Impact of Segmentation and Annotation in French end-to-end Synthesis

Martin Lenglet, Olivier Perrotin, Gérard Bailly

11th ISCA Speech Synthesis Workshop (SSW 11)
26-28 August 2021, Budapest, Hungary

Impact of Segmentation and Annotation in French end-to-end Synthesis

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

Audio books are commonly used to train text-to-speech models (TTS), as they offer large phonetic content with rather expressive pronunciation, but number and sizes of publicly available audio books corpora differ between languages. Moreover, the quality and accuracy of the available utterance segmentations are debatable. Yet, the impact of segmentation on the output synthesis is not well established. Additionally, utterances are generally used individually, without taking advantage of text level structuring information, even though they influence speaker reading. In this paper, we conduct a multidimensional evaluation of Tacotron2 trained on different segmentations and text level annotations of the same French corpus. We show that both spectrum accuracy and expressiveness depend on the segmentation used. In particular, a shorter segmentation, in addition with the annotation of paragraphs, benefits to spectrum reconstruction at the detriment of phrasing. Multidimensional analysis of mean opinion scores obtained with a MUSHRA-experiment revealed that phrasing was relatively more important than spectrum accuracy in perceptual judgement. This work serves as evidence that particular attention must be given to models evaluation, as well as how to use the training corpus to maximize synthesis characteristics of interest.
Index Terms: Speech Synthesis, French TTS, mixed-inputs TTS, French dataset

1. Introduction

In recent years, deep learning met huge success in language-related applications. In particular, state-of-the-art text-to-speech (TTS) models [1, 2, 3] coupled with neural vocoders [4, 5] achieve synthesis quality close to natural speech. As always with deep learning, the quality of the output heavily depends on the dataset used for training. The common approach of neural TTS, seen in events like Blizzard Challenge [6], is to compare multiple models on the same corpus to evaluate the resulting synthesis quality. This process minimizes the importance of input data structuring, which ultimately shapes the output of any deep learning model. One complementary work is to evaluate multiple segmentations of data structuring on the same TTS model. This paper adopts this approach.

Publicly available corpora designed to train TTS [7, 8, 9] are generally composed of audio book extracts read by one or more speakers, segmented in thousands of utterances. Utterances' lengths vary between 1 to 20 seconds, with boundaries often matching sentences, but not always. Even if these databases have been used to train state-of-the-art speech models [3, 10], long utterances may not be the best candidates to train TTS: (i) Large batch size with long utterances rely on high computation memory. (ii) Learning long-term dependencies is a challenging task for sequential models [11]. (iii) Style control, which is an increasing demand of the field, massively uses utterance level style embeddings [12, 13], which means that the shorter the utterances, the finer it is possible to tune speech style at inference time. These reasons made us consider a shorter segmentation may be better suited to train TTS efficiently.

Proposing a new segmentation gives us the opportunity to integrate specific annotations in the input data to give models relevant context information regarding the corresponding speech to produce: (i) End of paragraph are generally associated with specific phrasing modifications from the speaker, and are then worth noticing during training. (ii) In French, silent letters and optional liaisons are common, which are additional difficulties to train a TTS model on orthographic inputs alone. The addition of phonetic annotations contributes to alleviate this issue, and has shown to benefit to both transcriptions [14].

This paper presents a multidimensional comparison between the proposed segmentation and annotation of the LibriVox French corpus [15] and the original segmentation from M-AILABS [7], used to train the same Tacotron2 [1]. We evaluate the phrasing and spectral accuracy of each model. These objective measurements are paired with mean opinion scores evaluated through a MUSHRA-like experiment [16].

2. Related Work

To our knowledge, there is no publicly available French Tacotron2. Recent studies published on French synthesis focus on concatenation based TTS [17] or use Deep Convolutional TTS (DCTTS) [18]. DCTTS is a fully convolutional neural TTS, whose initial purpose was to alleviate the need for high computational power, while enabling quick training on smaller database. Although synthesis reaches acceptable standards, the overall quality does not match more recent models [1, 2, 3].

