On Unsupervised Acoustic Unit Discovery using Hashing VAEs

SaiKrishna Rallabandi, Wenting Ye, Elizabeth Salesky, Steven Hillis and Alan W Black

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA

srallaba,wye2,esalesky,shillis,awb@andrew.cmu.edu

Abstract

In this paper, we present an approach to automatically discover acoustic-phonetic units from a speech utterance in an unsupervised fashion. We first present an analysis to show that incorporating latent random variables into the neural generative models using suitable priors allows us to control what gets encoded into the latent space. Based on this, we employ articulatory features as discrete prior bank in the latent space and obtain acoustic units that are speaker and language independent. Through experiments the proposed model achieves significantly better fidelity compared to the baseline model with a lower bit rate.

Index Terms: variational inference, disentanglement, dilated convolutions

1. Introduction

A major bottleneck in the progress of many data-intensive language processing tasks such as speech recognition and synthesis is scalability to new languages and domains. Building such technologies for unwritten or under-resourced languages is often not feasible due to lack of annotated data or other expensive resources. A fundamental resource required to build such a stack is a phonetic lexicon - something that translates acoustic input to textual representation. Having such a lexicon, even if noisy, can help bootstrap speech recognition models, synthesis, and other technologies. Typical approaches may involve a pivot language or bootstrapping or adapting from a closely related high-resource language. But, this can be a deceptively non-trivial task due to linguistic differences which can pose inherent difficulties. For instance, it may be unreasonable to analyze a Sino-Tibetan language using English as a source. Moreover, using an additional language might make the model learn unintended surface level associations or biases between the participating languages that prevent them from generalizing across languages. Associations between these languages over a set of units that may better generalize to other languages. Therefore, in this paper we are interested in discovering the appropriate acoustic phonetic units.

In ZeroSpeech Challenge[1] resynthesis is considered a good proxy task to evaluate the performance of systems when training using unsupervised approaches. To accomplish this we use neural generative models. Deep Neural Generative models have seen a tremendous amount of progress in the recent past. These models aim to model the joint probability of the data distribution and the conditioning information as a product of conditional distributions. Typical implementations of such models follow an autoregressive framework, although other formulations have been suggested as well. Such models have been shown very effective in addressing one of the major challenges with conventional vocoding techniques - fidelity. Neural generative models has been shown to generate speech that rivals natural speech when conditioned on predicted mel spectrum [2].

Speech has a lot of natural variations in terms of content, speaker, channel information, speaking style, prosodic variations, etc. Accordingly, we are interested in models which have flexibility to marginalize such variations but preserve the phonetic content and distinguish meaningful differences between phonetic units. To accomplish this, we employ sequence to sequence models with latent random variables (referred to as latent stochastic models hereafter). These models provide a mechanism to jointly train both the latent representations as well as the downstream inference network. They are expected to both discover and disentangle causal factors of variation present in the distribution of original data, so as to generalize at inference time. While training latent stochastic models, optimizing the exact log likelihood can be intractable. To address this, a recognition network is employed to approximate the posterior probability using reparameterization [3]. When deployed in encoder-decoder models, this approach is often subject to an optimization challenge referred to as KL-collapse [4], wherein the generator (usually an RNN) marginalizes the learnt latent representation. Typical approaches to dealing this issue involve annealing the KL divergence loss [4, 5], weakening the generator [6] and ensuring the recall using bag of words loss.

In our work, we present an approach to deal with the KL-collapse problem by vector quantization in the latent space. Building on [7, 8], we add additional constraints in the prior space forcing the latent representations to follow articulatory dimensions: The encoded representation is hashed to a latent code based on a articulatory prior bank designed using a discrete codebook. Our decoder is a conditional WaveNet using speaker embedding as global embedding trained to regenerate input audio using the code sequence as local information.

