Speaking Rate Control of end-to-end TTS Models by Direct Manipulation of the Encoder's Output Embeddings

Martin Lenglet, Olivier Perrotin, Gérard Bailly

Interspeech 2022
18-22 September 2022, Incheon, Korea

Speaking Rate Control of end-to-end TTS Models by Direct Manipulation of the Encoder's Output Embeddings

Martin Lenglet, Olivier Perrotin, Gérard Bailly

Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, France
{martin.lenglet,olivier.perrotin,gerard.bailly}@grenoble-inp.fr

Abstract

Since neural Text-To-Speech models have achieved such high standards in terms of naturalness, the main focus of the field has gradually shifted to gaining more control over the expressiveness of the synthetic voices. One of these leverages is the control of the speaking rate that has become harder for a human operator to control since the introduction of neural attention networks to model speech dynamics. While numerous models have reintroduced an explicit duration control (ex: FastSpeech2), these models generally rely on additional tasks to complete during their training. In this paper, we show how an acoustic analysis of the internal embeddings delivered by the encoder of an unsupervised end-to-end TTS Tacotron2 model is enough to identify and control some acoustic parameters of interest. Specifically, we compare this speaking rate control with the duration control offered by a supervised FastSpeech2 model. Experimental results show that the control provided by embeddings reproduces a behaviour closer to natural speech data.
Index Terms: speech synthesis, embeddings analysis, natural control, duration control

1. Introduction

Deep neural Text-To-Speech systems such as Tacotron [1, 2] or FastSpeech [3], combined with neural vocoders like WaveNet [4], WaveRNN [5] or WaveGlow [6] produce more realistic voices than ever. As a result, numerous studies now focus on the rendition of the expressiveness [7, 8], whose control remains a ongoing challenge. In particular, prosody is known to convey co-verbal information that is desirable to make the interaction with a synthetic voice as natural as possible [9]. The accurate manipulation of prosodic parameters of interest such as pitch, energy or speaking rate is therefore a requirement for an interactive TTS system.

One approach to enable the control of these parameters at inference time consists in adding layers to the model in order to learn how to explicitly retrieve this information from the input sequence [3, 10]. Doing so, this information can be modified before being reintroduced into the decoding layer, resulting in a finer control of the output prosody. While this method enables an independent control of these parameters, it requires various preprocessing to extract alignments and acoustic parameters beforehand. Additionally, the proposed independent control may not correspond to the natural behaviour of the voice.

An alternative is to use implicit representation to bias the model toward the desired prosody [7, 11]. Via a so-called prosodic or reference encoder of the target speech signals, style and speaker embeddings model residual loss not yet explained by text input. During inference, a target prosodic example can then be used to complement the input text. While control of style may capture subtle natural co-variations, the semantics of control parameters is often given a posteriori.

In this paper, we introduce a new control for end-to-end TTS models: *Embedding Bias*. By analysing the phonetic embeddings at the encoder's output, we identify acoustic and paralinguistic parameters that are encoded in these latent representations, as well as their co-variations with other phonetic dimensions, learnt from the training data. We show how this information can be used to bias phonetic embeddings in order to control the speaking rate of the model, without the need for any additional data during the training phase. We implement and investigate this duration control on the embedding spaces of both Tacotron2 and FastSpeech2 models, whose biased embeddings are then fed to their attention mechanism and duration predictor, respectively. We compare this control to the explicit duration control provided by FastSpeech2.

