INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Variational Attention using Articulatory Priors for generating Code Mixed Speech using Monolingual Corpora

SaiKrishna Rallabandi and Alan W Black

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
`{srallaba, awb} @ cs.cmu.edu`

Abstract

Code Mixing - phenomenon where lexical items from one language are embedded in the utterance of another - is relatively frequent in multilingual communities and therefore speech systems should be able to process such content. However, building a voice capable of synthesizing such content typically requires bilingual recordings from the speaker which might not always be easy to obtain. In this work, we present an approach for building mixed lingual systems using only monolingual corpora. Specifically we present a way to train multi speaker text to speech system by incorporating stochastic latent variables into the attention mechanism with the objective of synthesizing code mixed content. We subject the prior distribution for such latent variables to match articulatory constraints. Subjective evaluation shows that our systems are capable of generating high quality synthesis in code mixed scenarios.

Index Terms: Code Mixing, Bilingual speech, Variational Auto Encoder

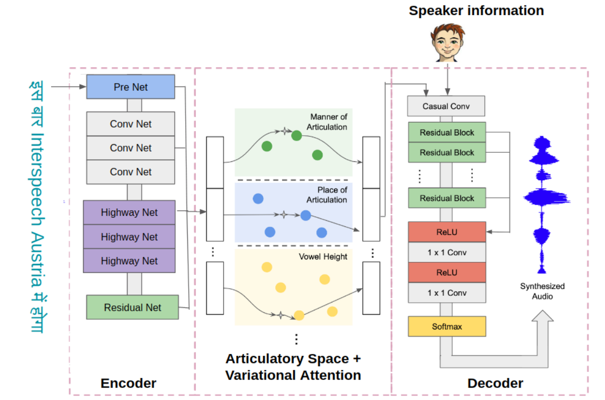

Figure 1: *Illustration of our procedure for generating a code mixed utterance. Text from different languages is converted into a common representation space by Tacotron encoder. The encoded representation is hashed to a latent code based on a discrete articulatory prior bank. The code is passed to the decoder, followed by a WaveNet using speaker embeddings as global conditioning that generates audio.*

1. Introduction

Code Mixing is a phenomenon where linguistic units such as phrases, words and morphemes of one language are embedded into an utterance of another language [1, 2]. This is quite common in multilingual societies such as in India where English has transitioned from the status of a foreign language to that of a second language. Today such mixing has manifested itself in various types of text ranging all the way from news articles through comments/posts on social media, leading to co-existence of multiple languages in the same sentence. In the context of Text to Speech (TTS), voice deployed in such contexts has to be able to synthesize mixed text without ignoring the content from one of the languages. Typical approaches for building such mixed lingual voices require bilingual recordings[3, 4, 5]: speech data from the speaker in both native language as well as the additional language. However, obtaining such data might not always be feasible. On the other hand, social media and web 2.0 has enabled an outburst of audiovisual content at an unprecedented rate. Therefore, it might be useful to design techniques that can leverage such resources. In this paper, we present initial steps in that direction.

We investigate training strategies for building code mixed voices subject to the availability of only monolingual data in participating languages. Specifically, we concern ourselves with two scenarios: (1) Mixing in the case of a sentence which is primarily Indic but interspersed with English words. Such sentences are found as a newspaper headlines (Ex: *Microsoft ki mobile devices unit ne apni nayee smart phone Lumia 640 aur uske badee screen wali variant 640 par se parda utha liya hai.*) (2) Mixing in the case of a sentence which is primarily English but has some Indic words. Such sentences are found as navigation instructions (Ex: *Proceed for 100 meters and then take a left at Sarojini Naidu Nagar Road, heading onto the Ballary chowrasta.*) Although building voices using such a combination of multilingual corpora appears as a simple extension of multispeaker or multilingual speech synthesis, generating code mixed content is a deceptively non trivial task since there is a mismatch between training and the testing scenarios: Even though the model has access to data from both the participating languages during training, code mixed content it is exposed to at test time - as seen from the example sentences - is a novel composition of linguistic units from both the languages. To assist the model in dealing with such mismatch, we incorporate latent stochastic variables into the training procedure.