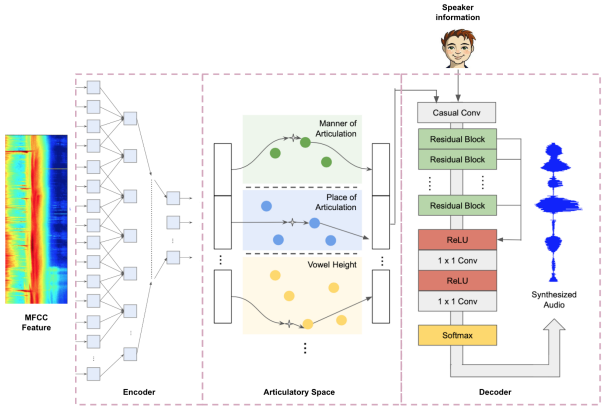

Figure 1: *Illustration of our procedure for automatically discovering acoustic units from a speech utterance. We pass the speech utterance through a downsampling encoder. The encoded representation is hashed to a latent code based on a discrete articulatory prior bank. The code is passed to the decoder, a WaveNet using speaker embeddings as global conditioning that regenerates audio.*

2. Background

2.1. Acoustic Unit Discovery

Let us consider a speech corpus X which consists of speakers $\{s_1, s_2..., s_n\}$. The goal of acoustic unit discovery is to come up with a set of units **U** that represent a speech utterance $x \subset$ X allowing robust resynthesis. The elements of such a set also might conform to desirable characteristics such as being injective, consistent and compact, i.e. that different inputs should have discriminant acoustic units, but expected variance such as speaker or dialect should produce the same acoustic units.

There have been numerous attempts to discover such acoustic units in an unsupervised fashion. In [9], authors presented an approach to modify the speaker diarization system to detect speaker-dependent acoustic units. [10] proposed a GMM-based approach to discover speaker-independent subword units. However, their system requires a separate Spoken Term Detector. Recently, due to the surge of deep generative model, using unsupervised method such as auto-encoder and variational auto-encoder (VAE). [11] designed a stacked AutoEncoder using backpropagation and then cluster the representations at the bottleneck layer. To avoid quick transitions leading to repeated units, they employed a smoothing function based on transition probabilities of the individual states. [12] extended the structured VAE to incorporate the Hidden Markov Models as latent model. [7, 8] proposed VQ-VAE and argue that by vector quantization the "posterior collapse" problem could be circumvented.

2.2. Disentanglement

There have been many attempts to interpret and manipulate the working of ELBO through analysis or exploitation of the latent space. For instance, in [13], authors decomposed the ELBO term and showed that there are terms measuring the total correlation between the latent variables. [14] introduced a generalization of ELBO by factorizing the latent representations into a hierarchy, while [15] presented an approach to accomplish disentanglement by modifying the co-variance matrix of the latent representations. Lastly, [16] augmented ELBO using the density ratio trick to accomplish disentanglement. Our work is similar to these in that we analyze ELBO to show that it is possible to control what gets disentangled.

Other works have focused on the prior distribution and causal factors, such as [17], which posited that to improve ELBO we must also improve the marginal KL e.g. we must have good priors, and [18], which showed that actively trying to disentangle the causal factors of variation is better than trying to pressure the model to forget the invariant representations. In [19], authors proposed incorporating a channel capacity term to promote disentanglement of these causal factors. We take inspiration from previous approaches that manipulate the prior distribution, but in our work, specifically incorporate articulatory constraints on the prior space. Doing so has additional benefits such as interpretation of the intermediate model outputs. Our implementation use an information bottleneck, which was shown in [20] to help models become robust to adversarial attacks as well. However, such analysis is beyond scope of the current study.

2.3. Neural Generative Models for Speech

Artificial generation of speech based on neural approaches has soared in the recent past. There have been continuous and significant improvements in both the aspects of speech generation - fidelity and flexibility. Autoregressive models such as [21], flow based models such as [22] have shown to generate audio that rivals the quality of natural speech. Approaches such as [23, 24] have shown ways to incorporate inductive biases into the generative process. [25] developed generic methods to enable the usage of distributional analysis of text at phone, word, and character levels in an unsupervised fashion. These techniques have been utilized in building highly flexible systems capable of generating different styles of speech and ability to build voices from noisy or very minimal data.