2. Related Work

The explicit prediction of low-level prosodic parameters such as F_0, duration and energy from the embedding space of encoder-decoder TTS models has led to excellent performance in disentangling these parameters [3, 10] at the expense of preserving the natural co-variations between them. Moreover, duration control usually applies a uniform gain to all phones, whereas variations of phone duration with speaking rate depends on its phonemic class and position in the sentence [12]. Whether the loss of both supra-segmental acoustic co-variations and non-linear duration variations at a segmental level degrades naturalness is still an open question and is investigated here. The opposite direction that consists in biasing the encoder output with an implicit representation of an audio sample learnt by a reference encoder (Global Style Tokens [7, 11, 13], Variational Auto Encoders [14, 15] or speaker encoders [16]) supposedly better preserves the co-variations of prosodic parameters. However, if most implementations allow to successfully identify dimensions in the obtained latent space to control low-level prosodic parameters, few quantitative studies had statistically analysed variations and co-variations of prosodic parameters introduced by an implicit control both at segmental and supra-segmental levels. Also, methods for systematic analyses of latent spaces are rarely given, with exceptions such as [17] who performed an *a posteriori* analysis using a crowd-sourced subjective evaluation of synthesis.

The difference between concatenation or addition of the style and text encoders outputs is not well described in the literature, yet the addition intuitively corresponds with a translation in the embedding space. Therefore, can we derive the appropriate translation for a given prosodic parameter modification from an analysis of the embedding space, without the need to train a reference encoder? Previous work on embedding space analysis showed promising results in terms of phonetic [18] and acoustic [19] structuring of the embedding space, but no control were yet identified from these analyses.

10.21437/Interspeech.2022-759

3. Proposed Method

We aim at performing an acoustic analysis of the latent space outputted by the encoder of an end-to-end TTS model, and use this analysis to exhibit an embedding bias that can monitor the speaking rate of the model. This method could be applied to any encoder-decoder architecture which uses an attention mechanism or a duration predictor. Both cases are implemented, taking Tacotron2 [2] and FastSpeech2 [3] as examples.

3.1. Encoder-decoder TTS models

Our implementation of Tacotron2 (*TC*) builds on the one shared by NVIDIA [20]. Following [21], *TC* uses a Gate Loss correction and is trained on both orthographic and phonetic transcripts, which are known to benefit to both types of inputs [22]. Additionally, the decoder generates two mel-spectrogram frames per step. Empirical analysis showed that generating 2 frames at a time did not degrade the overall quality of the synthetic speech, while speeding the inference process. FastSpeech2 (*FS*) strictly follows an early implementation [23]: the pitch predictor is trained on F_0 values instead of continuous wavelet transform in later versions. A Tacotron2-type post-net is added after the decoder. Also, pitch and energy values are averaged per phone instead of per frames, and normalised.

Both *TC* and *FS* are trained on a subset of the new segmentation of the French M-AILABS dataset provided by [24]. This subset includes 29557 utterances (more than 25h) of audiobook recordings from four novels uttered by Nadine Eckert-Boulet (NEB). 5% of this corpus (1477 utterances) was randomly picked as the test set. This dataset provides both orthographic and phonetic transcripts for every utterance. Only the phonetic transcripts (together with spaces and punctuation when associated with pauses) were used for *FS*, which was also provided a hand-checked phonetic alignment to train its duration predictor. Both models were trained until convergence, which took about 100 epochs. The post-net is bypassed during the first 10 epochs, while the learning rate is fixed at 10^{-3}. After this startup, the learning rate decreases exponentially until reaching 10^{-5} after 90 epochs. The batch size is set to 32 for both models. The vocoder used is WaveGlow [6].

3.2. Identification of Acoustic Parameters in Embeddings

After training, the entire test set was synthesised with both models, using the phonetic input. Together with the usual audio output, embeddings computed by the encoder of both models are saved for acoustic analysis, as well as the attention map from *TC* and the duration predictions from *FS*.

3.2.1. Automatic Segmentation of the Synthesised Audio Signal

In *TC*, durations of input phones are computed using the durations of their respective activations in the attention map [25]. The duration of output phones predicted with this method were compared to Ground-Truth phones duration. Syntheses were produced using teacher-forcing to ensure the same dynamic as the Ground-Truth. We measured a correlation of 0.88 on phones (durations of silences from punctuation marks are excluded), which made us consider this method for large scale acoustic analysis. Segmentation in *FS* is straightforward, the duration predictor providing the number of frames for each phone. An acoustic analysis of each phone is performed with Praat [26]. Several acoustic parameters are considered: phone duration, fundamental frequency (F_0), first three formants (F_1, F_2, and F_3), and energy (Sound Pressure Level).