The later TTS explore the well established encoder-decoder architecture: the encoder converts the input sequence into a hidden representation that the decoder uses to generate mel-spectrogram frames. As an interface between the two, Tacotron2 [1] employs a location-sensitive attention [19] module which computes a fixed length vector for each decoder step. The encoder adopts an approach that is similar to the classical language model processing pipeline: the input sequence is passed through three convolutional layers that compute local pattern, followed by bidirectional LSTM. Alternatively, Transformer TTS [2] and Fastspeech [3] introduce self-attention and multi-head attention layers as a replacement for recurrent units. These three models produce synthetic speech of similar quality [3]. We chose Tacotron2 for its relative ease to implement and straight training process. Additionally, Tacotron2 shows promising results for expressive control [12, 13], which is also one of our short term goal.

Although mean opinion scores are generally used to assess the global quality of TTS, this evaluation takes multiple aspects of speech into account: phonetic correctness and intelligibility,

10.21437/SSW.2021-3

spectral smoothness, expressiveness, etc. These clues may not vary conjointly, which means that the use of a single metric may not be sufficient. [20, 21] employ multidimensional scaling (MDS) [22] to extend the quality analysis of TTS models. This paper prolongs this perspective.

3. Proposed Method

This section presents the original baseline and the new segmentation proposed from the French LibriVox dataset, and the modifications added to the Tacotron2 implementation shared by NVIDIA[1]. Our implementation[2] and database[3] are available online.

3.1. Segmentation and Annotation

3.1.1. Original Database

We used the M-AILABS French dataset [7] as a starting point. This corpus includes more than 190h of recorded speech, segmented in utterances from 1s to 20s, given with corresponding orthographic transcripts. Recordings come from the free public domain audio books LibriVox database [15]. We selected a subset of the recordings made by Nadine Eckert-Boulet (NEB), for a total duration of 34h. Each book duration and corresponding number of utterances are given in Table 1. Audio files are originally sampled at 16000Hz, but we re-sampled them at 22050Hz.

Table 1: *Books duration (and number of utterances) for original and new segmentation of the M-AILABS French corpus.*

Book	**Original**	**New segmentation**
Les Mystères de Paris	22:31:27 (12285)	21:37:21 (25458)
Mme Bovary	11:39:50 (5775)	11:08:55 (12781)
Total	34:11:17 (18060)	32:46:16 (38239)

The orthographic transcript is given by the Gutenberg Project[4]. It is worth mentioning that NEB does not always strictly follow the original text. Some miss-spelling remain (for example: "precepteur" is said instead of "percepteur"), as well as some omissions. These miss-alignments correspond to 0.1% of the original corpus. We did not correct any of those transcripts for the baseline. Though, we spelled out all texts, including frequently used abbreviations in French ("M.": "Monsieur", "Mlle": "Mademoiselle", "n°": "numéro" and "etc": "et cetera"), and numbers ("1838": "dix-huit cent trente-huit"). Two punctuation marks were also replaced to stand as a single unique character: "..." was replaced by "$\sim$", "–" by "¬".

Each clip was originally bounded with 500ms of silence (zeros in the waveform) at the beginning and the end. These silences do not correspond to the recordings, but have been artificially added to each audio clip after segmentation. To limit the duration of initial and final silences in the synthesis, we truncated these silences at 130ms. This duration matches the initial and final silence lengths found in other speech databases such as LJspeech [8].

3.1.2. Re-segmentation

To reduce the average duration of utterances, we first restore the initial audio books chapters structure by aligning the original text from the Gutenberg Project with the recordings from LibriVox. As for the original segmentation, all texts are spelled out, but previously mentioned miss-spelling and omissions are now manually corrected. In addition, end of paragraphs are annotated with the punctuation mark "§", which is introduced after the last punctuation mark preceding each carriage return. Ends of paragraphs are accompanied by phrasing patterns of NEB, that are worth highlighting in the training corpus. For instance, Table 2 shows F0 and elongation of the final syllable before ends of paragraph vs. paragraph-internal periods, as well as their values for the following syllable. The last syllable is generally longer before the end of paragraph, and the F0 gap across the boundary is increased (6.45 vs. 5.11 semitones respectively).