2.4. Speech Chain

There have been attempts to combine the ASR model and TTS system to form a closed-loop speech chain inspired by their closely dependent nature. [26] proposed the first deep sequence-to-sequence model in close-loop architecture allows us to train our model on the concatenation of both labeled and unlabeled data.

3. *VACONDA*[1] - *VA*riational inference based *CON*trolled *D*isentanglement using *A*rticulatory priors

3.1. Analysis of optimization and disentanglement

WaveNet [27] is an autoregressive neural model with a stack of 1D convolutional layers that is capable of directly generating audio signal. It has been shown to produce generated speech that rivals natural speech when conditioned on predicted mel spectrum [2]. The input to WaveNet is subjected to corresponding gated activations while passing through each dilated convolutional layer and is classified by the final softmax layer into a μ law encoding. The concrete form of the residual gated activation function is given by following equation:

$$r_d(x) = tanh(W_f * x) \odot \sigma(W_g * x) \tag{1}$$

where x and $r_d(x)$ are the input and output with dilation d, respectively. The symbol $*$ is a convolution operator with dilation d and the symbol $\odot$ is an element-wise product operator. W represents a convolution weight. The subscripts f and g represent a filter and a gate, respectively. The joint probability of a waveform **X** can be written as:

$$P(X|\theta) = \prod_{t=1}^{T} P(x_t|x_1, x_2..x_{t-1}, \theta) \tag{2}$$

given model parameters θ. During implementation of WaveNet, the autoregressive process is realized by a stack of dilated convolutions. The final output y_t at time step t can be expressed mathematically as:

$$\hat{y_t} \sim \sum_{d=0}^{D} h_d * r_d(x) \tag{3}$$

where x, y represent input and output vectors; D is the number of different dilation used and d is the dilation factor; h_d is the convolution weights. This stack of convolutions is repeated multiple times in the original WaveNet. Optimization

[1] Phonetically similar to its namesake 'Wakanda' from Marvel Comics

in WaveNet is performed based on the error between predicted sample and the ground truth sample conditioned on previous samples in the receptive field alongside the local conditioning. Expressing the loss function being optimized mathematically the error at sample t is:

$$l_t = Div(\hat{y_t}||y_t) \tag{4}$$

Here, we define the divergence similar to the [28], To optimize this loss, the contribution from the individual convolution layers towards this global error function must be nullified. Now let us consider the expression for intermediate output for a single filter in Eqn 3:

$$x_{out}(t) = \sum_{\tau=0}^{t} h(\tau)x(t-\tau) \tag{5}$$

where τ is the receptive field covered by the model and $h(\tau)$ represents the discrete state representation at time t. Without loss of generality and dropping the term τ for brevity, the spectral representation generated by the model can be expressed as:

$$Y(z) = H(z)X(z) \tag{6}$$

Considering the discrete nature of input from Eqn 4, an interpretation of Eqn 6 is that the neural autoregressive model acts as the transfer function and is discretized by convolving with the samples from original signal. It has to be noted that this is similar to the formulation of source filter model of speech, specifically the periodic components aka voiced sounds. Voiced sounds typically represented as impulse train are convolved with the transfer function to generate spectral envelope. As a corollary, from Eqn 4 and 6, we posit that the optimization in WaveNet model is performed by minimizing the divergence between true and approximate spectral envelope. Note that latent stochastic models such as VAEs are aimed to minimize the divergence between true and approximate posterior distributions of input data. The advantage with such models is the presence of stochastic random variables that capture the causal factors of variation in input based on some prior information about the distributional characteristics of data. Techniques aimed at this [29] have shown that it is possible to effectively disentangle the factors of variation using stochastic variables. Hence, we postulate that it should be possible to augment WaveNet decoder with a suitable encoder and an appropriate prior distribution to disentangle the acoustic phonetic units from a given utterance.

