

USING VAES AND NORMALIZING FLOWS FOR ONE-SHOT TEXT-TO-SPEECH SYNTHESIS OF EXPRESSIVE SPEECH

Vatsal Aggarwal, Marius Cotescu, Nishant Prateek, Jaime Lorenzo-Trueba, and Roberto Barra-Chicote

Amazon, Cambridge UK

ABSTRACT

We propose a Text-to-Speech method to create an unseen expressive style using one utterance of expressive speech of around one second. Specifically, we enhance the disentanglement capabilities of a state-of-the-art sequence-to-sequence based system with a Variational AutoEncoder (VAE) and a Householder Flow. The proposed system provides a 22% KL-divergence reduction while jointly improving perceptual metrics over state-of-the-art. At synthesis time we use one example of expressive style as a reference input to the encoder for generating any text in the desired style. Perceptual MUSHRA evaluations show that we can create a voice with a 9% relative naturalness improvement over standard Neural Text-to-Speech, while also improving the perceived emotional intensity (59 compared to the 55 of neutral speech).

Index Terms— Text-to-speech, data efficiency, and semi-supervised learning

1. INTRODUCTION

Neural Text-to-Speech models, such as Tacotron and WaveNet [1, 2], are able to produce high quality speech with naturalness close to that of real humans. However, these models require several hours of data to produce high-quality voices.

This work focuses on creating high-quality expressive speech using only one utterance of a target style without requiring retraining. We show that we are able to create high-quality expressive voices at low cost and improve the naturalness of our baseline *state-of-the-art* neutral text-to-speech system. Our approach relies on an architecture combining Householder Flows and Variational AutoEncoders (VAE) in a novel way that provides better KL-Divergence (KLD) between an isotropic Gaussian prior and our VAE latent variables, alongside better reconstruction than existing methods. Reductions in KLD provide lower correlation between our latent variables. Consequently, our approach provides improved manipulation of the latents [3], substantial improvements in the disentanglement of the latent space, and raises perceptual metrics of synthesized expressive speech whilst using only one reference utterance.

To the authors knowledge, this work is the first to tackle the problem of one-shot synthesis of an unseen expressive

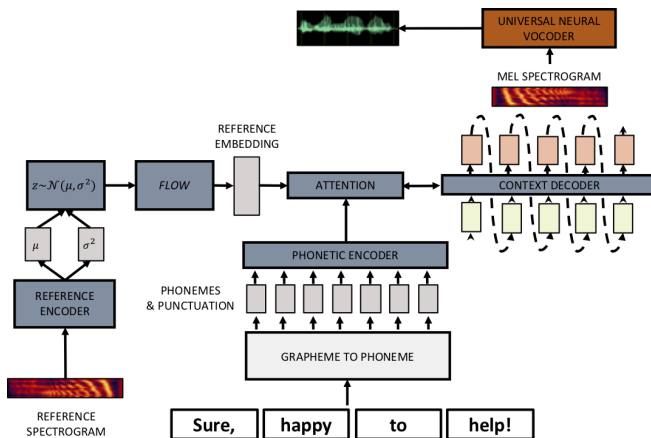


Fig. 1: Overview of the system’s architecture: a Variational seq-to-seq auto-encoder conditioned on the phonetic sequence. The Universal Neural Vocoder translates the generated spectrogram into the final speech signal.

style. Although there is existing work in the related fields of prosody generation [4] and prosody transfer [5], these approaches rely on either modelling expressive speech or transferring a seen expressive style to an unseen speaker, and are trained on a large corpora of expressive speech. In contrast, we use only one utterance to create an unseen style. Our problem is also distinct from that of speaker-adaptation [6, 7], as we wish to improve expressiveness by including emotion for existing speakers, not learn entirely new speakers. Existing work on emotional speech focuses on explicit emotional speech models [8], model adaptation [9] or voice conversion [10, 11], which again require significantly more data than the single utterance required by our proposed system.

2. PROPOSED APPROACH

For our proposed architecture, we extend a *state-of-art* sequence-to-sequence NTTS [2, 12, 13] system. This base seq2seq model uses an attention module to map the concatenated outputs of an RNN-based Phoneme Encoder and a mel-spectrogram conditioned Reference Encoder to mel-

spectrograms via an autoregressive decoder. The reference encoder is trained as the aggregate posterior of a VAE [14], and is conditioned on mel-spectrograms to predict the mean and variance of a 64-dimensional diagonal Gaussian Distribution (with similar architecture to [4]). Thus, the reference encoder and the decoder form a text-conditioned VAE, with the Phoneme Encoder providing text-conditioning to our network. We consider this as our baseline system, referred to as Vanilla VAE.

