Learning fast with fewer data samples using Neural HMMs

Shivam Mehta, Harm Lameris, Éva Székely, Jonas Beskow, Gustav Eje Henter
Department of Speech, Music & Hearing, KTH Royal Institute of Technology, Sweden
{smehta, lameris, szekely, beskow, ghe}@kth.se

Abstract

The neural TTS paradigm synthesises significantly better-quality speech than the previous paradigm of HMM-based statistical parametric speech synthesis (SPSS). However, it requires a large amount of time and a larger corpus to learn the alignments between text and speech because of the underlying non-monotonic attention mechanism. This paper presents the benefits of merging a neural TTS system with a Hidden Markov Model (HMM) thus mixing these two paradigms and getting the best of both worlds. We replace the underlying attention mechanism in a neural TTS with an autoregressive left-to-right no-skip HMM defined by a neural network. This results in a system which learns to speak 10 times faster, requires fewer training samples, does not break down into gibberish, is smaller in size, is fully probabilistic, and allows easy control over the speaking rate without compromising the naturalness of the audio.

Introduction

Text-to-Speech (TTS) also referred to as speech synthesis, aims to synthesise human-like natural and intelligible speech from text. Over the past decade, there has been a paradigm shift from the statistical parametric speech synthesis (SPSS) (Zen et al., 2009) to Neural TTS (Tan et al., 2021). This transition mainly occurred because of the superior synthesis quality of Deep Neural Network (DNN)-based TTS systems, along with the possibility of synthesising without the need for extensive feature engineering. Generally, Neural TTS is split into two elements: an acoustic model and a neural vocoder. The acoustic model is responsible for generating an intermediate audio representation from the text, while the vocoder is responsible for transforming those intermediate audio representations into a waveform. Some state-of-the-art neural TTS systems like Tacotron 2 (Shen et al., 2018), Glow TTS (Kim et al., 2020), etc. use mel-spectrograms as an intermediate audio representation and combine it with neural vocoders like WaveNet (Oord et al., 2016), HiFi-GAN (Kong et al., 2020), etc. to generate high-quality speech.

Initially integrating deep neural networks into HMM-based speech TTS increased the naturalness, but an externally forced alignment was obligatory for them to synthesise good quality speech (Watts et al., 2016). This issue of forced alignment was resolved with the use of an attention mechanism in the neural sequence-to-sequence TTS systems (Shen et al., 2018; Wang et al., 2017), where attention was applied between the context vector generated from the input text and the corresponding mel-spectrogram frame. Attention takes a long time to form, however, and it requires a substantial amount of training data. Additionally, it is non-monotonic in nature and does not enforce a sequential ordering of speech sounds, which causes synthesis that is susceptible to babbling, stuttering, or even unintelligible gibberish. While most components of a Neural TTS system, the front-end encoder, the intermediate audio representation (mel-spectrograms), and acoustic feedback of autoregression grant a major performance boost in the synthesis quality, attention impedes the ability to learn to generate good quality speech quickly (Watts et al., 2019).

In this paper, we replaced the attention mechanism in Tacotron 2 (Shen et al., 2018) with a left-to-right no-skip hidden Markov model. The resulting system obtained is thus smaller in parameter size, requires less training data, maximises the exact likelihood of the data, learns to speak and align much quicker, does not babble and allows for easy control over the speaking rate with a quantile-based transition. For synthesised audio examples and code please visit our demo page: https://shivammehta007.github.io/Neural-HMM/

Material and methods

Neural HMMs use the best elements of the Neural TTS based system and hidden Markov model-based Statistical Parametric Speech Synthesis (SPSS). The design choices were made based on the analysis of the pros and cons of both paradigms as described in the paper (Watts et al., 2019), which analysed both systems and realised the importance of different components in both paradigms. We mixed those components to get the best of both worlds. The architecture of the mixed TTS system consists of:

1. A learned front-end encoder instead of a rule-based system of SPSS. This aids in the coarticulation of words improves the prosody and allows the system to handle out-of-vocabulary words without writing explicit rules for them.
2. Autoregression helps the model provide feedback from the previously generated frames resulting in higher acoustic quality and improved frame-level positional encoding, whereas HMMs had the problem of constant statistics per state resulting in inconsistent durations and worse acoustic quality.
3. Mel-spectrogram as an acoustic feature provided the ability to generate better quality waveforms with neural vocoders while in the case of HMMs the acoustic features used were the vocoder features with F0, which resulted in over averaging of the pitch.
4. Instead of attention, neural HMMs use a left-to-right no-skip hidden Markov model which enforces monotonic alignment and provides a more natural way of ending the synthesis. While the neural TTS system (Tacotron 2) has a stop token which might not generate a large enough value to stop the synthesis thus making the system babble.

A detailed architecture of neural HMM is presented in Figure 1. In the further section, we will shed more light on the mentioned design choices and their benefits.