Table 2: *Comparison of F0 and elongation of syllable [23] around ends of paragraph (.§) and intermediate periods (.).*

		Syllable	
		Previous	**Following**
Elongation (%)	**.**	+184	+21
	.§	+218	+24
F0 (semitone)	**.**	1.96	7.01
	.§	0.96	7.41

Chapters are then segmented based on silences of at least 400ms. This duration usually corresponds to pauses made between speaking turns in conversations [24]. 94.56% of silences coincide with punctuation marks. For the others, a comma is added at the end of the utterance. 130ms of ambient silence from the recording are kept at the beginning and the end of each utterance. Timestamps were hand-checked for each utterance to ensure optimal segmentation. Table 1 shows duration and number of utterances of the obtained segmentation. Note that the proposed segmentation is 01:25:01 shorter than the original, due to the reduction of intra-utterance silences, but that reduction does not impact either the text read nor the speaking rate.

Fig.1 gives the distribution of utterances length of the original and the proposed segmentation. Median utterance length (resp. first and third quartiles) are reduced from $6.44s$ ($3.88s$ and $9.26s$) to $2.77s$ ($1.89s$ and $3.95s$). 82.5% of utterances of the new segmentation last between $1s$ and $5s$, and 0.25% of utterances last more than $10s$. 1336 utterances are unchanged, which corresponds to 7.4% and 3.5% of the original and new segmentation respectively.

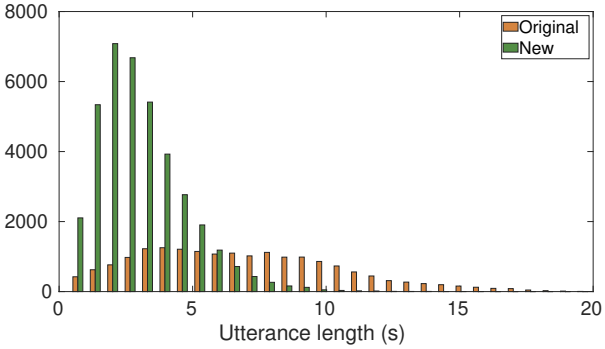


Figure 1: *Distribution of utterances length of original and new segmentation.*

[1] https://github.com/NVIDIA/tacotron2
[2] https://github.com/MartinLenglet/Tacotron2
[3] https://zenodo.org/record/4580406#.YI_qIyaxXmE
[4] https://www.gutenberg.org/

3.1.3. Phonetic Annotation

Training of both orthographic and phonetic transcripts, called representation mixing, enables to use both input types in the same utterance at inference time, and thus remove some ambiguities on particular issues, without the need for the whole phonetic transcript of the speech to synthesize. For instance, NEB performs numerous optional liaisons (22999 liaisons in the corpus of which 9597 [z], 9029 [t] and 3412 [n]), in particular bridging 844 infinitives and prepositions with [ʁ/]. Yet, these liaisons are not systematic, and adding the possibility to choose if the liaison is being made at inference time (as part of a style component) would be interesting. To study the impact of phonetic annotation, hand-crafted phonetic alignment is performed on the whole new segmentation.

3.2. Modifications of Tacotron2

3.2.1. Representation mixing

We introduce the mixed embedding matrices described in [14] in our model to give the possibility to train with both types of inputs. Contrary to [14], when the training includes phonetic inputs, input types are not mixed within the same utterance. The number of utterances is simply doubled, with the same audio file corresponding to both the orthographic and the phonetic input.