Models with latent random variables (referred to as latent stochastic variable models hereafter) provide flexibility to jointly train the latent representations as well as the downstream 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. However, while training latent stochastic variable models, optimizing the exact log likelihood can be intractable. To address this, a recognition network is employed to approximate the posterior probability using reparameterization [6]. We make an observation that articulatory information about speech production presents a discrete set of independent constraints. For instance, manner and place of articulation are two articulatory dimensions characterized by discrete sets(labial vs dental, etc). Based on this, we condition the recognition network in latent stochastic variable

http://dx.doi.org/10.21437/Interspeech.2019-1103

models to conform to articulatory prior space by using a bank of discrete prior distributions. We show that such priors help encode language independent information thereby facilitating synthesis of code mixed content.

2. Background

2.1. Synthesis of Code Mixed Text

Synthesis of code mixed text using monolingual data [7, 5] has been addressed primarily at the linguistic level: by either mapping the words/phones of the foreign language with the closest sounding phones of the native language or by using transliteration [8, 9]. However, these methods have been shown to generate foreign accents [10, 11, 12]. In our work, we borrow the central idea from the works - the requirement of a common linguistic space - and apply this as a constraint on the representations learnt in our latent stochastic variable model. In [13], the authors follow a two step procedure to address the issue with accented speech. They first warp the source speakers' speech parameter trajectories (in L1) towards the target speaker and then 'tile' them with the data (in L2) to form a pseudo training corpus which is subsequently used to train a bilingual speech synthesis system. Similar practices can be found in the literature for voice adaptation [14, 15, 16, 17, 18] and voice conversion [19]. Although we do not explicitly aim to transfer acoustic parameters, our decoder is engineered to work with a global speaker embedding that learns speaker specific information. Therefore, our approach can be seen as analogous to these works. Our work is closest to [20, 21] in that we use monolingual recordings. However, we explicitly work in the latent prior space while [20] operate at the level of encoding individual languages and [21] begin with an average voice and refine it using phoneme informed attention.

2.2. Disentanglement

In [22], authors decompose Evidence Lower Bound (ELBO) and show that there are terms measuring the total correlation between the latent variables. In [23], authors propose incorporating a channel capacity term to promote disentanglement of causal factors of variation in the data. Our work is similar to these in that we analyze ELBO to show that it is possible to control what gets disentangled. In [24], authors present a generalization of ELBO by factorizing the latent representation into a hierarchy. In [25], authors present an approach to accomplish disentanglement by modifying the co-variance matrix of the latent representations. In [26] authors augment ELBO using the density ratio trick to accomplish disentanglement. In [27], authors posit that to improve ELBO we must also improve the marginal KL, meaning we must have good priors. In [28] authors show that actively trying to disentangle the causal factors of variation is better than trying to pressurize the model to forget the invariant representations. We take inspiration from these approaches that manipulate the prior distribution and impose domain specific constraints - based on intuitions from articulatory features - on the prior space. Manipulating prior space has other benefits such as interpreting the intermediate stage outputs of the model. However, such analysis is beyond scope of the current study.

3. Proposed Approach

In this section, we first present the differences between soft attention and variational attention. We then highlight the role of priors in latent stochastic variable models. Based on these analyses, we present our proposed approach.

3.1. Soft Attention vs Variational Attention in Seq2Seq TTS

Let us consider a speech corpus X consisting of languages $\{l_1, ..., l_n\}$, where each l_i might comprise of multiple speakers. Let y_1,...,y_n denote acoustic frames in the target sequence y while x_1,...,x_n denote the encoded text sequence x from one of the languages. A typical attention based encoder decoder network such as Tacotron factorizes the joint probability of acoustic frames as product of conditional probabilities. Mathematically, this can be shown as below:

$$P(y|x) = \Pi_{t=1}^{t=n} P(y_t|x_1...x_m, s_t) \tag{1}$$

where s_t is a decoder state summarizing y_1,...y_{t-1}. Parameters θ of the model are set by maximizing either the log likelihood of training examples or the divergence between predicted and true target distributions. At each time step t in these models, an attention variable a_t is used to denote which encoded state of $x_1...x_m$ aligns with y_t. The most common form of attention used is soft attention, a convex combination from encoded representation of input text. It has to be noted that soft attention in such scenarios is essentially a latent deterministic variable that computes an expectation over the alignment between input and output sequences. Empirically, soft attention provides surprisingly good alignment often correlating with human intuitions. Having said that, to synthesize code mixed speech at test time, the generative process needs to disentangle appropriate individual language attributes from observed data X_{obs} and also compose them to form a coherent utterance in the voice of desired speaker. However, presence of deterministic alignment method limits the ability of models to generalize to such scenario.