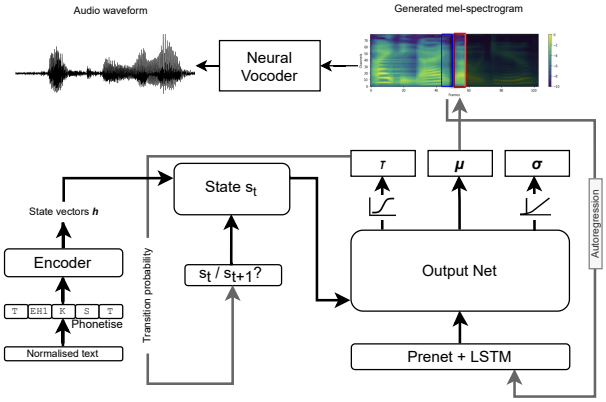

Figure 1. Neural HMM architecture

Phonemes

To synthesise speech in a monotonic nature we used phones as input to the Neural HMM instead of graphemes. This helped us to enforce a monotonic constraint as graphemes can have silent characters or different pronunciations altogether. Another factor for this design choice was the research (Fong et al., 2019) which listed the advantages of using phonemes instead of graphemes for a better synthesis quality.

Autoregression

Autoregression, along with having multiple states per phone, helped the model to define better sub-phone positioning. Looking at the previously generated frame, not only generates consistent harmonics but also solves the issue of constant statistics for each state in hidden Markov models. Figure 2 displays the effect of autoregression on mel-spectrograms. The top part is synthesised from a non-autoregressive model therefore, the statistic of an individual state remains constant and harmonics in higher frequency were not generated because of over smoothing across time while the bottom part is synthesised from an autoregressive neural HMMs thus there are different statistics even for a single state, therefore solving the problem of constant statistics per state of HMMs.

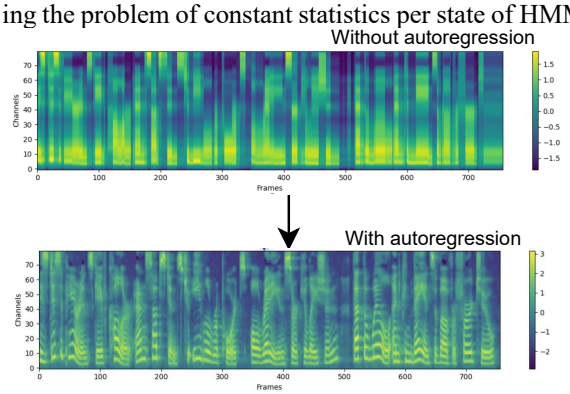

Figure 2. Effect of autoregression on the synthesised mel-spectrogram

Autoregression not only improves the synthesised mel-spectrograms but also improves the transition probability distribution of switching from one state to another. Without autoregression, it follows a geometric distribution when it arbitrarily switches to the next state resulting in inconsistent speech sound durations. On the contrary, with autoregression, the model is now aware of its position during a definite utterance of a phone and increases the probability of transitioning to the next state as it approaches the end of the phone's utterance. Figure 3 shows this phenomenon. We can see with autoregression, the transition probability for each state differs at each synthesis timestep, gradually increasing towards the end of the state's utterance while without autoregression, it generated a constant transition probability per state.

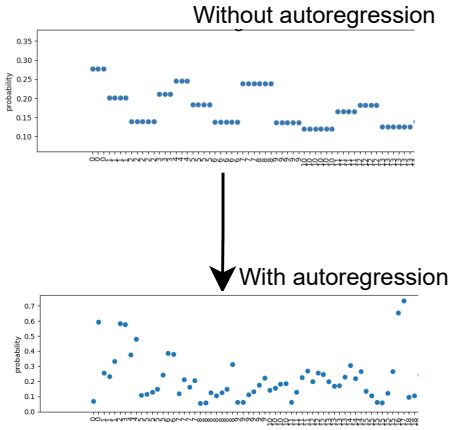

Figure 3. Effect of autoregression on transition probability of states

Left-to-right No-skip HMM

Since we used phones as the input to the Neural HMM TTS, we could reap the benefit from their acoustic property and enforce monotonic constraints on the alignments between text and speech. This enables the alignments to be learned 10x faster compared to cumulative attention-based Tacotron 2. This speedup helps the model to learn to speak faster and generate intelligible speech in as little as two hours of training. Since we use the forward algorithm of HMM it also computes the exact likelihood of the data and optimises that with gradient-based optimisation. During our experimentation, we found that using multiple states per phone improved the synthesis quality of utterance. This is useful, especially in the case of plosives where it is challenging for only one state to define the acoustic sound of the phone.

Experimentation Setup

We used the same configuration of two models as described in the paper (Mehta et al., 2022) this resulted in a 15.3M parameter neural HMM with 2 states per phone (NH2) configuration and a comparable configuration of Tacotron 2 without the post-net (T2-P) having 23.8M parameters. Both systems were trained for 30000 iterations with the training set of the LJSpeech dataset (Ito & Johnson, 2017). We extended the experiments with the use of a HiFi-GAN vocoder which was fine-tuned on mel-spectrograms of Tacotron 2 trained with LJSpeech as the neural vocoder for subjective evaluation, but any other compatible neural vocoder could also be used. The demo webpage has samples from both vocoder WaveFlow as in the original paper and HiFi-GAN.