Our proposed system extends this base model by adding a Householder Normalizing Flow [15] (as presented in Figure 1) to the text-conditioned VAE. Householder Flows consist of multiple steps of easily invertible affine transformations that transform samples from a diagonal Gaussian to samples from a full co-variance Gaussian. We hypothesise that Householder Normalizing Flows allow samples from the reference encoder to better match the diagonal normal Gaussian prior while allowing the samples after the flow to become more co-related. We believe that this allows for both improved disentanglement and better reconstruction by the decoder. To investigate this claim, we tried a variety of architectures differing only in how the Householder vectors used to define each step of the householder flow are predicted. We now summarize our different architectures (see also Figure 2).

- Architecture-1: We follow the original implementation of Householder Flows [15] - predicting the first householder vector as an additional output of the reference encoder, with all other vectors predicted via subsequent affine transformations.
- Architecture-2: We predict all the Householder vectors as separate outputs of the reference encoder.
- Architecture-3: We propose a novel architecture where Householder vectors are globally shared across all utterances. This architecture is distinct from all existing implementations of Householder Flows, as the vectors are not conditioned on the input to the VAE encoder in any way (e.g. Architecture-1).

Each architecture is evaluated using 2, 4, 8 and 16 householder vectors as hyperparameters. Across all our architectures, samples from the output of the flow are concatenated with the output of the phoneme encoder instead of samples from the output distribution of the reference encoder.

During training, we use a large training corpus (Section 4) and feed the target spectrogram into the reference encoder. At synthesis time, we instead use a spectrogram from the 1-shot data-set (also described in Section 4) as input to the reference encoder and feed phonemes of the target text into the phoneme encoder. To complete the TTS pipeline, we use the vocoder of [16] to convert the mel-spectrograms to audio waveforms.

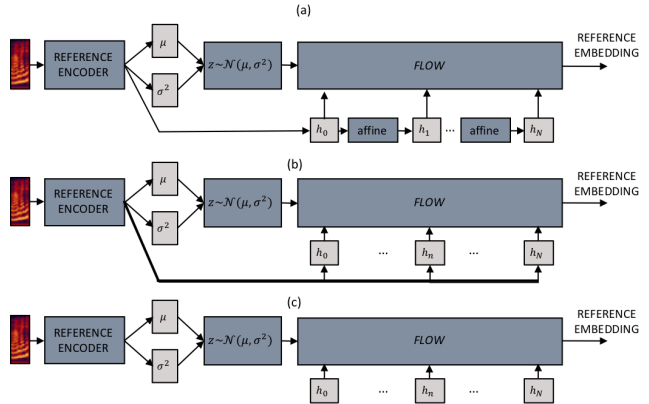


Fig. 2: Diagram of proposed architectures: Architecture-1 (a), Architecture-2 (b), and Architecture-3 (c). h_i refers to the i^{th} householder vector.

3. CORPORA

We train our models on a training data-set combining two corpora: a high-quality proprietary multi-speaker corpus, and a subset of the VCTK [17] corpus. Both data-sets contained natural speech waveforms from English speakers speaking in non emotional styles. Our internal corpus contains 181 hours of speech from 13 speakers (each contributing between 5 and 35 hours), which we combine with 21 English VCTK speakers (each with 23 minutes of recordings).

Alongside this training corpus, we also have access to a small corpus of expressive recordings for one of our internal speakers. This speaker exhibits the ‘excited’ emotion across three intensity levels: low, medium and high. We randomly picked one utterance for each intensity level to condition the reference encoder in our 1-shot experiments. Furthermore, we selected 50 text prompts to be synthesized for perceptual evaluations, ensuring that each prompt corresponds to an appropriate emotional and intensity level given its content.

4. RESULTS

We evaluated each system using both objective and perceptual metrics. Our objective metrics are the KL-divergence of our reference encoder and the final teacher-forced L2 loss achieved by our the decoder. To choose systems for perceptual evaluations, we compare our system against other *state-of-the-art* methods with HouseHolder Flows (across multiple hyper-parameter configurations), and choose the best model with and best model without flows. Our perceptual evaluations target speech naturalness, emotional intensity and signal quality.