On the other hand, variational attention[29] provides a mechanism to factorize this alignment and mediate the generative process of y through a stochastic variable z. In addition, both soft and hard attention mechanisms can be shown as special cases of ELBO[29]. Therefore, incorporating latent stochastic variables allows us to directly optimize ELBO. In this context, model parameters are set by maximizing the log marginal likelihood of the training samples. But direct maximization of this marginal in the presence of latent variable is often difficult due to expectation involved. To address this, a recognition network q is employed to approximate the posterior probability using reparameterization. It is interesting to note that the encoder in a deterministic Seq2Seq network functions as the recognition network in latent stochastic variable models and is incentivized to search over variational distributions to improve ELBO. Intuitively, the lower bound is tight when the inferred variational distribution is closer to the true posterior of the data. In this paper, we make an assumption that the true posterior of speech distribution is governed by the articulatory properties of speech. Based on this insight, we constrain the prior distribution to model the articulatory space.

3.2. Role of Priors in Latent Stochastic Variable Models

The choice of priors plays a significant role in optimization within latent stochastic models. In this subsection, we present an analysis to show that priors control the disentanglement of causal factors of variation in such models. Let us consider the ELBO being optimized:

$$E_{q_\phi(z|x,c)}[log p_\theta(x|c,z)] - |D_{KL}(q_\phi(z|x,c)||p_\theta(z|c))| \tag{2}$$

where the first term is the reconstruction error while the second is the divergence between approximate and true posteriors. Here are the four phenomenon that are manifested due to choices of priors:

(1) *Disentanglement or Factorization of causal factors of variation*
The KL divergence forces the posterior distribution output by encoder to follow an appropriate prior about the data generation process. Typically, prior space is assumed to be continuous distribution and a unit Gaussian. The global optimum value for the divergence in such cases is 0 and is reached only when both the distributions exactly match each other. Since the prior information about the data generation process typically involves some causal factors of variation of the data, this naturally is assumed to translate to a constraint on the encoder to track such factors. Thus, such models have potential to disentangle or factorize the causal factors of variation in the distribution.

(2) *Marginalization of Nuisance Factors of Variation*
It has to be noted that during training optimization is performed in expectation over minibatches. Therefore, the expectation of KL divergence can be rewritten as related to the amount of mutual information between the latent representation and the data distribution [30]. As this divergence decreases, the amount of information the encoder can place in the latent space also decreases. As a result, encoder is forced to discard some nuisance factors that may not have contributed to the generation of data. Thus, KL divergence also forces the model to marginalize the nuisance variables.

(3) *Posterior Collapse due to simple priors*
Consider the scenario where the prior is too simplistic, such as the aforementioned unit normal distribution. In such cases, the model is incentivized to force the posterior distribution to closely follow the Gaussian distribution [31]. Typically the decoders in variational models are implemented using universal approximators such as RNNs. In the context of a TTS systems, decoder segment of the acoustic model along with the neural vocoder act as the decoders. Since such decoders are very powerful, they are able to learn or ignore the priors about data distribution themselves and hence marginalize out the latent representation input from the encoder. In other words, the prediction of next sample is based solely on the marginal distribution at the current timestep which can be implemented by learning a dictionary per time step. Therefore, the encoder is no longer forced to track the causal factors of variation in the data. This is referred to as posterior collapse or mode collapse.

(4) *Loss of output fidelity due to complex priors*
A reasonable and intuitive solution to posterior collapse is making the prior space more complex thereby pressurizing the posterior distribution to track the prior space more closely. For instance, [32] attempt to accomplish this by adding a hyperparameter β to promote disentanglement and gradually increasing channel capacity, something that increases loss. However, it has to be noted that simply making the prior distribution arbitrarily complex also perhaps leads to unreasonable constraints on the decoder. For instance, in scenarios that have categorical distribution as their output (tasks such as language modeling, machine translation, 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 in such tasks. Having such strong priors directly affets the reconstruction ability in these models.

