Spontaneous Neural HMM TTS with Prosodic Feature Modification

Harm Lameris, Shivam Mehta, Gustav Eje Henter, Ambika Kirkland, Birger Moëll, Jim O'Regan, Joakim Gustafson, Éva Székely
Department of Speech, Music & Hearing, KTH Royal Institute of Technology, Sweden
{lameris, smehta, ghe, kirkland, bmoell, joregan, jkgu, szekely}@kth.se

Abstract

Spontaneous speech synthesis is a complex enterprise, as the data has large variation, as well as speech disfluencies normally omitted from read speech. These disfluencies perturb the attention mechanism present in most Text to Speech (TTS) systems. Explicit modelling of prosodic features has enabled intuitive prosody modification of synthesized speech. Most prosody-controlled TTS, however, has been trained on read-speech data that is not representative of spontaneous conversational prosody. The diversity in prosody in spontaneous speech data allows for more wide-ranging data-driven modelling of prosodic features. Additionally, prosody-controlled TTS requires extensive training data and GPU time which limits accessibility. We use neural HMM TTS as it reduces the parameter size and can achieve fast convergence with stable alignments for spontaneous speech data. We modify neural HMM TTS to enable prosodic control of the speech rate and fundamental frequency. We perform subjective evaluation of the generated speech of English and Swedish TTS models and objective evaluation for English TTS. Subjective evaluation showed a significant improvement in naturalness for Swedish for the mean prosody compared to a baseline with no prosody modification, and the objective evaluation showed greater variety in the mean of the per-utterance prosodic features.

Introduction

The advent of end-to-end neural text-to-speech (TTS) systems, such as Tacotron 2 (Shen et al., 2018), and FastSpeech 2 (Ren et al., 2020), has improved the quality of TTS compared to hidden Markov model-based TTS systems, and the quality now rivals human speech (An et al., 2021). Especially in conversational systems, however, these systems face issues relating to the ecological validity of the training data (Wester et al., 2016), which often consists of read-speech data or read conversational data (Kim et al., 2020; Shen et al., 2018). These TTS architectures treat naturally occurring prosodic variation as noise that is averaged for the output prosody (Raitio et al., 2020; Ram Mohan et al., 2021), and which do not allow for any control over the produced prosody (Raitio et al., 2022).

Concurrently, spontaneous speech has been increasingly used in speech synthesis (Gustafson et al., 2021). Although spontaneous speech is the most ecologically valid data to model conversational prosody, disfluencies, such as fillers, hesitations, and large variability make spontaneous speech challenging to model (Székely et al., 2019). Another challenge is the lack of structure, which includes unconventional sentence structure as well as overlap between the different speakers (Székely et al., 2019).

In this paper we investigate the use of spontaneous data for speech synthesis with prosodic feature control using neural HMM TTS (Mehta et al., 2022). Neural HMM TTS utilizes the monotonic statewise nature of left-right no-skip hidden Markov models (HMMs), which has two benefits: first, it helps in learning to synthesize coherent speech from small amounts of (disorderly) data, and secondly, it enables us to synthesize disfluencies, as the HMM can learn a specific state for the annotated disfluencies. We extend neural HMM TTS (Mehta et al., 2022) to learn a latent space to explicitly model the speech rate (sr) and the fundamental frequency (f0).

We train an English and a Swedish base speech synthesis model on read-speech datasets which are then fine-tuned on spontaneous speech datasets. We perform a subjective evaluation and an objective evaluation on the speech synthesizers. For the subjective evaluation, we asked participants to rate synthesized utterances for naturalness. For the objective evaluation, we compare the distribution of the per-utterance mean and per-utterance standard deviation of the prosodic features for the read-speech and spontaneous datasets, as well as the synthesized utterances.

Speech synthesizers

We use a modified version of neural HMM TTS (Mehta et al., 2022) for the experiments. Neural HMM TTS is an autoregressive TTS system that synthesizes mel-spectrograms conditioned on the input text. It has an encoder-decoder architecture based on Shen et al. (2018), but uses a no-skip left-to-right HMM in lieu of cumulative attention to enforce monotonic alignments. To enable the prosodic modification, we add a single feed-forward layer before the CNN-Bi-LSTM encoder, which projects the per-utterance speech rate and per-utterance mean fundamental frequency into a latent space. The output of the feed-forward layer is added to the phone embeddings to help learn the relation between text and prosodic features.

We trained separate speech synthesizers for the subjective and objective evaluations. For the subjective evaluation, we trained two speech synthesizers, one baseline voice and one voice with enabled prosody modification for each language for 15000 iterations. For the objective evaluation we use a further trained version at 37000 iterations of the English baseline and modifiable speech synthesizers. All synthesizers were trained on one NVIDIA GEFORCE RTX 3090 GPU.