3.2.2. Gate loss correction

Synthesizing short utterances, typically one or two words, has been shown to be a challenging task for TTS models [25]. Recurrent artifacts are repetition of the last syllable, or unintelligible words. With our proposed segmentation, 5% of utterances last less than 1s, which might cause some issues during inference. To avoid this, we fine tune the training of each model with 2 modifications: (i) 9 frames of recorded ambient silence are added at the end of each utterance, in which the end-of-sequence probability is set to 1. This silence originates from the pause following each utterance. (ii) a multiplying factor is added to the gate loss error before back-propagation. We empirically found that these modifications correct previously mentioned artifacts, and improve the overall synthesis quality. The benefits of these modifications are evaluated in section 4.

4. Experiments and Results

4.1. Experimental Setup

The 6 models trained for this experiment are presented below:

- O **and** O_g are trained on the original segmentation from M-AILABS for 200 epochs.
- N **and** N_g are trained on the new segmentation proposed in section 3.1.2, with only orthographic inputs for 200 epochs.
- P **and** P_g are trained on the new segmentation proposed in section 3.1.2, with both orthographic and phonetic inputs for 100 epochs, since each epoch corresponds to twice the number of utterances of the orthographic models.

Models annotated $_g$ are fine-tuned with the gate loss correction. The multiplying factor is set to 10 for these models. This correction is introduced for the last quarter of the training epochs. Before that separation, only one model is trained using warm-start from the English model trained on LJSpeech shared by NVIDIA. The postnet is bypassed during the first 10 epochs, and the learning rate is fixed at 10^{-3}. This phase enables the model to initiate a coarse transition from English to French. Then the postnet is reactivated and the learning rate decreases exponentially until reaching 10^{-5} at 90 epochs. The batch size is limited to 32, due to memory limitations with long utterances of the original segmentation, and thus is set to 32 for all models. Batches are randomly picked among utterances of approximate same length.

We pick 5% of the original corpus as test set. To ensure a fair comparison between models, these 903 utterances are randomly selected among the 1336 common utterances between the original and the new segmentation. Thus, the amount of speech seen by each model during training is rigorously the same. Only the orthographic transcript of the test set is used in this section, even for models P and P_g. Note that this test set does not favor the new segmentation: phonetic inputs and paragraphs markers are not used.

The vocoder used is WaveRNN [5]. WaveRNN is faster and demands less resources than the original WaveNet [4] used by [1], and still provides a good voice quality [26]. We trained WaveRNN from scratch for 1000 epochs on the new segmentation from Table 1 with a learning rate of 10^{-4}. Then we fine-tuned the model with 520 more epochs at a learning rate of 10^{-5}.

4.2. Objective measurements

4.2.1. Accuracy

We evaluate the spectral accuracy of each model through the proximity of the generated spectra with the *vocoded* ground truth (GT). Since syntheses differ in length, mel-spectrograms are first aligned by dynamic time warping (DTW) [27]. Mean squared error (MSE) on aligned spectrograms are then computed and averaged on the test set; results are shown in Fig. 2.

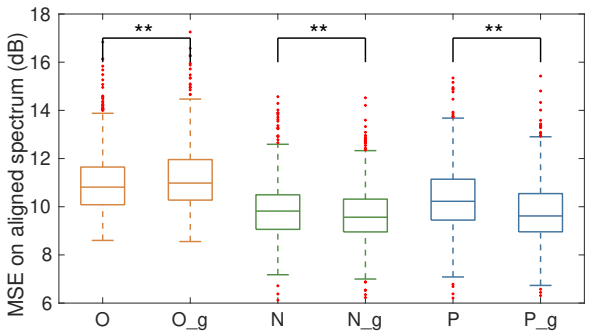


Figure 2: *Mean squared error between models and ground truth, calculated on mel-spectrograms aligned by dynamic time warping. ** indicates a significant effect of the gate loss correction according to Tukey-Kramer test ($p < 0.05$).*

The model has a statistical effect on the computed distances according to a one-way ANOVA ($F = 246.5$, $p < 0.001$). Tukey-Kramer multiple comparisons show that all pairs are statistically different, except P_g/N and P_g/N_g. The gate loss correction has a significant impact on all models. The new segmentation decreases the spectral distortion, with a beneficial contribution of the gate loss correction in this case. On the other hand, this correction decreases the spectral accuracy of the model trained on the original segmentation.