Table 1: *Correlation coefficients between acoustic features predicted from MDS coordinates and measured on synthesis. log(d) = logarithm of the duration ; E = Energy.*

Model	**log(d)**	**F_0**	**F_1**	**F_2**	**F_3**	**E**
Tacotron2	0.83	0.51	0.70	0.93	0.75	0.67
FastSpeech2	0.89	0.86	0.84	0.91	0.74	0.87

3.2.2. Acoustic Analysis of the Embedding Space

The synthesis of the entire test set provides a total of 51746 phone embeddings that encode contextual information introduced by the encoder of each model. To consider the voiced-dependent acoustic features (section 3.2.1), only the 22528 vowels of the test set are studied in this section. The relationship between embeddings and acoustic features measured on the corresponding synthesised audio segments is derived as follows: 1) Dimensional reduction of the embedding space with Multi-dimensional Scaling (MDS) [27]. A distance matrix between embeddings is first calculated using cosine distance. 2) A projection matrix is derived to enable transitions between the initial embedding space and the reduced MDS space. 3) All the acoustic features are individually approximated by least square multilinear regression from embeddings coordinates in the MDS. This procedure is similar to [19], but is applied on phone embeddings instead of utterance-wise style embeddings.

The approximation of acoustic parameters from the MDS coordinates is compared to the measured acoustic features on the synthesised signals and correlation coefficients are shown in table 1. Phone durations are computed in logarithmic scale, because this gave better correlations. Same goes for F_0, F_1, F_2 and F_3 which are expressed in semitones for better approximations. These correlations indicate that most of these acoustic features are well encoded in the embeddings. Note that a lower correlation does not mean that the model does not implement this acoustic feature, but rather that this feature is not encoded in the phone embeddings alone (note that duration, F0 and energy encoders of *FS* further contextualise embeddings with CNNs) or not correlated in a linear way. As a result, this feature is less likely to be easily controllable by modifying the embeddings before passing through the decoder. On the contrary, high correlations emphasise the features that are encoded in this latent space: phone duration is well encoded by every model, as well as spectral clues such as formants. *FS* has better correlations of prosodic measurements like F_0 and energy, which are trained to be predicted by the model from the very same embeddings.

3.3. Acoustic control

From the regressions described in section 3.2.2, the gradient of each acoustic feature in the MDS is computed. This vector, called *embedding bias*, is the leverage used to control one particular feature at a time: a translation along this vector is added to all the embeddings of an utterance before passing through the decoder. The regression is used to evaluate the magnitude of translation needed to induce the desired modification of the acoustic feature. This study specifically evaluates the control given by the duration embedding bias, expressed in log-duration. Hence the addition of a bias in the log domain is equivalent to applying a multiplying factor on phone duration. We empirically identified that a correcting factor k was needed to achieve the desired modification of phone duration, resulting in a corrected translation of $k * log(m)$ to multiply phone duration by m. $k = 2.94$ and $k = 2.33$ for *TC* and *FS* respectively.

In the case of *FS*, the embedding bias is applied before duration prediction, and predicted duration from the biased embeddings is used for decoding, without any external input. We showed in a preliminary study that an embedding bias computed on vowel embeddings alone is more efficient in inferring duration modification in the synthesis signal, supported by the fact that vowels duration show more variability than consonants [12]. In the following, the bias is derived from the vowel embeddings space but applied on all input phone embeddings at inference.

