $P(\boldsymbol{z})$ is written as $P(\boldsymbol{z}) = \sum_{\boldsymbol{s}} P(\boldsymbol{z}|\boldsymbol{s})P(\boldsymbol{s})$. Since our VAE is designed to perform statistical phrase/accent component estimation (SPACE), we call it "VAE-SPACE".

3.3. Sequential modeling with gated CNN

To capture long- and short-term dependencies in F_0 contours, we use a gated CNN [11] to construct both the encoder and decoder networks of the VAE. Gated CNNs are CNNs equipped with gated linear units (GLUs) as activation functions instead of regular rectified linear units (ReLUs) [15] or Tanh activations. The output of the l_{th} hidden layer of a gated CNN is described as a linear projection $\boldsymbol{H}_{l-1} * \boldsymbol{W}_l + \boldsymbol{b}_l$ modulated by an output gate $\sigma(\boldsymbol{H}_{l-1} * \boldsymbol{V}_l + \boldsymbol{c}_l)$

$$\boldsymbol{H}_l = (\boldsymbol{H}_{l-1} * \boldsymbol{W}_l + \boldsymbol{b}_l) \otimes \sigma(\boldsymbol{H}_{l-1} * \boldsymbol{V}_l + \boldsymbol{c}_l), \quad (3)$$

where $\boldsymbol{W}_l$, $\boldsymbol{V}_l$, $\boldsymbol{b}_l$ and $\boldsymbol{c}_l$ are the network parameters to be trained, σ is the sigmoid function and $\otimes$ indicates the element-wise product. Here, the input to the 1_{st} layer is $\boldsymbol{H}_0 = \boldsymbol{x}$ for the encoder and $\boldsymbol{H}_0 = \boldsymbol{z}$ for the decoder whereas the output from the l_{th} layer is $\boldsymbol{H}_l = [\boldsymbol{\mu}_z; \log \boldsymbol{\sigma}_z^2]$ for the encoder and $\boldsymbol{H}_l = [\boldsymbol{\mu}_x]$ for the decoder. Similar to LSTMs, the output gate multiplies each element of $\boldsymbol{H}_{l-1} * \boldsymbol{W}_l + \boldsymbol{b}_l$ and control what information should be propagated through the hierarchy of layers in a data-driven manner.

4. EXPERIMENTS

4.1. Experimental Conditions

Datasets: For the speaking voice F_0 contours, we used the ATR speech database [16], 429 sentences (around 0.5 hours) uttered by one male speaker to train a deep generative model, and the remaining 53 sentences (around 3 minutes) to evaluate the performance. For singing voice F_0 contours, we used real singing data paired with the musical score, 42 songs (around 15 minutes) sung by 6 singers including female and male classical, pop, and amateur singers (namely, each singer recorded 7 songs). To evaluate the performance of singing voice F_0 contour modeling, we adopted the leave one out cross validation strategy over singer dependent models.

F_0 **extraction:** we adopted TEMPO [17] as an F_0 analyzer. Based on the label data, the frame shifts were set to 8 and 5 ms for the speaking and singing voices, respectively. Note that we carefully checked the actual extracted F_0 contours and excluded data that have failed to analysis. The final number of data after excluding was described in **Datasets**.

Model architecture: The model setting of the conventional stochastic model of F_0 generative process (**SPACE**) was the same as reported in [4]. Table 1 details the network architectures of our proposed model (**VAE-SPACE**) for speaking voice F_0 contour. The stride of each convolution was set to 1. For speaking voice, we set the number of channels over latent space, K, to 4 indicating the mean and variance values of the phrase and accent components. We canceled the baseline value 60 Hz of the speaking voice F_0 contour, in advance. For singing voice, we set K to 2 indicating the mean and variance values of the musical score. We normalized the singing voice F_0 contour, its musical score, and its backward difference to zero-mean and unit-variance using their training sets, respectively. We optimized the model parameters using the Adam optimizer [18] with a mini-batch of size 32. The learning parameters α, β_1, β_2 were set to 0.0001, 0.9, and 0.99, respectively.

Table 1. Details of network architecture.

