# So-to-Speak: an exploratory platform for investigating the interplay between style and prosody in TTS

*Éva Székely, Siyang Wang, Joakim Gustafson*

Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden

szekely@kth.se, siyangw@kth.se, jkgu@kth.se

## Abstract

In recent years, numerous speech synthesis systems have been proposed that feature multi-dimensional controllability, generating a level of variability that surpasses traditional TTS systems by orders of magnitude. However, it remains challenging for developers to comprehend and demonstrate the potential of these advanced systems. We introduce So-to-Speak, a customisable interface tailored for showcasing the capabilities of different controllable TTS systems. The interface allows for the generation, synthesis, and playback of hundreds of samples simultaneously, displayed on an interactive grid, with variation both low level prosodic features and high level style controls. To offer insights into speech quality, automatic estimates of MOS scores are presented for each sample. So-to-Speak facilitates the audiovisual exploration of the interaction between various speech features, which can be useful in a range of applications in speech technology.

**Index Terms**: speech synthesis, TTS, prosody, speaking style

## 1. Introduction

As neural text-to-speech (TTS) systems have grown increasingly natural-sounding, research into style and prosody-controllable TTS has garnered significant attention [1, 2, 3, 4]. These advanced systems offer a vast range of expressive possibilities, making them candidates for various applications. However, understanding and evaluating their capabilities using traditional crowd-sourced perceptual evaluations presents a unique challenge due to the sheer variety of speech outputs they can produce. We here propose an interactive tool to help developers adequately assess the full potential of these systems, and to explore and evaluate the diverse landscape of style and prosody control in synthetic speech. Using this tool, hundreds of speech samples can be displayed on an interactive grid, arranged according to a flexible range of features. Acoustic-prosodic feature control is represented on the grid's axes, while higher-level controls, such as style, can be adjusted using a slider that navigates between grids. Each sample is colored corresponding to an automatic MOS score, to indicate an estimation of the synthesis quality.

So-to-speak aims to provide an interactive environment where users can generate, modify, and compare synthetic speech samples across a range of controllable prosodic features and speaking styles. By presenting these samples in a visually accessible interactive grid format, this platform enables users to explore the complex relationships between these factors and gain insights into how they shape speech perception. The codebase for the interface is made available[1].

In this demonstration we chose to include an example neural TTS model which is controllable in the prosodic features speech rate and f0, and it is possible to interpolate between read speech and spontaneous conversational style. But it is easy to adapt the interface to other combinations of prosody and style, including energy, spectral tilt [2] or voice quality features like creak [4] combined with genre-specific styles [5] such as podcast, impromptu monologue, formal presentation; situational context-dependent styles such as whisper or Lombard speech [3], or affective styles like smiling voice or emotional speech.

## 2. Synthesis Model Example

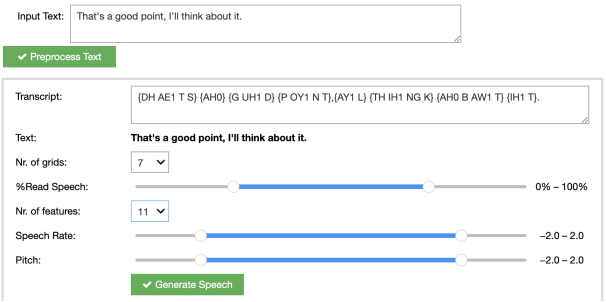
### 2.1. Data

The speech data was obtained from a publicly available multimodal corpus of 15 interactions between a human moderator and two participants that were given the task of decorating an apartment using a GUI on a large touch screen [6]. Speech data was extracted from the moderator, a male speaker of General American English, which was automatically segmented into multi-breath groups of maximum 11 seconds [7], totaling 5h 40 min. The spontaneous speech data were supplemented with 2 h 30 min of read-speech audio of the same voice talent reading sentences originating from both fiction and non-fiction texts. The prosodic features f0 and speech rate (syllables per second) were extracted and summarised for the spontaneous speech at breath group level and for the read speech at utterance level.

