On the Role of Dialogue Context in Predicting Speaking Style

Vincenzo Scotti and **Roberto Tedesco**
DEIB, Politecnico di Milano – Via Golgi 42, 20133, Milano (MI), Italy
vincenzo.scotti@polimi.it   roberto.tedesco@polimi.it

Text to Speech (TTS) synthesis is a problem almost as old as *Natural Language Processing* (NLP). The focus of this problem is on creating tools capable of generating a voice uttering a given text. These tools have many applications, from assistive technologies to chatbots and virtual assistants. Older solutions to build TTS tools relied on *concatenative synthesis* (either *diphone* or *unit selection*) (Jurafsky and Martin, 2009). Nowadays, *deep learning* powered solutions have shown impressive generative capabilities are becoming the standard technology to build TTS tools (Tan et al., 2021): models like *Tacotron* (Skerry-Ryan et al., 2018), *DeepVoice* (Ping et al., 2018), or *FastSpeech* (Ren et al., 2019) can generate incredibly natural voices. The naturalness of the generated speech is also helped by the use of neural *vocoders* that convert spectrograms into fluent and natural speech waveforms, replacing to the *Griffin-Limm algorithm* (Zhu et al., 2007).

Most recent deep learning models try to go beyond mere speech generation from the text. As a result, newer models try to factorise the probability predicted by these generative models to the condition it on various aspects: *speaker's voice*, *speaking style*, or *prosody* (Tan et al., 2021). While in some cases, as speaker conditioning, it is possible to build separate and re-usable models to extract the latent representation that encoders the desired information (Jia et al., 2018), in others, like speaking style prediction, the encoder of the latent representation is an integral part of the TTS model (Wang et al., 2018), yielding to a strong coupling between style encoder and TTS that makes the sub-modules hardly re-usable.

In this work, we focused on speaking style conditioning and detaching the style prediction from the actual speech synthesis, to promote the re-usability and modularity of these models. State-of-the-art TTS models use an unsupervised approach called *Global Style Tokens* (GSTs) that extracts the style representation from a given reference audio (Wang et al., 2018). This representation is computed internally by the TTS as a weighted sum of some learnt latent vectors. Changing the weights in the combination, the resulting style vector used to condition the spectrogram generation changes and as a result the speaking style changes (usually, these vector affects aspects like speech rate or intensity). These weights are computed from a reference audio clip (the one we want to emulate the speaking style); consequently, the resulting model is bound by the need for a reference database of speaking styles to choose from and a way to retrieve the reference style audio from the database, introducing the need of a human in the loop selecting manually the reference audio clip for the style. Alternatively, it would be necessary to have a separate module predicting the speaking style latent vector from the text the TTS needs to utter. This prediction problem is the object of this study.

Our work focuses on predicting the speaking style from the given text inside a conversation, for the application to conversational agents and chatbots. To this end, we developed and trained a neural network module working as a connection module between the textual component and the speech synthesis component of a chatbot. While some works have already addressed this problem more generically (Stanton et al., 2018), we are interested in understanding the role of the context of the conversation (i.e., the preceding turns in the dialogue) in the choice of the appropriate response speaking style.

We developed this connection module to help improve the human-likeness of chatbots and conversational agents. In fact, the choice of the appropriate tone and style while talking during a conversation is a consequence of the empathetic capabilities that characterise humans; thus, being able to control this aspect would improve the perception of the underlying agent. While this aspect does not seem to be crucial in assistive technologies applications, in other use cases like developing mental healthcare chatbots it is. The use case we propose for this technology is the development of counselling/psychotherapy chatbots, which require the simulation of empathetic traits to be perceived by the users as human thus favouring sympathy and openness from the user towards the agent.

To investigate the role of conversational context in the choice of the appropriate speaking style, we started from deep neural networks trained for dialogue language modelling –*DialoGPT* (Zhang et al., 2020) and *Therapy-DLDLM* (the latter is a custom dialogue model trained on open domain and therapy conversations)– and conditioned TTS model –namely *Mellotron TTS* (Valle et al., 2020), a variation of the Tacotron TTS–. The dialogue language models work as embedding models: we leverage their latent contextual representation of the dialogue as input to predict the speaking style. To explore the possible solutions, in the experiments we compared textual models having different complexities. Moreover, to understand the role of context, we compared, for each model, different encoding approaches: (i) response embeddings only, (ii) contextualised response embeddings, (iii) combined context-response embeddings. The latter pre-trained model provides a reference encoder for the GSTs. In fact, the employed TTS model encapsulates a reference encoder to extract the latent prosody representation used to compute the style latent vector.