Input	:	1ch.	1×1200
Encoder CNN 1	:	4 ch.	1×200, GLU
Encoder CNN 2	:	8 ch.	1×50, GLU
Encoder CNN 3	:	4 ch.	1×15, GLU
Encoder CNN 4	:	8 ch.	1×5, GLU
Encoder CNN 5	:	K ch.	1×50
Latent	:	K ch.	1×1200
Decoder CNN 1	:	16 ch.	1×10, GLU
Decoder CNN 2	:	1 ch.	1×10
Output	:	1 ch.	1×1200

Table 2. Results of subjective evaluation for F_0 contour similarity (# of evaluated samples: 42, # of evaluators: 7).

VAE-SPACE	Musical score or Fair	p-value
76.2 %	23.8 %	0.000312

4.2. Experimental Results

4.2.1. F_0 Contour Similarity Over Singing Voice

We subjectively evaluate singing voice synthesized with two types of F_0 contours. One is F_0 contour corresponding to the music score, and another is F_0 contour generated by using the only decoder part our proposed framework given the musical score after training both of the encoder and decoder parts of the VAE.

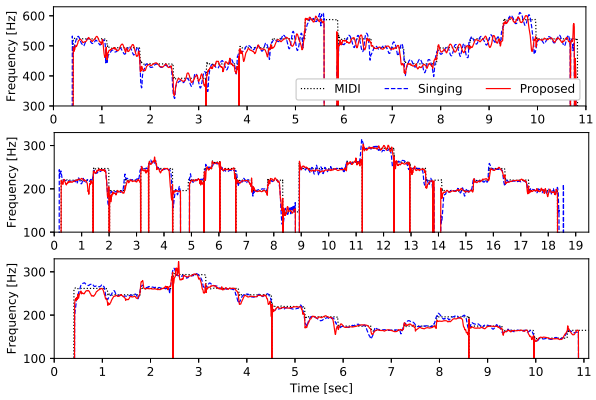

Fig. 1. Samples of F_0 contour and MIDI note for real singing data sung by female classical (top), male pop (middle), and male amateur (bottom) singers.

Table 3. Generation errors of F_0 contour and its underlying parameters for real speaking data (top) and "ideal" condition (bottom; # of evaluated samples: 53).

	VAE-SPACE	SPACE
F_0 contour	**0.0536**	0.0883
Phrase component	**0.0947**	0.123
Accent component	**0.0936**	0.122

F_0 contour	**0.0169**	0.0790
Phrase component	**0.0322**	0.131
Accent component	**0.0329**	0.110

Table 2 shows that our proposed model achieved to generate F_0 contours which are similar to those of real singing data. Note that the breakdown of "Musical score or Fair" is that "Musical score" and "Fair" are 7.14 % and 16.7 %, respectively. Considering the comments of evaluators, it seems that they have felt the fluctuations even if F_0 contours are monotonic, namely the musical score. One possible reason is that the singing style causes the fluctuations of acoustic features including not only F_0 contours but also spectral features and power information of the waveform. As shown in samples (Fig. 1), we confirmed that our proposed model made it possible to generate F_0 contours with the fluctuations, such as vibrato and overshoot.

4.2.2. Generation Error Over Speaking Voice

Calculating root mean square error (RMSE) between the reference and the generated one, we objectively evaluate the performance of our proposed model, **VAE-SPACE**. Note that not only actual F_0 contours observed in real speaking data but also F_0 contours reconstructed by the Fujisaki model parameters are used as the reference. Using F_0 contours reconstructed by the Fujisaki model parameters means the "ideal" condition.

Table 3 shows the performances in the case of training the model by using only actual F_0 contours observed in real speaking data and the "ideal" condition. Both of the results show that our proposed model successfully achieved higher performance compared with the conventional method. The major factors of getting high performance is the constraint of

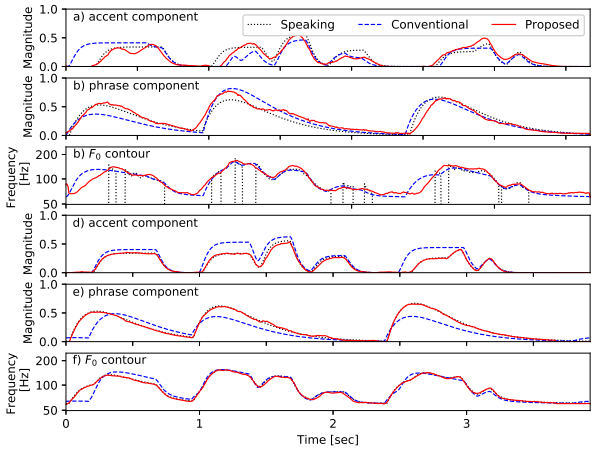

Fig. 2. Samples of F_0 contour and phrase and accent components for real speaking data (a $\sim$ c) and "ideal" condition (d $\sim$ f).