Therefore, priors in latent stochastic models play a significant role in the optimization and facilitate disentanglement of causal factors of variation on the one hand, as well as help the ability of the model to reconstruct the data distribution on the other. In this paper, we engineer the prior space to follow articulatory constraints. Since the articulatory features can be considered independent of language, we believe they facilitate encoder in disentangling language universal information important to compose code mixed utterance at the test time. As there are multiple articulatory dimensions (place vs manner), we implement them using a bank of discrete distributions as opposed to a single continuous/discrete distribution. We believe that this choice makes the prior space sufficiently complex to prevent posterior collapse while still being tractable for high fidelity output. In the following section, we explain the details of our approach.

3.3. *VACONDA*[1] - *Va*riational *A*ttention based *CON*trolled *D*isentanglement using *A*rticulatory priors

Table 1: *Articulatory Features*

Feature name	Possible Classes	Cardinality
vowel or consonant	+ - 0	3
vowel length	s l d a 0	5
vowel height	1 2 3 0 -	5
vowel frontness	1 2 3 0 -	5
lip rounding	+ - 0	3
consonant type	s f a n l r 0	7
place of articulation	l a p b d v g 0	8
consonant voicing	+ - 0	3

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 linguistics can be at different levels: phonemes, words, syllables, sub word units, etc. From the analysis presented in previous subsections, we posit that it helps encoder effectively disentangle the latent causal factors of variation if we use background knowledge about the data distribution while designing the priors. In other words, incorporating appropriated priors provides us with an opportunity to control what gets disentangled (or) decomposed (or) factorized in the latent space. In our context, an appropriate requirement from the encoder is to generate language agnostic yet phonetic representations such that a speaker dependent decoder can synthesize code mixed content. Therefore, we engineer our prior space to account for phonetic information in the utterance by representing the prior as a discrete latent variable bank, similar to 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 the table 1. Voice building procedure with these priors is depicted in figure 1. We have used the articulatory dimensions according to the definitions in Indic voice building process of [33]. Although some of them might be redundant, for this initial study we have retained all the articulatory dimensions. Without loss of generality, we assume

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

that the individual latent articulatory dimensions are independent of each other. The divergence between the true prior and approximate prior now becomes:

$$D_{KL}(q_{\phi}(z_{enc}|p)||p(z_{code})) = \sum_{i=1}^{N}[$$
$$E_{q_{\phi}(z^{i}_{enc}|p)}[logq_{\phi}(z^{i}_{enc}|p)] - E_{q_{\phi}(z^{i}_{enc}|p)}[logp(z^{i}_{code})]]$$

where N is the number of articulatory dimensions and i denotes the index of individual articulatory dimensions. z_{code} denotes the parameterized codebook and z_{enc} denotes the representation output by the encoder.

4. Experiments

4.1. Data

We have used speech and text data from three Indian languages Hindi, Telugu and Marathi released as a part of resources for Indian languages [34] to build our synthesis systems. From our baseline voice building process, we found male speaker from Hindi to be the most reliable voice in terms of quality. Therefore, all of our systems use English recordings from Mono segment of this speaker as English set - as a scaffolding. For other two languages, we use only monolingual data from the speakers. In other words, to generate code mixed Telugu sentence, the systems have access to English content but from a different speaker. As baseline for comparison, we have built a CLUSTERGEN voice using monolingual recordings employing phone mapping. Evaluation was performed in the form of listening tests with 20 native students following the convention of Blizzard Challenge evaluations using [35] with naturalness as criterion in terms of Mean Opinion Score (MOS) on a scale of 1(least natural) to 5(highly natural). All the listening tests involved test sentences generated using the Multilingual test set (ML) from [36]. The evaluation results are depicted in table 2.