However, this is a deceptively non-trivial task. If the prior is too simplistic, such as unit normal distribution, the model is trivially incentivized to force the posterior distribution to closely follow the Gaussian prior distribution [30], particularly early in training. This results in the decoder marginalizing out the latent variable completely, manifesting in poor reconstruction ability. On the other hand, making the prior distribution arbitrarily complex also leads to unreasonable constraints on the decoder. For instance, in scenarios that have categorical distributions as their output (tasks such as language modeling, machine translation, and image captioning among others) it is unintuitive to assume that the true prior that generates latent distribution is a Gaussian when the likelihood is based on discrete sequential data. We make an observation that dealing with speech presents a characteristic advantage - speech has both continuous as well as discrete priors. The generative process of speech assumes a Gaussian prior distribution which is continuous in nature. However, the language which is also present in the utterance can be approximated to be sampled from a discrete prior distribution. Exact manifestation of this in the linguistics can be at different levels: phonemes, words, syllables, subword units, etc. From the analysis presented in the previous section, we hypothesize that if we use background knowledge about the data distribution while designing the priors, we can help the encoder effectively disentangle the latent causal factors of variation in the data. In other words, this presents us with an opportunity to control what gets disentangled in the latent space by appropriately choosing a prior distribution. Therefore, we engineer our prior space to account for the phonetic information in the utterance by representing the prior as a discrete latent variable bank, similar to the filterbanks used for feature extraction from speech. Each discrete latent variable has a different set of states reflecting one of the articulatory dimensions. The specific design of our latent space is highlighted in Table 1.

Table 1: *Articulatory Features*

Feature name	Value	Details
vc	+ - 0	vowel or consonant
vlng	s l d a 0	vowel length
vheight	1 2 3 0 -	vowel height
vfront	1 2 3 0 -	vowel frontness
vrnd	+ - 0	lip rounding
ctype	s f a n l r 0	consonant type
cplace	l a p b d v g 0	place of articulation
cvox	+ - 0	consonant voicing
asp	+ - 0	consonant voicing
nuk	+ - 0	consonant voicing

4. Experiments

4.1. Dataset

4.1.1. ZeroSpeech 2019 dataset

ZeroSpeech Challenge 2019: TTS without T is to propose to build a speech synthesizer without any text or phonetic labels [31, 32, 1]. The systems are required to extract the symbolic representation of the raw audio, and then re-synthesize the audio using these discovered units. There are three datasets in total: (1) *Unit Discovery Dataset* provides audio from a variety of speakers and is used to unsupervised acoustic modeling, (2) *Voice Dataset* provides audio from the targeted speaker and is used for synthesizer modeling and (3) *Parallel Dataset* is intended for finetuning both the sub-systems. We have not utilized the parallel dataset for our observations in this study. The development language is English and the test language is Standard Indonesian. The system is constrained to not use any pre-existing resource or models. To ensure that the model generalizes out of the box, the hyperparameter will be fine-tuned only on the development dataset, and the model will be trained in test language under the same parameters.

4.2. Baseline System

We have a three-stage pipeline: (1) *Unit Discovery*: We hypothesize acoustic units given a speech utterance using latent Stochastic Models; (2) *Unit Alignment*: We fine-tune the alignment between the utterance and the proposed acoustic units ; (3) *Unit Synthesis*: We build a speech synthesizer using the acous-

tic units and the target voice.

As proposed in [33], we take the initially discovered transcription of the acoustic units for our speech corpus and train an ASR model on it. Then we re-encode the corpus using the ASR model, and train a TTS system on it. Here we using Bi-LSTM with CTC loss as our ASR model, and tacotron [34] as TTS system.