4.2.2. Phrasing

Pauses position and duration contribute to the expressiveness of speech [28]. We computed mean speech and silence duration across the whole synthesised test set for each model and for

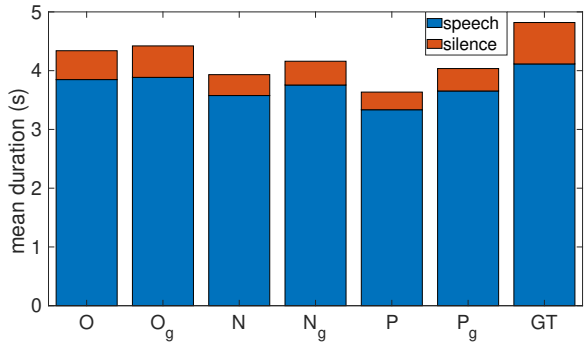


Figure 3: *Mean utterance duration on the whole test set for each model.*

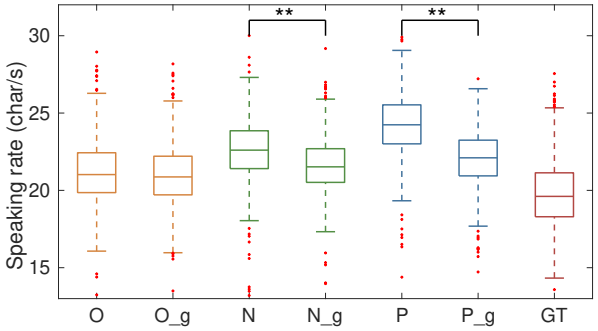


Figure 4: *Speaking rate of each model, calculated on each utterance of the test set. Speaking rate is estimated in characters per second, pause durations are not taken into account. ** indicates a significant effect of the gate loss correction ($p < 0.05$).*

GT. By extension, this calculation also enables us to estimate the speaking rate of each model on the test set. Mean utterance duration and speaking rate are shown in Fig. 3 and Fig. 4 respectively.

Models trained on the new segmentation do not exhibit the same temporal behavior than models trained on the original segmentation. Utterances mean duration is smaller with the new segmentation ($3.93s$ and $3.64s$ compared to $4.44s$ for N, P and O respectively). Silences duration are also proportionally smaller: 9.2%, 8.2% and 11.3% for N, P and O respectively. As a result, the speaking rate increases with the new segmentation. Note that the speaking rate of all models is significantly higher than GT. The gate loss correction tends to reduce the differences observed compared to GT. Not only silences duration are increased, but also speech duration, resulting in a lower speaking rate. This decrease is statistically significant for the new segmentation, but not for the original. All other pairs are significantly different according to Tukey-Kramer multiple comparisons.

Longer pauses observed with O and O_g may result from the intra-utterance pauses frequency and duration in the original segmentation provided by M-AILABS. In that case, models are trained on audio clips that sometimes contain pauses longer than 1s, and thus reproduce that behavior during inference. On the contrary, the re-segmentation processing avoids intra-silences longer than 400ms, resulting in a more straight-forward synthesis.

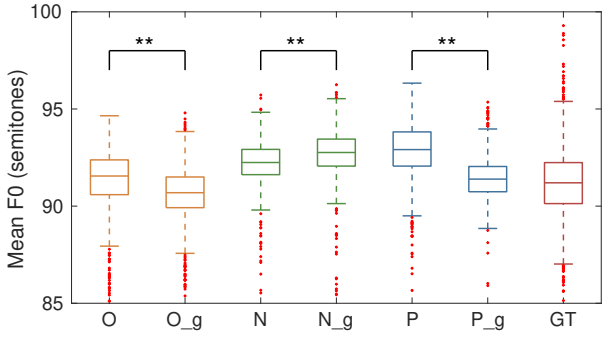


Figure 5: *Mean fundamental frequency calculated on voiced sections of each utterance of the test set. ** indicates a significant effect of the gate loss correction ($p < 0.05$).*