4. Experiments and Results

4.1. Models and test set

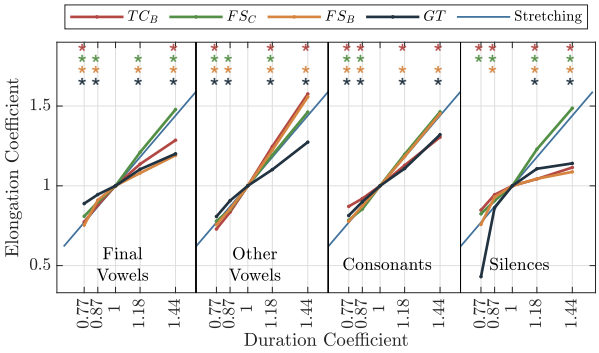
In this section we will investigate and compare the efficiency of the embedding bias control on *TC* and *FS*. In addition, two baselines are added: *FS* with explicit duration control (without embedding bias) and a simple linear time-interpolation of the mel-spectrogram output of an unbiased *TC* (resp. *FS*) to change the full duration of the signal before feeding it to the neural vocoder. In both baselines, a similar modification of duration is applied on all phones, but *FS* has the chance to make some acoustic modifications through the decoding process. In the following, TC_B, FS_B, FS_C and *stretching* refer to *TC* with embedding bias, *FS* with embedding bias, *FS* with explicit duration control, and mel-spectrogram interpolation, respectively.

The test set described in section 3.1 is synthesised with 4 duration coefficients, chosen to be representative of the phone rate distribution of the training dataset. These coefficients $m_i = \{0.77, 0.87, 1.18, 1.44\}$ are chosen to reach $i = \{+2, +1, -1, -2\}$ standard deviation around the mean phone rate, respectively.

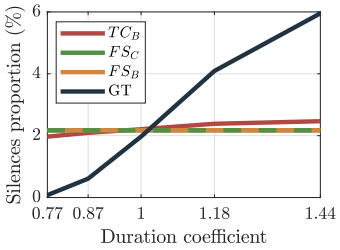
4.2. Non-linear duration modification

For each synthesised signal with a given duration coefficient, the duration of each phone is measured (see section 3.2.1), and divided by the mean duration of its phone class synthesised with the same model without duration control, to provide an elongation coefficient. Fig. 1a displays the average elongation coefficient per duration coefficient, model, and phone class. Final vowels are vowels just preceding a silence in the audio signal. For each phone class, the diagonal corresponds to the *stretching* condition, where the elongation coefficient equals the duration coefficient. The red, green and yellow curves correspond to TC_B, FS_C, FS_B, respectively. Moreover, average phone elongation coefficients were also calculated on the ground truth train database (*GT*) and reported in dark blue. A Kruskal-Wallis rank-sum test performed on the per-phone elongation coefficients showed a significant effect of both phone class and duration control ($p < 10^{-3}$). A post-hoc Wilcoxon rank-sum test then assessed for each phone class and duration coefficient whether each method significantly differs from the *stretching* conditions. Significance ($p < 10^{-3}$) is displayed by coloured stars above each data point. Fig. 1b shows the ratio between the number of pauses longer than 30 ms in the audio signal and the number of phones in the text input for each duration coefficient on TC_B, FS_C, FS_B and *GT* (by nature, this ratio do not vary with duration control for FS_C, FS_B and *stretching*).

Concerning elongation coefficients (Fig. 1a), FS_C follows the diagonal: as expected, frames are linearly duplicated through duration control for any class of phonemes. On the contrary, *GT* data displays non-linear behaviours that are consistent with [12] findings. These behaviours are partly followed by the embedding bias-controlled model. Looking first at slower



(a) *Elongation coefficient is the mean phone elongation compared to the unbiased voice. * indicates a significant difference with stretching.*



(b) *Silences proportion is the ratio between the number of silences in the audio signal and number of phones in the text input.*