4.3. VACONDA

The architecture of our model is built on top of VQ-VAE. It consists of three modules: an encoder, quantizer and a decoder. As our encoder, we use a dilated convolution stack of layers which downsamples the input audio by 64. The speech signal was power normalized and squashed to the range (-1,1) before feeding to the downsampling encoder. To make the training faster, we have used chunks of 2000 time steps. This means we get 31 timesteps at the output of the encoder. The quantizer acts as a bottleneck and performs vector quantization to generate the appropriate code from a parameterized codebook. We define the latent space $e \in R^{k \times d}$ to contain k d-dim continuous vector. Quantization is implemented using minimum distance in the embedding space. The number of classes was chosen to be 64, approximating 64 universal phonemes. We use a linear mapping to first project the 128 dimensional vector to 160 dimensions. We then perform comparisons with respect to individual articulatory dimensions each of which is 16 in size. Assuming $z_e(x)$ denotes the encoder output in the latent space, then the input of decoder $z_d(x)$ will be obtained by $\arg\min_j d(e_j, z_e(x))$, where d is a similarity function of two vectors. In this paper, we consider Euclidean distance as the similarity metric. Our decoder is an iterated dilated convolution-based WaveNet that uses a 256-level quantized raw signal as the input and the output from vector quantization module as the conditioning. Although using a Mixture of Logistics loss function might yield a better output, we have only used a 256 class softmax in this study. The decoder takes the output from the quantizer along with the speaker label as global conditioning and aims to reconstruct the input in an autoregressive fashion. Following IDCNNs, we have shared the parameters of all the stacks.

4.4. Analysis

In this section, we will discuss different design choices in the architecture, including input features and latent space constraints.

4.4.1. Acoustic Unit Discovery

Here we analyze the AUD performance of three different models in ZeroSpeech dataset as shown in Table 2. We only show the results in English since we don't have ground truth for the Indonesian language.

Table 2: *Performance of different systems in ZeroSpeech*

	English	
Model	ABX score	bitrate
Baseline	**27.46**	74.5
Three-stage Model	34.86	68.54
VACONDA	38	**58.19**

As in Table 2, the VACONDA achieves the best bit rate among three models. With such small number of unit, we could resynthesize and even convert the speech in a very high quality.

4.4.2. Speech Resynthesis and Conversion

The proposed model supports synthesizing the same speech in both the same speaker and a different speaker. Here we show a sample in the test dataset of Indonesian language in Figure 2. When we feed the decoder with the same speaker identification, the decoder will generate the original audio. Otherwise, it will perform speech convertion. The three audio shares similar structure. However, the converted audio has denser waveform, suggesting it's a different speaker. For the sampled audio, please visit the our website.

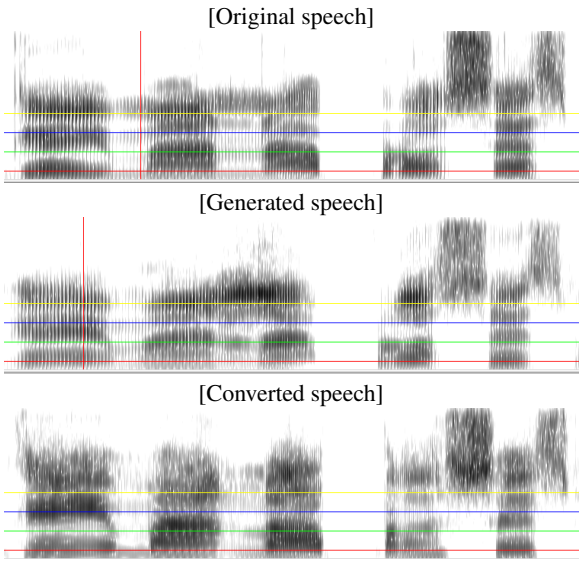

Figure 2: *Spectrograms of original, generated, and converted speech. The source speaker is female while the target speaker is male.*

5. Conclusion

In this paper, we present an approach to automatically discover acoustic-phonetic units from a speech utterance in an unsupervised fashion. We first present an analysis to show that incorporating latent random variables into neural generative models using suitable priors allows us to control what gets encoded into the latent space. Based on this, we employ articulatory features as a discrete prior bank in the latent space and obtain acoustic units that are speaker and language independent. To validate effectiveness of the discovered units, we perform discriminability tests as part of ZeroSpeech Challenge 2019.

6. References

[1] D. Ewan *et al.*, "Zerospeech 2019: Tts without t," in *Interspeech 2019*, 2019, pp. 3442–3446.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," *arXiv preprint arXiv:1712.05884*, 2017.

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

[4] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," *arXiv preprint arXiv:1511.06349*, 2015.

[5] C. Zhou and G. Neubig, "Multi-space variational encoder-decoders for semi-supervised labeled sequence transduction," *arXiv preprint arXiv:1704.01691*, 2017.