4.1. Objective Evaluation

We compare the Vanilla VAE system summarized in the Section 2, with the three other architectures predicting house-

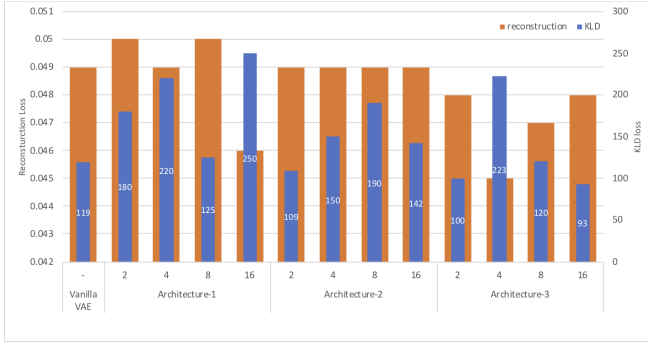


Fig. 3: KL-Divergence (blue bars) and Reconstruction Objective (orange bars) Metrics. We plot the average loss for the final epoch when training is stopped for each architecture and hyper-parameter presented in the Proposed Approach section.

holder vectors in different ways. Our proposed architecture, Architecture-3, achieves the lowest combination of KL-divergence and Reconstruction Loss across all systems (Figure 3). Furthermore, it is the only architecture which provides lower KL-divergence, reducing it by 22%, or improved disentanglement, while also providing better reconstruction when comparing against Vanilla VAE. From informal listening, this model indeed provided better quality in limited scenarios as compared to the other models, with flows, that we tried. We believe this is because the flow in our proposed architecture is now able to learn transformations of factors that are relevant for speech in general rather than having to learn how to transform factors that could change with different inputs. We will refer to this architecture with the best hyper-parameter (16 householder vectors) as VAE+FLOW for the rest of this paper.

In the next section, we verify that the superior disentanglement achieved by our model (without a loss in reconstruction loss), allows for convincing 1-shot generalization of expressive speech.

4.2. Perceptual Evaluations

To evaluate our model with respect to naturalness, emotional strength, and signal quality, we ran three Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) tests [18] across 40 listeners from Amazon Mechanical Turk.

For naturalness, we asked our listeners “Please rate the naturalness of the systems. Make sure that at least 1 system per screen is always rated 100.” and provided a scale from 0 (Very poor) to 100 (Completely natural). To measure emotional strength, we asked “Please rate the strength of the emotion of the systems. Consider the neutral sounding sample (there is one provided for each utterance) as 0 and an extremely emotional sounding sample as 100.” and provided a scale from 0 (Neutral) to 100 (Very Emotional). Finally, to measure signal quality, we asked “Please rate the systems in terms of audio quality. Try to ignore the content of the speech

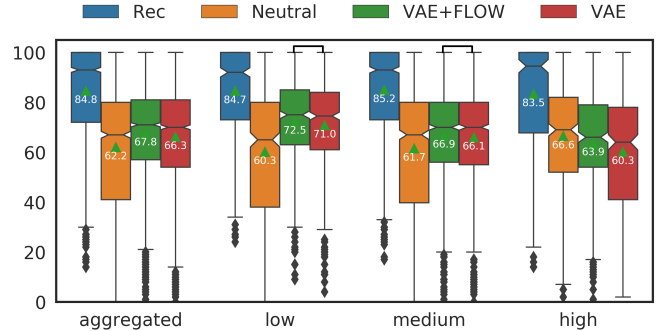


Fig. 4: Perceptual Results of naturalness for low, medium, high and aggregated emotional intensities. We plot MUSHRA responses of listeners as box-plots for each intensity. The number in white and location of the green triangle represents the mean of the listeners response for that system.

and the expressivity and instead focus on the quality of the audio signal (e.g glitches, clicks, noise...)” and provided a scale from 0 (Very poor) to 100 (Very good).

Each MUSHRA consisted of 4 systems: recordings of Emotional Speech (hidden higher anchor), Neutral TTS (hidden lower anchor), Vanilla VAE and VAE+FLOW. We show each listener 50 utterances with excited emotions with prompts as described in the Corpus section. To measure statistical significance, we ran t-tests with Bonferroni-Holm corrections at a 95% significance level. Our null-hypothesis is that the mean MUSHRA scores of the listening population for the two systems is equal.