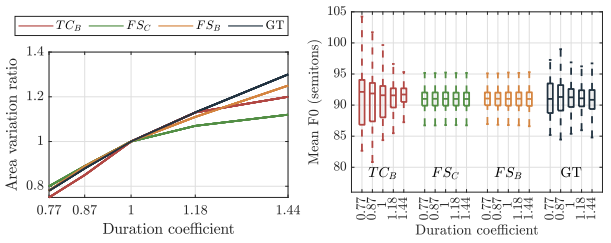
Figure 1: *Impact of duration control for each model and GT.*

speaking rates ($m_i > 1$), *GT* displays a saturation for final vowels and silences whose mean durations are large for average speaking rate (125 ms and 213 ms, respectively) and weakly lengthened as the speaking rate decreases. This behaviour has been learnt by TC_B and FS_B. Regarding other vowels and all consonants, *GT* shows a linear lengthening with duration control but to a lesser extent than *stretching*. This is compensated by the introduction of pauses in the *GT* signals: Fig. 1b displays three times more pauses in *GT* when the speaking rate is 1.44 times slower. Conversely, FS_B is unable to add any pauses in the signal, and the effect is negligible for TC_B. Alternatively, both models compensate by expanding the vowels longer than the stretch (Fig 1a). On consonants, TC_B seems to have learnt *GT* behaviour, wile FS_B follows the *stretching* trend. Looking now at higher speaker rates ($m_i < 1$), *GT* final vowels are preserved while silences are dramatically shortened or deleted (Fig. 1b). This behaviour was not replicated by any model. For other vowels and consonants, *GT* and all models follow a linear shortening of phones matching *stretching*.

Globally, *GT* duration modification is mainly performed with pauses addition and deletion, that are hardly managed by the embedding bias-controlled models. Regarding the observed non-linearity per class of phonemes, TC_B follows at best the *GT* behaviours, even though it compensates for the lack of pause addition by vowel lengthening. Both TC_B and FS_B follows the saturation of final vowels and pauses that are imposed by the data distribution, but FS_B mainly follows the *stretching* behaviour otherwise, showing that *TC* better models non-linearities in duration modification than *FS*, when using a similar embedding-bias control policy.

4.3. Co-variations of acoustic parameters

To investigate the co-variations of acoustic parameters with duration control, we first derived F_1 and F_2 values for all /a,i,u/



(a) *Vocalic triangle area.* (b) *Fundamental frequency.*

Figure 2: *Acoustic parameter variations by model and by speed.*

vowels present in the synthesised signals with the different models and duration coefficients. We then derive the area between the three vowels on the F_1-F_2 plane. The ratios between the area obtained for each duration coefficient and without duration modification are reported in Fig. 2a for each model and *GT*. For higher speaking rate, a similar linear compression of the vocalic triangle is observed for all models and *GT*, typical of an undershooting of the vowel targets. For lower speaking rates, *GT* displays an expansion of the vocalic triangle, which is successfully replicated by FS_B and TC_B, with a slight saturation for the 1.44 coefficient. FS_C shows a stronger saturation.

Fig. 2b shows mean utterance F_0 values per model and duration coefficient. *GT* data shows an increase in F_0 median and variability with highest speaking rates, which are well replicated by TC_B. Conversely, none of the *FS* models display any co-variation of F_0 with duration control. Overall, co-variations of features learnt in an unsupervised way, like formants, are well replicated by both models, while the the F_0 and duration prediction tasks implemented in *FS* lead the latter to ignore the co-variations between those parameters.

4.4. Listening Experiment

To investigate the effect of segmental and supra-segmental variations and co-variations of prosodic parameters on perception, we conducted a listening experiment where each model was evaluated against the *stretching* method. A CMOS protocol was followed [28], where participants were presented a (model,*stretching*) pair and asked which of these voice speed renderings felt the most natural. Each pair consisted of one sentence synthesised with one of the three models (FS_C, FS_B, TC_B) and one of the four duration coefficients (0.77, 0.87, 1.18, 1.44) against its *stretching* counterpart. Order of presentation was counterbalanced. In total, 3 models $\times$ 4 duration coefficients $\times$ 18 sentences $\times$ 2 order of presentation = 432 pairs were evaluated. 42 participants recruited on Prolific [29] took part in the experiment, and each evaluated 72 stimuli following a Latin Square design so that every model, duration and sentence was equally seen by each subject. Table 2 reports the averaged CMOS obtained for each model and duration coefficient. A positive value indicates that the model was preferred over *stretching*, and conversely. A non-parametric Kruskal-Wallis test showed a significant effect of both duration control and models on the CMOS ($p < 0.001$). Post-hoc Wilcoxon tests by pairs were applied and a star in the Table indicates that the model shows a statistically different CMOS than the other two models for this duration coefficient ($p < 0.001$).