Evaluation of Neural HMM TTS

We performed an objective evaluation and a subjective evaluation for the sentences synthesised by a neural HMM TTS system and a comparable Tacotron 2 configuration.

Objective evaluation

We synthesised test utterances of the LJSpeech dataset and transcribed them through Google's commercial ASR

system. We used those transcribed texts to calculate the Word Error Rate (WER) every 500 iterations.

Subjective evaluation

We performed a MUSHRA like MOS evaluation to validate the naturalness of the synthesised audio, the participants were asked to listen to 3 sets of phonetically balanced Harvard sentences, where each set containing 10 utterances. The participants were asked to rate the naturalness of the synthesised speech from 0 to 5.

Results

When trained on a full dataset of LJSpeech, the model started synthesising intelligible speech starting from 1,500 iterations while attention based Tacotron 2 uttered gibberish until 15,000 iterations. But when the amount of training set was reduced to a mere 500 training utterances. Tacotron 2 never learned to speak while neural HMM TTS remained unaffected by the low amount of data, deeming neural HMMs is very effective in a low resource setting. The result of the objective evaluation is plotted in figure 4 where we compare the WER of natural speech v/s Tacotron 2 without the post-net (T2-P) along with its 500 training utterances configuration v/s Neural HMMs with 2 states per phone (NH2) along with its 500 training utterances configuration.

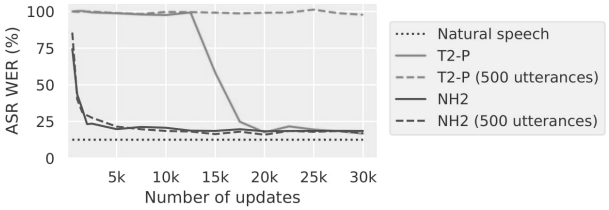

Figure 4. WER of systems at different updates during training

Tacotron 2 without the post-net (T2-P) gave 3.74 ± 0.06 MOS and neural HMMs with 2 states per phone (NH2) gave 3.5 ± 0.07 MOS while reducing the number of parameters by 43.48 % in the NH2 configuration. We suspect this discrepancy between MOS is mainly because of two factors: first, NH2 has relatively less modelling power compared to T2-P because of the smaller number of parameters in the former, and second, the vocoder HiFi-GAN was finetuned on the outputs of a Tacotron 2 system, therefore, the vocoder has more affinity to reduce noise and produce more natural-sounding waveforms while synthesising from a Tacotron 2 generated mel-spectrogram. It is also worth mentioning that no statistical difference was found between Tacotron 2 with post-net (T2+P) and Tacotron 2 without post net (T2-P) with HiFi-GAN, hinting that because of finetuning the HiFi-GAN is more robust to the artefacts present in mel-spectrograms synthesised by the autoregressive part of a Tacotron 2 system.

Conclusions

In this paper, we present the use of Neural HMMs for text-to-speech synthesis, where attention in the Neural TTS system is replaced by a left-to-right no-skip hidden Markov model. We experimented by replacing cumulative attention in Tacotron 2 with neural HMMs to synthesise mel-spectrograms and use HiFi-GAN as a neural vocoder to further generate waveforms. The resulting system learns to speak 10 times faster (within 2 hours with LJSpeech), is 43.48% smaller in size, allows control over the speaking rate and does not break into gibberish with comparable naturalness.

We believe such a system could speed up speech research and development as it not only provides faster iterations but also can work very well in a low-resources setup.

Acknowledgements

This research was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

Fong, J., Taylor, J., Richmond, K., & King, S. (2019). A comparison between letters and phones as input to sequence-to-sequence models for speech synthesis. In *Proceecdings of the 10th ISCA Speech Synthesis Workshop* (pp. 223–227).

Ito, K., & Johnson, L. (2017). *The LJ Speech Dataset*. https://keithito.com/LJ-Speech-Dataset/

Kim, J., Kim, S., Kong, J., & Yoon, S. (2020). Glow-TTS: A generative flow for text-to-speech via monotonic alignment search. In *Advances in Neural Information Processing Systems 33* (pp. 8067–8077).

Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (pp. 17022–17033).

Mehta, S., Szekely, E., Beskow, J., & Henter, G. E. (2022). Neural HMMS are all you need (for high-quality attention-free TTS). In *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 7457–7461).

Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499.*

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y., & Wu, Y. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4779–4783).

Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021). A survey on neural speech synthesis. *arXiv preprint arXiv: 2106.15561.*

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. In *Proceedings of the 2017 Annual Conference of the International Speech Communication Association* (pp. 4006–4010).

Watts, O., Henter, G. E., Fong, J., & Valentini-Botinhao, C. (2019). Where do the improvements come from in sequence-to-sequence neural TTS? In *Proceedings of the 10th ISCA Speech Synthesis Workshop* (pp. 217–222).

Watts, O., Henter, G. E., Merritt, T., Wu, Z., & King, S. (2016). From HMMS to DNNS: Where do the improvements come from? In *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5505–5509).

Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, *51*(11), 1039–1064.