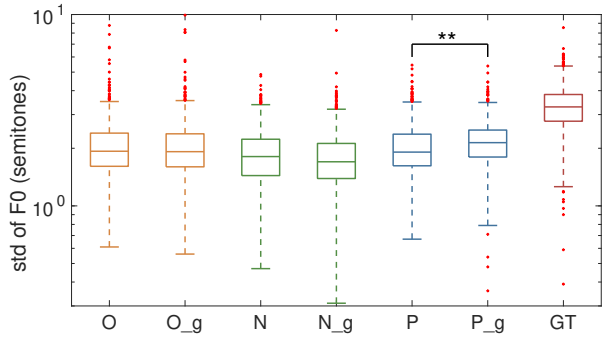


Figure 6: *Standard deviation of fundamental frequency calculated on each utterance of the test set. ** indicates a significant effect of the gate loss correction ($p < 0.05$).*

4.2.3. Pitch

As additional prosody measurements, we evaluate the pitch of each model using the Praat software [29]. The mean fundamental frequency (F0) and standard deviation of F0 is measured on voiced sections for every utterance of the test set. Results are given in Fig. 5 and Fig. 6 respectively.

One-way ANOVA shows a statistical effect of the model on both mean F0 and standard deviation of F0. Regarding mean F0, Tukey-Kramer multiple comparisons show that all pairs differ significantly, except O/P_g, O/GT and N_g/P. As to standard deviation of F0, only phonetic models P and P_g exhibit a significant effect of the gate loss correction, while both P and P_g are not statistically different from O and O_g. N and N_g have significantly lower standard deviation than all other models.

The new segmentation increases mean F0, but this effect is partially compensated when training the model on mixed inputs with gate loss correction. Similarly, the gate loss correction induces a lower mean F0 when training on the original segmentation. None of the presented models show standard deviation of F0 similar to GT, which might lead to less expressive synthetic voices.

4.3. Subjective evaluation

In accordance with objective measurements presented in section 4.2, 3 models were selected to evaluate the mean opinion scores through a MUSHRA-like experiment [16]. We keep only models that have been fine-tuned with gate loss correction, as they generally exhibit the closest proximity with GT behavior.

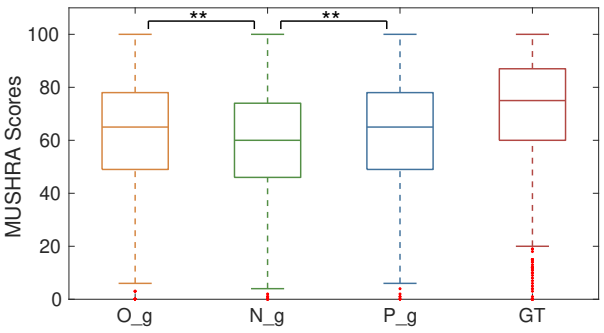
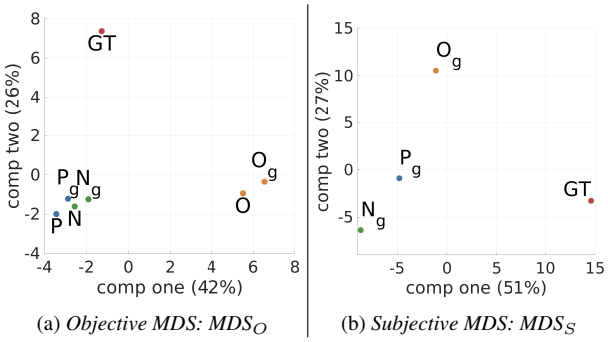


Figure 7: *MUSHRA results. ** indicates a significant difference between models ($p < 0.05$).*



(a) *Objective MDS: MDS_O* (b) *Subjective MDS: MDS_S*

Figure 8: *Multidimensional scaling of distances between pairs conditions. Left and right graphs show objective and subjective distances respectively. Proportions of variance explained are given for each component.*