Table 2: *CMOS of duration control methods against stretching of unbiased synthesis from the same model.*

Model	0.77	0.87	1.18	1.44
TC_B	**-0.818***	**-0.544***	**0.525***	-0.004
FS_C	-0.079	0.048	**0.171***	**0.623***
FS_B	-0.075	-0.048	**-0.175***	-0.159

Overall, FS_B was considered as similar as *stretching* while TC_B shows more contrasting results, supporting that subjects were sensitive to the segmental and supra-segmental co-variations that are globally better modelled by TC_B. For higher speaking rates, TC_B was significantly less preferred than *stretching*. The prosodic parameters analysis highlighted a difference in F_0 variability between models for higher speaking rate may explain this failure. A further analysis of the training set showed that highest speaking rates often correspond to the expressive reading of dialogs between characters. Without any residual encoder to segment this paralinguistic information apart from text input, *TC* may have learnt an averaged representation of these characters, resulting in an unnatural speech depreciated by participants. By contrast, TC_B is preferred to *stretching* with the 1.18 duration coefficient. With this coefficient, the main difference between models lays in the non-linearity of phone duration (Fig. 1a), where TC_B closely matches the behaviour of *GT*. This is a case where the learning of co-variation is in favour of naturalness. Reaching the 1.44 duration coefficient, both embedding bias-controlled models are equally rated as *stretching*, while FS_C is preferred. We showed that at this speaking rate the addition of pauses in the signal is essential to prevent the over-lengthening of vowels sounds observed for TC_B and FS_B that could have been perceived as unnatural. Conversely, even though FS_C cannot add supplementary pauses, it has the ability to lengthen them to a greater extent. The preference of FS_C over *stretching* could also be due to a better conservation of phone transitions, that is yet to be verified.

5. Conclusions and Discussion

We proposed a method for the analysis of the embedding space of an encoder-decoder TTS model to derive an embedding bias that is applied to control a given prosodic parameter. It aims at 1) explicitly targeting a specific prosodic parameter, in opposition to reference encoders; 2) preserve the segmental and supra-segmental variations and co-variations in speech, contrary to learnt prosodic control models. Evaluation was performed on the control of speaking rate on both attention-based (*TC*) and duration predictor based (*FS*) methods. Objective analyses showed that while the prosodic parameters estimation implemented in *FS* cleared its embedding space of most of their corresponding segmental and supra-segmental co-variations, *TC* successfully modelled this information, and this was well perceived in a listening test. The possibility to add or remove pauses while modifying the speaking rate appears essential in order to model the natural behaviour of speech. Models that use explicit phonetic inputs (ex: *FS*) negate this phenomenon. Future works should elaborate on how to give this degree of freedom to synthesis models. This multi-dimensional segmental and supra-segmental prosodic parameter variations introduced by the embedding bias control invites to propose more feature-centred evaluations in the future, in conjunction with the control of other prosodic parameters.

6. Acknowledgements

This research has received funding from the BPI project THERADIA and MIAI@Grenoble-Alpes (ANR-19-P3IA-0003). This work was granted access to HPC/IDRIS under the allocation 2021-AD011011542R1 made by GENCI.

7. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proceedings of Interspeech*, Stockholm, Sweden, August 21-24 2017, pp. 4006–4010. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1452

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *ICASSP*. IEEE, 2018, pp. 4779–4783.

[3] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[4] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[5] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2410–2419.

[6] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP*. IEEE, 2019, pp. 3617–3621.

[7] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," *arXiv preprint arXiv:1803.09017*, 2018.