4.2. Implementation Details

We have built two systems employing variational attention: VQTacotron with vanilla vector quantization and VACONDA - with articulatory prior on the latent space. The architecture of our models continues from [37], with some modifications. We have used WaveNet[38] as our decoder. Following [39], we have shared the parameters of all the residual layers with common dilation factors. We use Mixture of Logistics loss to train the model and the number of logistics was set to 10. Speech signal was power normalized and squashed to the range (-1,1). To make the training faster, we have used chunks of 8000 time steps. Our quantizer performs vector quantization to generate the appropriate code from a parameterized codebook. We define the latent space $e \in R^{k \times d}$ contains k d-dim continuous vector. Quantization is implemented using minimum distance in the embedding space. We have used 128 dimensions to perform the comparison in system VQTacotron. The number of classes was chosen to be 64, approximating 64 universal phonemes. For system VACONDA, we use a linear mapping to first project the 128 dimensional vector to 160 dimensions. We then perform comparison with respect to individual articulatory dimensions each of which is 16 in size. The speaker embedding is shared between the decoder of our acoustic model and WaveNet. We have noticed the lengths of utterances in the Indic datasets being too big to train attention from scratch. Therefore we have initialized attention using alignments performed within Festvox using HMM aligner. All the models were built at phone level since that was observed to be the most stable configuration even though our phones do not cover all the variants (ex. we do not have explicit phones for geminates). We have used quantization penalty and commitment loss terms as mentioned in [40]. In addition, we have also normalized each latent embedding vector to be on a unit sphere.

Table 2: *MOS Scores for Naturalness in prosodic modeling based experiments*

Config	Clustergen	VQTacotron	VACONDA
Hi-Eng (Male)	3.9	**4.31**	4.28
Tel-Eng(Female)	3.6	3.9	**4.1**
Mar-Eng(Male)	3.7	4.0	**4.0**
Mar-Eng(Female)	3.4	3.9	**4.0**

4.3. Observations

An informal analysis on the outputs from the proposed systems revealed that the characteristics of the English speaker were retained in certain areas within the utterance, resulting in a slightly stylized version[2]. We want to investigate this further and hope to uncover techniques that can provide more control. While most of the systems using CLUSTERGEN [4] make errors in the prosodic features such as irregular duration shifts at the boundaries between languages, the proposed approaches have smooth transitions at the boundaries. However, we have observed marked differences in the pronunciations by the proposed approaches. For instance, the phone 'S' from the word 'Stanford' when heard in isolation is indistinguishable from other fricative sounds. Since we specifically deal with articulatory priors in VACONDA, a reasonable assumption to make is that this issue will be bypassed by the model. However, this characteristic is common across voices built using both VQTacotron as well as VACONDA.

5. Conclusion

In this paper, we investigated approaches to build mixed-lingual speech synthesis systems based on separate recordings and present systems at three different levels. Specifically we present a way to incorporate stochastic latent variables into attention mechanism. We subject the latent variables to match articulatory constraints. Subjective evaluation shows that our systems are capable of generating high quality synthesis in code mixed scenarios. From evaluations, we have identified interesting issues specific to the proposed approaches and different from errors observed in any of the previous methods. We are investigating them as an ongoing work and hope to understand them as well as formulate better techniques to handle the code mixed text.

6. Acknowledgements

We thank AFRL for funding this research, student volunteers for taking part in the listening evaluations and reviewers for such valuable feedback.

[2] The samples can be found here `http://www.cs.cmu.edu/~srallaba/IS2019_CodeMixedTTS/`.

7. References

[1] P. Muysken, *Bilingual speech: A typology of code-mixing*. Cambridge University Press, 2000, vol. 11.

[2] S. Gella, K. Bali, and M. Choudhury, "ye word kis lang ka hai bhai? testing the limits of word level language identification," in *International Conference on Natural Language Processing (ICON)*, 2014.

[3] C. Traber, K. Huber, K. Nedir, B. Pfister, E. Keller, and B. Zellner, "From multilingual to polyglot speech synthesis," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[4] S. Rallabandi and A. W. Black, "On building mixed lingual speech synthesis systems," in *Proceedings of Interspeech*, 2017.

[5] K. R. Chandu, S. K. Rallabandi, S. Sitaram, and A. W. Black, "Speech synthesis for mixed-language navigation instructions," in *Proceedings of Interspeech*, 2017.

[6] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations*, 2014.