Table 4. Processing time to solve inverse mapping from F_0 contour to its underlying parameters (Unit: [sec]).

	VAE-SPACE	SPACE
53 sentences	**0.0126** $\pm$ 0.0002	2712.080
average	—	5.67

VAE that is the training of the parametric encoder in combination with the generator network. As shown in Fig. 2, the F_0 contours and their underlying parameters generated by our proposed model are more closer to the reference compared with those generated by the conventional method. In particular the "ideal" condition, the underlying parameters estimated by our proposed framework are truly close to the references.

4.2.3. Processing Time to Solve Inverse Mapping

To demonstrate the use of CNN architecture, we measure the processing time to estimate the underlying parameters of F_0 contours given actual obtained F_0 contours. Although our proposed model **VAE-SPACE** enables to work on a GPU, the conventional model **SPACE** works on only a CPU. The CPU and GPU are "Intel® Xeon® Processor E5-2699 v3" and "NVIDIA Corporation GK210GL [Tesla K80] (rev a1)", respectively.

Table 4 shows that our proposed model makes it possible to work in real time.

5. CONCLUSIONS

In this paper, we have presented a unified approach to model both of speaking and singing voice F_0 contours. The key role of our approach is a learning-based mapping to realize complex mapping, which is really difficult to fully elucidate, between F_0 contours and their underlying parameters. Experimental results revealed that the presented approach significantly outperforms the conventional stochastic model of F_0 generative process.

Acknowledgements: This work was supported by JSPS KAKENHI 17H01763.

6. REFERENCES

[1] Hiroya Fujisaki, "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour," *Vocal physiology: Voice production, mechanisms and functions*, pp. 347–355, 1988. 1, 2

[2] Siu Wa Lee, Shen Ting Ang, Minghui Dong, and Haizhou Li, "Generalized f0 modelling with absolute and relative pitch features for singing voice synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 429–432. 1

[3] Yasunori Ohishi, Hirokazu Kameoka, Daichi Mochihashi, and Kunio Kashino, "A stochastic model of singing voice f0 contours for characterizing expressive dynamic components," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012. 1, 2

[4] Hirokazu Kameoka, Kota Yoshizato, Tatsuma Ishihara, Kento Kadowaki, Yasunori Ohishi, and Kunio Kashino, "Generative modeling of voice fundamental frequency contours," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1042–1053, 2015. 1, 2, 3

[5] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013. 1

[6] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther, "Ladder variational autoencoders," in *Advances in Neural Information Processing Systems*, 2016, pp. 3738–3746. 1

[7] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling, "Semi-supervised learning with deep generative models," in *Advances in Neural Information Processing Systems*, 2014, pp. 3581–3589. 1

[8] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin, "Variational autoencoder for deep learning of images, labels and captions," in *Advances in Neural Information Processing Systems*, 2016, pp. 2352–2360. 1

[9] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio, "Generating sentences from a continuous space," *arXiv preprint arXiv:1511.06349*, 2015. 1

[10] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al., "Conditional image generation with pixelcnn decoders," in *Advances in Neural Information Processing Systems*, 2016, pp. 4790–4798. 1

[11] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier, "Language modeling with gated convolutional networks," *arXiv preprint arXiv:1612.08083*, 2016. 1, 2, 3

[12] Takuhiro Kaneko, Hirokazu Kameoka, Kaoru Hiramatsu, and Kunio Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," *Proc. Interspeech 2017*, pp. 1283–1287, 2017. 2

[13] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997. 2

[14] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014. 2

[15] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814. 3

[16] Akira Kurematsu, Kazuya Takeda, Yoshinori Sagisaka, Shigeru Katagiri, Hisao Kuwabara, and Kiyohiro Shikano, "Atr japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990. 3

[17] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999. 3

[18] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 3