Data

For the English read-speech model, the baseline in the objective evaluation, we used the LJSpeech corpus (Ito & Johnson, 2017). This corpus contains 24 hours or 11,300 utterances of a single female speaker of US English reading passages from seven non-fiction books published between 1884 and 1964. The utterances range in length from one to ten seconds. The spontaneous English model was trained on a corpus created from the audio of the Trinity Speech-Gesture dataset (Ferstl & McDonnell, 2018). The dataset consists of 25 impromptu monologues of approximately 10 minutes long by a male voice

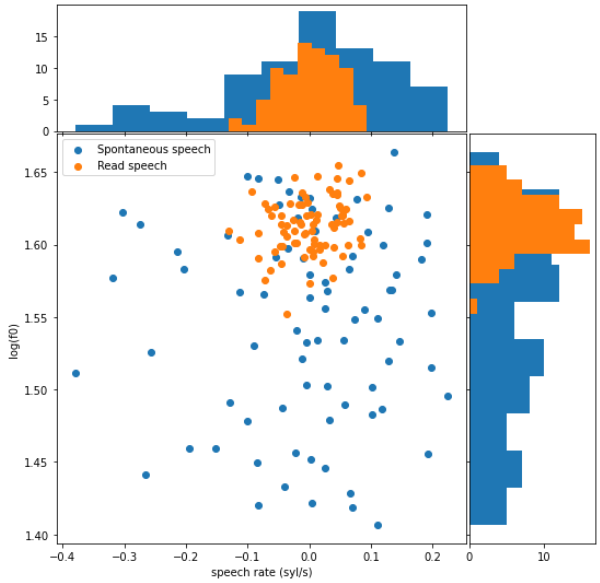

Figure 1. The per-utterance speech rate and per-utterance mean log(f0) centred around the dataset mean for the synthesized utterances.

actor speaking Hiberno-English. The monologues concern topics such as hobbies, daily activities, and interests.

The data for the non-spontaneous Swedish model is an open-source TTS corpus from the Norwegian Språkbanken ("NST Swedish Speech Synthesis (44 Khz)," 2021) containing read speech. This dataset was recorded by a male professional actor and consists of 11 hours of audio or 5200 utterances. The spontaneous Swedish corpus was compiled from six hours of audio extracted from a podcast in conversational style recorded by a male Swedish comedian. In the podcast, the speaker and his co-host prepare and evaluate sandwiches from a cookbook. Each podcast is 30-50 minutes long and contains discussions about food and exchanges of amusing stories from both interlocutors. Even though the most extreme speech behaviours were manually removed, such as singing, laughing, shouting, and speaking while eating, the resulting TTS corpus contains a lot of speaking style variation.

Experiments

Subjective evaluation

For the subjective evaluation, we performed a MUSHRA-like MOS test. The speech synthesizers were pre-trained on the read-speech corpus of the respective language for 10k iterations. The speech synthesizer without modifiable prosody was fine-tuned on the spontaneous speech corpus for 5k iterations. The training regimen for the prosodically modifiable speech synthesizer was identical apart from the addition of the prosodic features during fine-tuning. The selected utterances consisted of an equal mix of sentences omitted from training from the read-speech and spontaneous speech corpora. For the subjective evaluation, we compared an unmodified neural HMM TTS system baseline with the neural HMM TTS speech synthesizer with enabled prosody modification.

English

We recruited 20 participants using Prolific. The participants were required to be native speakers of English and reside in English-speaking countries. The participants were paid £3.13 for an average completion time of 14 minutes and 10 seconds.

We presented the participants with 20x6 matched stimuli. These stimuli consisted of one baseline unmodified stimulus and five stimuli where each feature, either speech rate or fundamental frequency, was the mean feature, or lowered or raised by 1 standard deviation from the mean. The MUSHRA-like MOS scores can be seen in Table 1. A one-way ANOVA shows no significant difference between the unmodified baseline and the prosody modification.

Swedish

For the Swedish subjective evaluation, we recruited 15 participants using Prolific. The people were required to be native speakers of Swedish residing in Sweden. Participants were paid £3.76 and completed the evaluation in an average of 15 minutes and 10 seconds.

The participants were presented with 20x6 matched stimuli which consisted of a baseline unmodified stimulus and five stimuli in which the speech rate or fundamental frequency were -1, 0, or 1 standard deviations from the mean. The MUSHRA-like MOS scores can be seen in Table 1.

Table 1. The MUSHRA-like MOS scores for the subjective evaluation.

Features	Standard deviation	English	Swedish
SR	-1	3.07 ±1.14	2.69±1.14
	1	3.16±1.19	2.66±1.07
f0	-1	2.97±1.18	2.46±1.07
	1	3.12±1.08	2.54±1.03
Mean feats.	0	3.13±1.08	**2.71±1.10**
No feats.	-	3.01±1.12	2.43±1.13

Objective evaluation

We performed an objective evaluation by comparing 80 generated utterances for the read-speech and spontaneous speech synthesizers for English, for which the spontaneous speech synthesizer has enabled prosody modification. We trained both voices for 37000 iterations. The prosody modifiable spontaneous voice was trained on the read-speech dataset for 20000 iterations without feature modification, before being fine-tuned on the spontaneous dataset for 17000 iterations. The synthesized utterances for the prosody modifiable spontaneous voice were synthesized with a balanced set of features ranging from -2 to 2 standard deviations from the mean. The results can be found in Figure 1.