[6] T. Zhao, R. Zhao, and M. Eskenazi, "Learning discourse-level diversity for neural dialog models using conditional variational autoencoders," *arXiv preprint arXiv:1703.10960*, 2017.

[7] A. van den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 6306–6315.

[8] J. Chorowski, R. J. Weiss, S. Bengio, and A. v. d. Oord, "Unsupervised speech representation learning using wavenet autoencoders," *arXiv preprint arXiv:1901.08810*, 2019.

[9] M. Huijbregts, M. McLaren, and D. Van Leeuwen, "Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection," in *2011 IEEE international conference on Acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 4436–4439.

[10] A. Jansen, S. Thomas, and H. Hermansky, "Weak top-down constraints for unsupervised acoustic model training," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8091–8095.

[11] L. Badino, C. Canevari, L. Fadiga, and G. Metta, "An auto-encoder based approach to unsupervised learning of subword units," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7634–7638.

[12] J. Ebbers, J. Heymann, L. Drude, T. Glarner, R. Haeb-Umbach, and B. Raj, "Hidden markov model variational autoencoder for acoustic unit discovery." in *INTERSPEECH*, 2017, pp. 488–492.

[13] 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*, 2018, pp. 2615–2625.

[14] B. Esmaeili, H. Wu, S. Jain, A. Bozkurt, N. Siddharth, B. Paige, D. H. Brooks, J. Dy, and J.-W. van de Meent, "Structured disentangled representations," *stat*, vol. 1050, p. 12, 2018.

[15] A. F. Ansari and H. Soh, "Hyperprior induced unsupervised disentanglement of latent representations," *arXiv preprint arXiv:1809.04497*, 2018.

[16] H. Kim and A. Mnih, "Disentangling by factorising," *arXiv preprint arXiv:1802.05983*, 2018.

[17] M. D. Hoffman and M. J. Johnson, "Elbo surgery: yet another way to carve up the variational evidence lower bound," in *Workshop in Advances in Approximate Bayesian Inference, NIPS*, 2016.

[18] E. Banijamali, A.-H. Karimi, A. Wong, and A. Ghodsi, "Jade: Joint autoencoders for dis-entanglement," *arXiv preprint arXiv:1711.09163*, 2017.

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

[20] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *arXiv preprint arXiv:1612.00410*, 2016.

[21] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR abs/1609.03499*, 2016.

[22] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," *arXiv preprint arXiv:1811.00002*, 2018.

[23] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, "Voiceloop: Voice fitting and synthesis via a phonological loop," *arXiv preprint arXiv:1707.06588*, 2017.

[24] ——, "Voice synthesis for in-the-wild speakers via a phonological loop," *arXiv preprint arXiv:1707.06588*, pp. 1–11, 2017.

[25] O. Watts, "Unsupervised learning for text-to-speech synthesis," Ph.D. dissertation, University of Edinburgh, 2012.

[26] A. Tjandra, S. Sakti, and S. Nakamura, "Listening while speaking: Speech chain by deep learning," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 301–308.

[27] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[28] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, "Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications," *arXiv preprint arXiv:1701.05517*, 2017.

[29] 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," 2016.

[30] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel, "Variational lossy autoencoder," *arXiv preprint arXiv:1611.02731*, 2016.

[31] S. Sakti, R. Maia, S. Sakai, T. Shimizu, and S. Nakamura, "Development of hmm-based indonesian speech synthesis," in *Proc. Oriental COCOSDA*, 2008, pp. 215–219.

[32] S. Sakti, E. Kelana, H. Riza, S. Sakai, K. Markov, and S. Nakamura, "Development of indonesian large vocabulary continuous speech recognition system within a-star project," in *Proceedings of the Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST)*, 2008.

[33] S. Sitaram, S. Palkar, Y.-N. Chen, A. Parlikar, and A. W. Black, "Bootstrapping text-to-speech for speech processing in languages without an orthography," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7992–7996.

[34] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in Neural Information Processing Systems*, 2018, pp. 4480–4490.