GT is added as high anchor for the MUSHRA. This perceptive test was performed online using the webMUSHRA framework [30]. Utterances containing less than 7 words and more than 23 words were excluded from this test to keep only the central 90% of the test set length distribution. 60 utterances were randomly selected in the remaining test set, with equivalent representation of utterance lengths in the selection. 13 of the selected utterances contained one phonetic mistake (5 in O_g, 3 in P_g, and 5 in all models), and were replaced before the experiment. Participants were separated in 2 groups, each group listened to 30 out of the 60 selected utterances. For each utterance, participants were given the original text input, and were asked to evaluate the 4 given conditions (3 models + GT) according to the voice quality. No explicit reference was given during the listening. The experiment began with 5 minutes of training during which participants listened to a variety of synthesis that they were about to hear during the experiment and learned how to use the webMUSHRA interface. Audio examples are available online [5]. 44 participants recruited on Prolific [31] and aged 18-65 took part in the experiment. Participants were French native speakers, and had little or no previous experience with listening tests. Results of the MUSHRA are given in Fig.7

We compared the median score of each model using a Wilcoxon rank sum test. Differences are significant if $p < 0.05$. GT exhibits a significantly higher score than the 3 evaluated models. N_g scores significantly lower than all other models. No statistical differences are shown between O_g and P_g.

4.4. Multidimensional analysis

Despite the differences on specific expressiveness clues measured in section 4.2, subjective evaluation performed in section 4.3 does not exhibit a clear perceptive preference for one of the models O_g or P_g. To explore implicit dimensions of the evaluation of the models, we use a multidimensional analysis of the distances computed between each model and GT. These distances are evaluated on both objective and subjective measurements:

- **Subjective distances:** absolute score differences between all possible condition pairs evaluated in the MUSHRA, averaged across all participants and all utterances.
- **Objective distances:** MSE between all possible conditions pairs computed on mel-spectrograms aligned by DTW [27]. Objective distances are averaged across all 903 utterances of the test corpus.

[5] http://www.gipsa-lab.fr/~martin.lenglet/segmentation_impact/index.html

Table 3: *Correlation coefficients between objective measurements and components of MDS. * and ** indicate $p < 0.1$ and $p < 0.05$ respectively. ASE: aligned spectrum error, SR: Speaking rate, PD: pauses duration.*

MDS	**Dim**	**objective measurements**				
		ASE	SR	PD	mean F0	std F0
Obj	1	**0.90****	-0.47	0.44	**-0.71***	-0.06
	2	0.63	**-0.74***	**0.89****	-0.43	**0.97****
Subj	1	0.89	**-0.93***	**0.96****	-0.50	**0.98****
	2	0.97	-0.02	0.13	-0.83	-0.17

Then, we projected the two obtained distances matrices in two independent 2-dimensions space using classical Multidimensional scaling (MDS) [22]. To give a better idea of the impact of the gate loss correction, both corrected and non-corrected models were included in the objective MDS. Subjective and objective MDS (named MDS_S and MDS_O respectively in the following) are given in Fig.8.

Correlations between objective measurements computed in section 4.2 and the components of both MDS are estimated. Correlations coefficients are given in Table 3. Note that GT is not considered for correlation with aligned spectrum error (ASE). Correlation coefficients indicate that prosodic clues like pauses duration and standard deviation of F0 are closely related to the second component of MDS_O, but to the first component of MDS_S. On the other hand, spectral accuracy measurements ASE and mean F0 are correlated to the first component of MDS_O, and similarly for the second component of MDS_S, even if this tendency is not significant. Two main dimensions emerge in both evaluations: spectrum accuracy and expressiveness. The axis inversion (and associated portion of variance explained) tends to show these dimensions are not given the same importance in the perceptive judgement than in the objective measurement. As a result, the proximity of spectrum quality observed between GT and models trained on new segmentation on the first component of Fig.8a is downgraded to the second component of Fig.8b. Respectively, expressiveness is given more importance in the perceptive test than it is in the objective measurements, resulting in O_g being closer to GT in the first component of Fig.8b. Fig.8a emphasizes the benefits of the proposed gate loss correction, as all models annotated $_g$ are closer to GT on the expressiveness dimension.