4.2.1. Naturalness

MUSHRA evaluations of naturalness are presented in Figure 4. When aggregated across all emotional intensities, there is a statistically significant difference between all pairs of systems. Both 1-shot based expressive systems (VAE+FLOW and VAE) obtain higher naturalness (66.3 and 67.8 respectively) than the neutral system (62.2). When we break down the results per intensity conveyed in the reference utterance, this improvement in naturalness of both the 1-shot based expressive systems with respect to the neutral system is maintained for low and medium intensities. However, the difference between the two 1-shot expressive systems is not statistically significant. In contrast, when using high intensity utterances as the reference, there is a statistically significant degradation respect to neutral (66.6) for both expressive systems: partially reduced by VAE+FLOW (63.9) as compared to the vanilla VAE (60.3).

4.2.2. Emotional Strength

For emotional strength (Figure 5), we found significant differences between all pairs of systems, in the aggregated results as well as results split by intensity. As expected, listeners

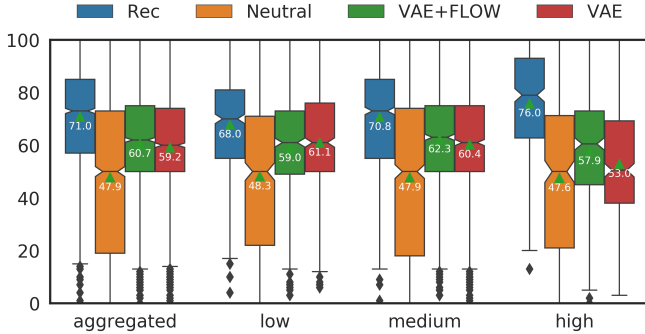


Fig. 5: Perceptual Results for emotional strength across low, medium, high and aggregated emotional intensities. Legend and metrics are reported as in Figure 4.

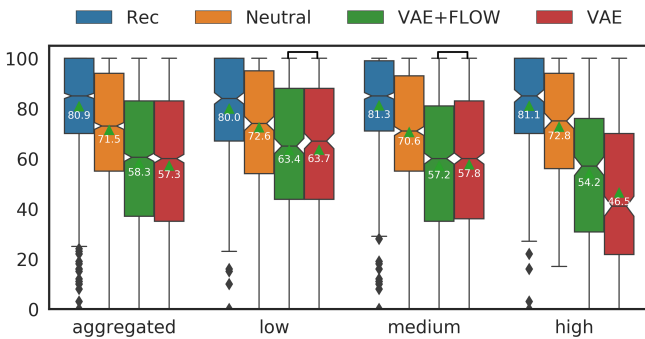


Fig. 6: Perceptual Results for signal quality for low, medium, high and aggregated emotional intensities. Legend and metrics are reported as in Figure 4.

scored the emotional strength of the neutral system consistently (and lower than rest of the systems) across all the evaluations and both the 1-shot expressive systems were scored with a higher mean score than neutral.

4.2.3. Signal Quality

Figure 6 shows the perceptual evaluations of signal quality. When aggregated across all emotional intensities, we see a significant difference between all pairs of system. Furthermore, for low and medium intensities the difference between flow and no-flow is not significant, whereas differences between all other system pairs are significant. In the high intensity case, the difference between all pairs of systems is significant.

5. DISCUSSION

Overall, both Vanilla VAE and VAE+FLOW provide a method to perform one-shot adaptation of a new expressive style. Vanilla VAE provided a statistically significant improvement in emotional strength (23% relative), and naturalness (2% relative) over Neutral TTS despite listeners detecting a degradation in signal quality (20% relative).

Our proposed VAE+FLOW architecture, leveraging a simple computationally inexpensive flow, is able to statistically significantly improve further upon Vanilla VAE in emotional strength (2.5% relative) naturalness (2.2% relative), and signal quality (1.7% relative).

After breaking down the perceptual results by emotional intensity, we found that the relative improvement achieved by VAE+FLOW over Vanilla VAE increases with emotional intensity. In this scenario, a higher intensity means a bigger acoustic divergence from the neutral speech. We, therefore, hypothesize that our proposed VAE+FLOW architecture performs better at the extreme cases of high intensity due to improved disentanglement, leading to better generalization capabilities.

This hypothesis is supported by our perceptual tests on emotional strength. All systems suffer a drop in naturalness when synthesizing high excitement (due to the degradation in signal quality caused by trying to synthesize speech very different to the training corpus). However, the improvement in generalization by the VAE+FLOW system provides a smaller naturalness degradation, compared to Neutral TTS, than the Vanilla VAE system in this case (4% and 9.5% respectively). Finally, we note that while we’re consistently increasing emotional strength in the low and medium case, these results are not translating proportionately to listeners perception of naturalness despite natural recordings of those emotions being rated highly. We hypothesize this is due to the limited signal quality that 1-shot systems can currently achieve.