### 2.2. TTS Architecture

For the demo we use a TTS voice based on a modified Tacotron2 [8] architecture.Speaking style is identified in the model through an 8-dimensional speaker-like embedding, set up following [1]. This embedding is appended to each utterance's encoded text and passed to the attention and decoder blocks from the model. Using such an embedding allows for straightforward interpolation between styles. Control of the prosodic features *speech rate* and *pitch* is implemented similar to [2] by appending normalised breath group average feature values to the encoded text. Normalisation is performed by aligning the 1st and the 99th percentile points of the input data to the values of $-1$ and 1 respectively, while allowing outliers to go outside of that range, and also to prompt for features outside the normal range at inference, in which case the system attempts to model a voice style to fit the features. A model is first initialised on a pretrained read speech model, and trained for 70k iterations on the corpus with two embeddings, indicating whether the utterance is from the *read* or *spontaneous* part of the corpus. This model is then further trained including the prosodic features for an additional 100k iterations.

[1] https://github.com/evaszekely/So_To_Speak



Input Text: That's a good point, I'll think about it.

Transcript: {DH AE1 T S} {AH0} {G UH1 D} {P OY1 N T},{AY1 L} {TH IH1 NG K} {AH0 B AW1 T} {IH1 T}.

Text: **That's a good point, I'll think about it.**

Nr. of grids: 7

%Read Speech: 0% – 100%

Nr. of features: 11

Speech Rate: –2.0 – 2.0

Pitch: –2.0 – 2.0

Figure 1: *Input interface where the user can specify settings for the prosody and style control.*

# 3. Interface

## 3.1. Feature modification and display

The interface is built in a Jupyter Notebook using primarily the ipywidgets library for visualisation and control features. The whole process, from grapheme-to-phoneme conversion, through synthesis and MOS prediction is executed within the Notebook, and can be performed on a single GPU. The design is such that the prosodic feature variation is displayed on the axes of the interactive grid on which the samples can be played by selection, whereas the style function can be varied interactively, where a slider allows the user to scroll through the grids with varying levels of conversational and read speech styles.

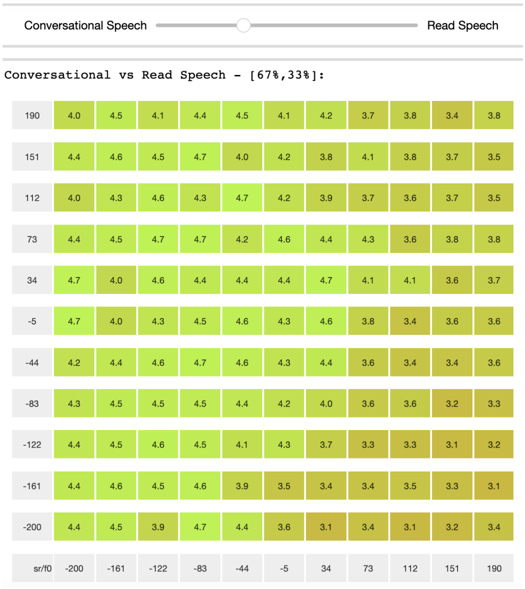
## 3.2. MOS prediction

To provide users with an automatic estimation of the synthesis quality, we use an off-the-shelf MOS prediction system [9] [2]. It is trained with synthesised read-speech samples and has shown strong generalisation capability in a large-scale comparison of MOS prediction systems [10].

Depending on the application area So-to-Speak is used in, the automatic MOS score display on the grid can be exchanged to an acoustic feature extractor or to an emotion recognition network to provide users with a visual overview of how the modified prosody and style affects for example voice quality, or an estimate of how they affect perceived emotion.

# 4. Use-case scenario examples

The firsthand use-case scenario of the So-to-Speak tool is TTS development itself. It can be helpful both to carry out rapid informal evaluations during the development stage as well as to aid in the selection of specific settings to generate samples for crowd-sourced listening tests. Human-computer interaction researchers could use So-to-Speak to create customisable voice profiles that might suit a conversational system or character in a given context. So-to-Speak can also be used to generate stimuli for perceptual experiments, and likely also to aid scientists in forming new hypotheses about perceptual effects of the interaction between prosodic features and speaking styles. Further potential application areas of modified versions of this platform are among others second language learning, Alternative and Augmentative Communication (AAC) and speech therapy.



```
Conversational vs Read Speech - [67%,33%]:
```