Besides the two pre-trained models for text analysis and spectrogram generation, we leveraged also a pre-trained vocoder model –namely *WaveGlow* (Prenger et al., 2019)– to convert the Mel spectrogram synthesised by the TTS into a raw waveform. This allowed us to assess qualitatively that the audio synthesised by the TTS was still intelligible even if generated from the predicted GSTs, rather than the GSTs of a reference audio clip.

Following similar works on GSTs prediction (Stanton et al., 2018), we addressed the problems in two ways: (i) predicting the combination weights to compose the style vector, (ii) predicting directly the raw embedding vector The difference in the two targets results in slightly different approaches to the problem. In fact, the former approach requires a model working as a classifier, yielding a probability distribution that corresponds to the combination weights; we used the *Kullback–Leibler divergence* (KL divergence) of the predicted distribution from the target one as loss function. The latter approach, instead, requires a model doing regression, in this case, we trained the model minimising the Euclidean norm of the difference between the predicted and the target style vector. Note that to train the model, independently from the approach, we need to have a spoken dialogue data set providing the transcriptions. From the audio clips in the training data set, we can extract the target weights distribution or the target style vector we want to learn to predict on new data.

Table 1: Results of the two approaches to reconstruct the GST (lower is better).

Model	**No. of parameters**	**MSE**			**KL-Divergence**		
		Response	**Response from context**	**Context and response**	**Response**	**Response from context**	**Context and response**
DialoGPT	117M	0.0473	0.0515	0.0559	0.1238	0.1377	0.1526
	345M	0.0451	0.0449	0.0478	0.1148	0.1055	0.1249
	762M	0.0492	0.0557	0.0599	0.1236	0.1450	0.1540
Therapy-DLDLM	762M	0.0427	**0.0399**	0.0441	0.1047	**0.0958**	0.1109

We conducted the experiment training and evaluating the style prediction module on the *IEMOCAP corpus* (Busso et al., 2008), which offers scripted and spontaneous dialogues in English displaying different emotions that result in a wide variety of speaking styles[1]. Given this highly variated corpus, the neural network module could learn how to choose the appropriate speaking style given the output text and the conversational context, simulating the desired empathetic behaviour we were looking for. We evaluate the considered models by computing the prediction error of the speaking style (using both weight prediction and embedding prediction approaches) on the test split of IEMOCAP. The resulting figures, reported in Table 1, indicate an important role of context and language model complexity in achieving good predictive capabilities. In fact, the model with the highest number of parameters using contextualised response embeddings achieves the best results with both approaches.

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 42(4):335–359, 2008.

Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez-Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 4485–4495, 2018.

Dan Jurafsky and James H. Martin. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition*. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International, 2009.

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan Ömer Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 3617–3621. IEEE, 2019.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3165–3174, 2019.

R. J. 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. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4700–4709. PMLR, 2018.

Daisy Stanton, Yuxuan Wang, and R. J. Skerry-Ryan. Predicting expressive speaking style from text in end-to-end speech synthesis. In *2018 IEEE Spoken Language Technology Workshop, SLT 2018, Athens, Greece, December 18-21, 2018*, pages 595–602. IEEE, 2018.

Xu Tan, Tao Qin, Frank K. Soong, and Tie-Yan Liu. A survey on neural speech synthesis. *CoRR*, abs/2106.15561, 2021.

Rafael Valle, Jason Li, Ryan Prenger, and Bryan Catanzaro. Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 6189–6193. IEEE, 2020.

Yuxuan Wang, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A. Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5167–5176. PMLR, 2018.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DIALOGPT : Large-scale generative pre-training for conversational response generation. In Asli Celikyilmaz and Tsung-Hsien Wen, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 270–278. Association for Computational Linguistics, 2020.

Xinglei Zhu, Gerald Beauregard, and Lonce L. Wyse. Real-time signal estimation from modified short-time fourier transform magnitude spectra. *IEEE Trans. Speech Audio Process.*, 15(5):1645–1653, 2007.

[1]Links to the source code. Dialogue GST: `https://github.com/vincenzo-scotti/dialoguegst`, Mellotron TTS API: `https://github.com/vincenzo-scotti/tts_mellotron_api`, Thearapy-DLDLM: `https://github.com/vincenzo-scotti/dldlm`