[8] D. Stanton, Y. Wang, and R. Skerry-Ryan, "Predicting expressive speaking style from text in end-to-end speech synthesis," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 595–602.

[9] D. Potdevin, "Vers des agents conversationnels animés sociaux: Quelle influence de l'intimité virtuelle sur l'expérience utilisateur et la relation-client?" Ph.D. dissertation, Université Paris-Saclay, 2020.

[10] D. S. R. Mohan, V. Hu, T. H. Teh, A. Torresquintero, C. G. Wallis, M. Staib, L. Foglianti, J. Gao, and S. King, "Ctrl-P: Temporal Control of Prosodic Variation for Speech Synthesis," in *Proc. Interspeech 2021*, 2021, pp. 3875–3879.

[11] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," *arXiv preprint arXiv:1803.09047*, 2018.

[12] N. Campbell, "Multi-level timing in speech," Ph.D. dissertation, Sussex University, U.K. Department of Experimental Psychology, 1992.

[13] M. Kim, S. J. Cheon, B. J. Choi, J. J. Kim, and N. S. Kim, "Expressive Text-to-Speech Using Style Tag," in *Proc. Interspeech 2021*, 2021, pp. 4663–4667.

[14] Y.-J. Zhang, S. Pan, L. He, and Z.-H. Ling, "Learning latent representations for style control and transfer in end-to-end speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6945–6949.

[15] W.-N. Hsu, Y. Zhang, R. J. Weiss, Y.-A. Chung, Y. Wang, Y. Wu, and J. Glass, "Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization," in *ICASSP*. IEEE, 2019, pp. 5901–5905.

[16] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *arXiv preprint arXiv:1806.04558*, 2018.

[17] P. van Rijn, S. Mertes, D. Schiller, P. M. Harrison, P. Larrouy-Maestri, E. André, and N. Jacoby, "Exploring Emotional Prototypes in a High Dimensional TTS Latent Space," in *Proc. Interspeech 2021*, 2021, pp. 3870–3874.

[18] A. Perquin, E. Cooper, and J. Yamagishi, "An investigation of the relation between grapheme embeddings and pronunciation for tacotron-based systems," *arXiv preprint arXiv:2010.10694*, 2020.

[19] N. Tits, F. Wang, K. E. Haddad, V. Pagel, and T. Dutoit, "Visualization and interpretation of latent spaces for controlling expressive speech synthesis through audio analysis," *arXiv preprint arXiv:1903.11570*, 2019.

[20] NVIDIA, "Tacotron2 implementation." [Online]. Available: https://github.com/NVIDIA/tacotron2

[21] M. Lenglet, O. Perrotin, and G. Bailly, "Impact of segmentation and annotation in french end-to-end synthesis," in *11th ISCA Speech Synthesis Workshop*. ISCA, 2021, pp. 13–18.

[22] K. Kastner, J. F. Santos, Y. Bengio, and A. Courville, "Representation mixing for tts synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5906–5910.

[23] C.-M. Chien, "FastSpeech2 implementation." [Online]. Available: https://github.com/ming024/FastSpeech2

[24] G. Bailly, O. Perrotin, and M. Lenglet, "Ressources for End-to-End French Text-to-Speech Blizzard challenge," Mar. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4580406

[25] M. Lenglet, O. Perrotin, and G. Bailly, "Modélisation de la parole avec tacotron2 : Analyse acoustique et phonétique des plongements de caractère," in *Actes des Journées d'Etudes sur la Parole (JEP)*, Noirmoutiers, France, June 13-17 2022.

[26] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9, pp. 341–345, 2001.

[27] J. B. Kruskal, *Multidimensional scaling*. Sage, 1978, no. 11.

[28] I. T. Union, "Methods for objective and subjective assessment of quality," International Telecommunication Union, Tech. Rep. ITU-T P.800, 1998.

[29] S. Palan and C. Schitter, "Prolific.ac—a subject pool for online experiments," *Journal of Behavioral and Experimental Finance*, vol. 17, pp. 22–27, 2018.