Discussion

Subjective evaluation

For the subjective evaluation for English, a one-way ANOVA showed no significant difference between naturalness scores for the unmodified baseline compared to the prosodically modified stimuli. This suggests that the

prosody modification does not hinder training for the speech synthesizer, or result in deteriorated naturalness, even at 15k training steps.

For Swedish, a one-way ANOVA showed a difference in the means for the stimuli. A post-hoc Tukey indicated that the mean speech rate and mean fundamental frequency modification resulted in a significant improvement in the naturalness. This suggests that the model benefits from being conditioned on the mean speech rate and fundamental frequency.

The MOS scores are not straightforward to compare to other spontaneous speech synthesizers, as, e.g. Székely et al. (2019) were evaluated for appropriateness of speaking style and authenticity rather than naturalness. The Swedish MOS scores are lower than the English MOS scores, which we explain by the particularly challenging nature of the Swedish spontaneous corpus; the Swedish data was taken from spontaneous conversation and contains a large number of disfluencies, while the English corpus was extracted from a set of monologues.

Objective evaluation

Figure 1 shows greater variation for both the speech rate and the fundamental frequency. This indicates that at 37000 the latent space for the prosodic features has been demonstrably learned, and that it results in greater variation than synthesizing the read speech without features. ¨

Noteworthy is that while for the speech rate the greater variation for the spontaneous speech occurs both for slower and faster speech, the greater variation for the fundamental frequency mostly occurs for lower-pitched speech. Another identifiable difference is the apparent normal distribution of the speech rate and fundamental frequency for the unmodified synthesized utterances, compared to the more sporadic distribution for the spontaneous synthesized utterances.

Conclusion

In this study, we examined the use of spontaneous speech data for synthesizing prosodically modifiable speech. Our study demonstrated that the synthesized speech is more diverse than for synthesized read speech, and that the naturalness is either equal or better than unmodified utterances even after a very short training time. Additionally, it shows that modifiable prosody is achievable with relatively little GPU usage.

More research is needed, however, before synthesized spontaneous speech is of the same quality as synthesized read speech in order to tap the vast amount of spontaneous data available for speech synthesis and to achieve the prosodic variability of spontaneous speech.

References

An, X., Soong, F. K., Yang, S., & Xie, L. (2021). Effective and direct control of neural TTS prosody by removing interactions between different attributes. *Neural Networks*, *143*, 250–260. https://doi.org/10.1016/J.NEUNET.2021.06.006

Ferstl, Y., & McDonnell, R. (2018). Investigating the use of recurrent motion modelling for speech gesture generation. *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, 93–98.

Gustafson, J., Beskow, J., & É Székely. (2021). Personality in the mix-investigating the contribution of fillers and speaking style to the perception of spontaneous speech synthesis. *Proceedings of the 11th ISCA Speech Synthesis Workshop*. https://www.speech.kth.se/~jocke/publications/ssw11_personality.pdf

Ito, K., & Johnson, L. (2017). *The 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 H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 8067–8077). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/file/5c3b99e8f92532e5ad1556e53ceea00c-Paper.pdf

Mehta, S., Szekely, E., Beskow, J., & Henter, G. E. (2022). Neural HMMS Are All You Need (For High-Quality Attention-Free TTS). *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7457–7461. https://doi.org/10.1109/ICASSP43922.2022.9746686

NST Swedish speech synthesis (44 khz). (2021). In *NST Swedish Speech Synthesis (44 kHz)*. https://www.nb.no/sprakbanken/resource/nst-swedish-speech-synthesis-44-khz/

Raitio, T., Li, J., & Seshadri, S. (2022). Hierarchical Prosody Modeling and Control in Non-Autoregressive Parallel Neural TTS. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7587–7591. https://doi.org/10.1109/ICASSP43922.2022.9746253

Raitio, T., Rasipuram, R., & Castellani, D. (2020). Controllable neural text-to-speech synthesis using intuitive prosodic features. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2020-October*, 4432–4436. https://doi.org/10.48550/arxiv.2009.06775

Ram Mohan, D. S., Hu, V., Teh, T. H., Torresquintero, A., Wallis, C. G. R., Staib, M., Foglianti, L., Gao, J., & King, S. (2021). Ctrl-P: Temporal Control of Prosodic Variation for Speech Synthesis. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *5*, 3361–3365. https://doi.org/10.48550/arxiv.2106.08352

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2020). *FastSpeech 2: Fast and High-Quality End-to-End Text to Speech*. https://doi.org/10.48550/arxiv.2006.04558

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; Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/ICASSP.2018.8461368

Székely, É., Henter, G., Beskow, J., & Gustafson, J. (2019). Spontaneous Conversational Speech Synthesis from Found Data. *Interspeech*. https://www.isca-speech.org/archive_v0/Interspeech_2019/pdfs/2836.pdf

Wester, M., Watts, O., & Henter, G. E. (2016). *Evaluating comprehension of natural and synthetic conversational speech*. https://doi.org/10.21437/SpeechProsody.2016-157