6. CONCLUSIONS

We proposed a novel approach using householder flows within a text-conditioned VAE that overcomes the common trade-off between disentanglement and reconstruction in a state-of-the-art VAE implementation. The proposed flow-based VAE system significantly reduced KL-divergence (22% relative reduction) whilst providing a 2% relative improvement in reconstruction. Furthermore, we showed that pre-existing methods can perform 1-shot adaption to a new expressive style with only one utterance. Our VAE+FLOW model improved naturalness and emotional strength over an existing state-of-the-art system particularly when generalizing across more extreme prosody differences (with respect to the neutral training data) using only one utterance.

We have shown that signal quality is the bottleneck preventing joint improvement of naturalness and emotional strength. Future investigations are necessary to create voices with limited resources across a wider variety of expressive styles, seeking to further improve naturalness without harming expressiveness.

7. ACKNOWLEDGMENT

We would like to thank Pablo Garcia Moreno, Jasha Droppo, Henry Moss, Andrew Breen and Thomas Drugman for insightful discussions relevant to this work.

8. REFERENCES

- [1] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.
- [2] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.
- [3] Wei-Ning Hsu, Yu Zhang, Ron J Weiss, Heiga Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathan Shen, et al., “Hierarchical generative modeling for controllable speech synthesis,” *arXiv preprint arXiv:1810.07217*, 2018.
- [4] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J Weiss, Rob Clark, and Rif A Saurous, “Towards end-to-end prosody transfer for expressive speech synthesis with tacotron,” *arXiv preprint arXiv:1803.09047*, 2018.
- [5] Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo, “Expressive speech synthesis via modeling expressions with variational autoencoder,” *arXiv preprint arXiv:1804.02135*, 2018.
- [6] Eliya Nachmani, Adam Polyak, Yaniv Taigman, and Lior Wolf, “Fitting new speakers based on a short untranscribed sample,” *arXiv preprint arXiv:1802.06984*, 2018.
- [7] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui 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.
- [8] Jaime Lorenzo-Trueba, Gustav Eje Henter, Shinji Takaki, Junichi Yamagishi, Yosuke Morino, and Yuta Ochiai, “Investigating different representations for modeling and controlling multiple emotions in dnn-based speech synthesis,” *Speech Communication*, vol. 99, pp. 135–143, 2018.
- [9] Xixin Wu, Yuewen Cao, Mu Wang, Songxiang Liu, Shiyin Kang, Zhiyong Wu, Xunying Liu, Dan Su, Dong Yu, and Helen Meng, “Rapid style adaptation using residual error embedding for expressive speech synthesis,” in *Interspeech*, 2018, pp. 3072–3076.
- [10] Jian Gao, Deep Chakraborty, Hamidou Tembine, and Olaitan Olaleye, “Nonparallel emotional speech conversion,” *arXiv preprint arXiv:1811.01174*, 2018.
- [11] Ravi Shankar, Jacob Sager, and Archana Venkataraman, “A multi-speaker emotion morphing model using highway networks and maximum likelihood objective,” *Proc. Interspeech 2019*, pp. 2848–2852, 2019.
- [12] Nishant Prateek, Mateusz Lajszczak, Roberto Barra-Chicote, Thomas Drugman, Jaime Lorenzo-Trueba, Thomas Merritt, Srikanth Ronanki, and Trevor Wood, “In other news: A bi-style text-to-speech model for synthesizing newscaster voice with limited data,” *NAACL HLT 2019*, pp. 205–213, 2019.
- [13] Javier Latorre, Jakub Lachowicz, Jaime Lorenzo-Trueba, Thomas Merritt, Thomas Drugman, Srikanth Ronanki, and Viacheslav Klimkov, “Effect of data reduction on sequence-to-sequence neural tts,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7075–7079.
- [14] Diederik P Kingma and Max Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
- [15] Jakub M Tomczak and Max Welling, “Improving variational auto-encoders using householder flow,” *arXiv preprint arXiv:1611.09630*, 2016.
- [16] Jaime Lorenzo-Trueba, Thomas Drugman, Javier Latorre, Thomas Merritt, Bartosz Putrycz, Roberto Barra-Chicote, Alexis Moinet, and Vatsal Aggarwal, “Towards achieving robust universal neural vocoding,” 2019.
- [17] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonal, et al., “Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit,” *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.
- [18] ITUR Recommendation, “Bs. 1534-1. method for the subjective assessment of intermediate sound quality (mushra),” *International Telecommunications Union, Geneva*, 2001.