[7] N. K. Elluru, A. Vadapalli, R. Elluru, H. Murthy, and K. Prahallad, "Is word-to-phone mapping better than phone-phone mapping for handling english words?" in *ACL*, 2013.

[8] S. Sitaram, S. K. Rallabandi, S. Rijhwani, and A. Black, "Experiments with cross-lingual systems for synthesis of code-mixed text," in *9th ISCA Speech Synthesis Workshop*, 2015, pp. 76–81.

[9] S. Sitaram and A. W. Black, "Speech synthesis of code-mixed text." in *LREC*, 2016.

[10] L. M. Tomokiyo, A. W. Black, and K. A. Lenzo, "Foreign accents in synthetic speech: development and evaluation." in *Proceedings of Interspeech*, 2005.

[11] N. Campbell, "Talking foreign," in *Proceedings of Eurospeech*, 2001.

[12] L. Badino, C. Barolo, and S. Quazza, "Language independent phoneme mapping for foreign tts," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[13] Y. Qian, J. Xu, and F. K. Soong, "A frame mapping based hmm approach to cross-lingual voice transformation," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.

[14] A. Kain and M. Macon, "Personalizing a speech synthesizer by voice adaptation," in *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998.

[15] M. Kurimo *et al.*, "Personalising speech-to-speech translation in the emime project," in *Proceedings of the ACL System Demonstrations*, 2000.

[16] K. Oura, K. Tokuda, J. Yamagishi, S. King, and M. Wester, "Unsupervised cross-lingual speaker adaptation for hmm-based speech synthesis," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010.

[17] J. Yamagishi *et al.*, "Robust speaker-adaptive hmm-based text-to-speech synthesis," in *IEEE TASLP*, 2010.

[18] J. Latorre, K. Iwano, and S. Furui, "Polyglot synthesis using a mixture of monolingual corpora," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[19] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan, "Text-independent voice conversion based on unit selection," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006.

[20] Y. Cao, X. Wu, S. Liu, J. Yu, X. Li, Z. Wu, X. Liu, and H. Meng, "End-to-end code-switched tts with mix of monolingual recordings," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[21] L. Xue, W. Song, G. Xu, L. Xie, and Z. Wu, "Building a mixed-lingual neural tts system with only monolingual data," *arXiv preprint arXiv:1904.06063*, 2019.

[22] 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.

[23] 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.

[24] 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," in *International Conference on Artificial Intelligence and Statistics*, 2018.

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

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

[27] M. D. Hoffman and M. J. Johnson, "Elbo surgery: yet another way to carve up the variational evidence lower bound," in *NeurIPS*, 2016.

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

[29] Y. Deng, Y. Kim, J. Chiu, D. Guo, and A. Rush, "Latent alignment and variational attention," in *Advances in Neural Information Processing Systems*, 2018.

[30] A. Makhzani and B. J. Frey, "Pixelgan autoencoders," in *Advances in Neural Information Processing Systems*, 2017.

[31] 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.

[32] 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.

[33] A. W. Black, "Clustergen: a statistical parametric synthesizer using trajectory modeling." in *Proceedings of Interspeech*, 2006.

[34] A. Baby, "Resources for Indian languages," in *CBBLR workshop, International Conference on Text, Speech and Dialogue*, 2006.

[35] A. Parlikar, "TestVox: web-based framework for subjective evaluation of speech synthesis," *Opensource Software*, 2012.

[36] K. Prahallad, A. Vadapalli, S. Kesiraju, H. Murthy, S. Lata, T. Nagarajan, M. Prasanna, H. Patil, A. Sao, S. King *et al.*, "The blizzard challenge 2014," in *Proceedings of Blizzard Challenge workshop*, 2014.

[37] P. Baljiker, S. K. Rallabandi, and A. Black, "An investigation of convolution attention based models for multilingual speech synthesis of indian languages," in *Proceedings of Interspeech*, 2018.

[38] A. Van Den Oord *et al.*, "Wavenet: A generative model for raw audio," *CoRR abs/1609.03499*, 2016.

[39] E. Strubell *et al.*, "Fast and accurate entity recognition with iterated dilated convolutions," in *ACL*, 2017.

[40] 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.