5. Conclusions and Discussion

We have proposed a shorter segmentation of the French M-AILABS corpus and compared the training of Tacotron2 on both original and new datasets. Through multi dimensional evaluation, we have shown that the way speech data are segmented impacts both quality and expressiveness factors in opposite directions. Future works should elaborate on how to combine the advantages of both segmentation with curriculum training. An important contribution of this work is the addition of the gate loss correction as a fine tuning of the model, which contributes to improve prosodic aspects of the synthesized speech. The use of multidimensional analysis of mean opinions scores introduces relevant nuances to the MUSHRA results. The structuring of the subjective notation latent space, as well as the prediction of positions in this space thanks to objective measurements should be the focus of future works.

6. Acknowledgments

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] 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.

[2] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.

[3] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[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] X. Zhou, Z.-H. Ling, and S. King, "The blizzard challenge 2020," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 1–18.

[7] I. Solak, "The M-AILABS speech dataset," https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/, 2019.

[8] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[9] P.-E. Honnet, A. Lazaridis, P. N. Garner, and J. Yamagishi, "The siwis french speech synthesis database. design and recording of a high quality french database for speech synthesis," Idiap, Tech. Rep., 2017.

[10] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang *et al.*, "A comparative study on transformer vs rnn in speech applications," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 449–456.

[11] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber *et al.*, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," in *A field guide to dynamical recurrent neural networks*, Y. Hochreiter, Sepp Bengio, P. Frasconi, J. Kolen, and S. Kremer, Eds. IEEE Press, 2001.

[12] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[13] 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*. IEEE, 2019, pp. 6945–6949.

[14] 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.

[15] J. Kearns, "Librivox: Free public domain audiobooks," *Reference Reviews*, 2014.

[16] I. BS, "1534-1, method for the subjective assessment of intermediate quality level of coding systems," *International Telecommunications Union, Geneva, Switzerland*, vol. 14, 2003.

[17] M. Shamsi, J. Chevelu, N. Barbot, and D. Lolive, "Corpus design for expressive speech: impact of the utterance length," in *Speech Prosody*. ISCA, 2020, pp. 955–959.

[18] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *ICASSP*. IEEE, 2018, pp. 4784–4788.

[19] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *arXiv preprint arXiv:1506.07503*, 2015.

[20] C. Mayo, R. A. Clark, and S. King, "Multidimensional scaling of listener responses to synthetic speech," in *Interspeech*. ISCA, 2005, pp. 1725–1728.

[21] F. Hinterleitner, S. Möller, C. Norrenbrock, and U. Heute, "Perceptual quality dimensions of text-to-speech systems," in *Interspeech*, 2011.

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

[23] P. Barbosa and G. Bailly, "Characterisation of rhythmic patterns for text-to-speech synthesis," *Speech Communication*, vol. 15, no. 1, pp. 127–137, 1994.

[24] G. Bailly and C. Gouvernayre, "Pauses and respiratory markers of the structure of book reading," in *Interspeech*, 2012, pp. 2218–2221.

[25] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *arXiv preprint arXiv:1710.07654*, 2017.

[26] P. Govalkar, J. Fischer, F. Zalkow, and C. Dittmar, "A comparison of recent neural vocoders for speech signal reconstruction," in *Proc. 10th ISCA Speech Synthesis Workshop*, 2019, pp. 7–12.

[27] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1. IEEE, 1993, pp. 125–128.

[28] E. Godde, G. Bailly, D. Escudero, M.-L. Bosse, and E. Gillet-Perret, "Evaluation of reading performance of primary school children: Objective measurements vs. subjective ratings," in *International workshop on child computer interaction (WOCCI)*, 2017.

[29] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, pp. 341–345, 01 2001.

[30] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webmushra—a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, 2018.

[31] 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.