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 190 | 4.0 | 4.5 | 4.1 | 4.4 | 4.5 | 4.1 | 4.2 | 3.7 | 3.8 | 3.4 | 3.8 |
| 151 | 4.4 | 4.6 | 4.5 | 4.7 | 4.0 | 4.2 | 3.8 | 4.1 | 3.8 | 3.7 | 3.5 |
| 112 | 4.0 | 4.3 | 4.6 | 4.3 | 4.7 | 4.2 | 3.9 | 3.7 | 3.6 | 3.7 | 3.5 |
| 73 | 4.4 | 4.5 | 4.7 | 4.7 | 4.2 | 4.6 | 4.4 | 4.3 | 3.6 | 3.8 | 3.8 |
| 34 | 4.7 | 4.0 | 4.6 | 4.4 | 4.4 | 4.4 | 4.7 | 4.1 | 4.1 | 3.6 | 3.7 |
| -5 | 4.7 | 4.0 | 4.3 | 4.5 | 4.6 | 4.3 | 4.6 | 3.8 | 3.4 | 3.6 | 3.6 |
| -44 | 4.2 | 4.4 | 4.6 | 4.7 | 4.6 | 4.3 | 4.4 | 3.6 | 3.4 | 3.4 | 3.6 |
| -83 | 4.3 | 4.5 | 4.5 | 4.5 | 4.4 | 4.2 | 4.0 | 3.6 | 3.6 | 3.2 | 3.3 |
| -122 | 4.4 | 4.5 | 4.6 | 4.5 | 4.1 | 4.3 | 3.7 | 3.3 | 3.3 | 3.1 | 3.2 |
| -161 | 4.4 | 4.6 | 4.5 | 4.6 | 3.9 | 3.5 | 3.4 | 3.4 | 3.5 | 3.3 | 3.1 |
| -200 | 4.4 | 4.5 | 3.9 | 4.7 | 4.4 | 3.6 | 3.1 | 3.4 | 3.1 | 3.2 | 3.4 |
| sr/f0 | -200 | -161 | -122 | -83 | -44 | -5 | 34 | 73 | 112 | 151 | 190 |

Figure 2: *An example grid with a sentence synthesised with 7 style settings and 11 prosodic feature steps, totalling 847 unique speech samples. The audios play upon clicking on a cell. Moving the style slider on top updates the grid to the requested style.*

# 5. Acknowledgements

Supported by: Swedish Research Council proj. Perception of speaker stance (VR-2020-02396), Digital Futures proj. AAIS, Riksbankens Jubileumsfond proj. CAPTivating (P20-0298).

# 6. References

[1] R. Valle, J. Li, R. Prenger, and B. Catanzaro, “Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens,” in *Proc. ICASSP*, 2020.

[2] T. Raitio, R. Rasipuram, and D. Castellani, “Controllable neural text-to-speech synthesis using intuitive prosodic features,” in *Proc. Interspeech*, 2020, pp. 4432–4436.

[3] Q. Hu, T. Bleisch, P. Petkov, T. Raitio, E. Marchi, and V. Lakshminarasimhan, “Whispered and Lombard neural speech synthesis,” in *Proc. SLT*, 2021, pp. 454–461.

[4] H. Lameris, M. Włodarczak, J. Gustafson, and É. Székely, “Neural speech synthesis with controllable creaky voice style,” in *Proc. ICPhS*, 2023.

[5] É. Székely, J. Edlund, and J. Gustafson, “Augmented prompt selection for evaluation of spontaneous speech synthesis,” in *Proc. LREC*, 2020, pp. 6368–6374.

[6] D. Kontogiorgos, V. Avramova, S. Alexanderson, P. Jonell, C. Oertel, J. Beskow, G. Skantze, and J. Gustafson, “A multimodal corpus for mutual gaze and joint attention in multiparty situated interaction,” in *Proc. LREC*, 2018, pp. 119–127.

[7] É. Székely, G. E. Henter, and J. Gustafson, “Casting to corpus: Segmenting and selecting spontaneous dialogue for TTS with a CNN-LSTM speaker-dependent breath detector,” in *Proc. ICASSP*, 2019, pp. 6925–6929.

[8] J. Shen, R. Pang, R. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen *et al.*, “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in *Proc. ICASSP*, 2018.

[9] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, “Generalization ability of mos prediction networks,” in *Proc. ICASSP*, 2022.

[10] W.-C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, “The VoiceMOS Challenge 2022,” in *Proc. Interspeech*, 2022, pp. 4536–4540.

[2] `https://github.com/nii-yamagishilab/mos-finetune-ssl.git`