The MARA corpus: Expressivity in end-to-end TTS systems using synthesised speech data

Adriana STAN[1], Beáta LŐRINCZ[1,2], Maria NUȚU[1,2], Mircea GIURGIU[1]
[1]Technical University of Cluj-Napoca, Romania
[2]Babeș-Bolyai University, Cluj-Napoca, Romania
{adriana.stan, beata.lorincz, maria.nutu, mircea.giurgiu}@com.utcluj.ro

***Abstract*—This paper introduces the MARA corpus, a large expressive Romanian speech corpus containing over 11 hours of high-quality data recorded by a professional female speaker. The data is orthographically transcribed, manually segmented at utterance level and semi-automatically aligned at phone-level. The associated text is processed by a complete linguistic feature extractor composed of: text normalisation, phonetic transcription, syllabification, lexical stress assignment, lemma extraction, part-of-speech tagging, chunking and dependency parsing.**
Using the MARA corpus, we evaluate the use of synthesised speech as training data in end-to-end speech synthesis systems. The synthesised data copies the original phone duration and F_0 patterns of the most expressive utterances from MARA. Five systems with different sets of expressive data are trained. The objective and subjective results show that the low quality of the synthesised speech data is averaged out by the synthesis network, and that no statistically significant differences are found between the systems' expressivity and naturalness evaluations.

***Index Terms*—speech synthesis, expressive speech corpus, end-to-end systems, deep neural networks, HTS, Merlin, Romanian**

I. Introduction

Over the past few years, with the advancement of deep learning algorithms and their application in text-to-speech synthesis (TTS), the naturalness of the synthetic voices became comparable to that of a human speaker [1]. Yet this naturalness comes at the cost of using a large training dataset. However, extended expressive speech corpora adequate for training high-quality speech synthesis systems are still scarcely found. While in the mainstream languages such data is constantly released [2], for the languages which have fewer native speakers, the development of these resources is not a focus of the research community.

When such a dataset is not available, other methods to improve the quality or the expressivity of a TTS system need to be employed. Depending on the TTS paradigm, the manner in which the naturalness or prosody of the synthetic voice can be enhanced differs. For naturalness, the common approach across all paradigms is to pretrain a model with data from multiple speakers and to fine-tune its parameters toward a target speaker [3], [4]. This makes it possible to use fewer data from the target speaker, as the model already learned an internal representation of the speech characteristics in general.

Different approaches are applied in the case of expressivity or prosody transfer from a source to a target speaker. In statistical parametric systems, only the duration and fundamental frequency (F_0) models were adapted [5]. However, in the recent end-to-end synthesis systems which commonly employ a sequence-to-sequence architecture, there is no explicit representation of the speech prosody. The models are left to learn an internal representation of it through additional observed or latent attributes. For example, [6] extends the Tacotron architecture with an embedding layer that learns a latent representation of the prosody extracted from an audio reference. The external reference contains the desired prosody style and is fed into the network during the synthesis step. Similar approaches are described in [7] and [8] where the prosody is learned by a specialised module, called Global Style Token layer. [9] applies variational autoencoders (VAEs) to learn an unsupervised representation of different speech emotions without access to the emotion labels. VAEs are also applied in [10] to inject a latent style representation provided by the recognition network of the VAE. [11] uses the same strategy to provide two levels of hierarchical latent variables which aim to explain dimensions not present in the data labels, such as speaking style, accent, background noise, and recording conditions.

In the previous studies, the latent representations require manual adjustment or audio reference during inference. Although this level of control is highly desirable, it might pose additional computational and usability problems. One different type of style or prosody transfer in neural network-based TTS systems uses specialised layers or additional linguistic information. [12] uses a hidden layer augmentation strategy which adds new neurons responsible for learning the additional features of the target speaker's style. On the other hand, [13] introduces contextual word embeddings and style id to generate a two-style neural TTS voice with limited training data.

In this study we introduce a newly developed Romanian expressive speech corpus called **MARA** and describe its contents and annotation process. Starting from this corpus, and with the aim of finding other reliable means of prosody transfer

This work was supported by a grant of the Romanian Ministry of Research and Innovation, PCCDI – UEFISCDI, project number PN-III-P1-1.2-PCCDI-2017-0818/73, within PNCDI III.

978-1-6654-2786-9/21/$31.00 ©2021 IEEE

in end-to-end systems, we evaluate the use of synthesised expressive speech as training data in a Tacotron-based [14] architecture.

The paper is organised as follows: Section II describes the MARA corpus' development and annotation. In Sections III and IV we present the methodology used to obtain synthesised expressive speech data and the speech synthesis systems built with this data, respectively. The objective and subjective evaluation of these systems is summarised and discussed in Section V, with conclusions drawn in Section VI

II. The MARA Speech Corpus

Within our group, freely-accessible Romanian text and speech corpora have been a priority in the recent years. We started with a single speaker dataset containing around 4 hours of high-quality recordings performed in an hemianechoic chamber (RSS) and aimed at developing text-to-speech synthesis systems [15]. We then extended the corpus with 16 new speakers to become one of the largest Romanian parallel speech datasets (SWARA) [16]. SWARA includes over 21 hours of data recorded in a studio environment, semi-automatically aligned at phone level. Both RSS and SWARA contain utterances read with a flat intonation pattern. To the best of our knowledge there are no other freely available Romanian speech resources adequate for TTS systems development.

Text-to-speech synthesis systems built with this data have been used in both academic and commercial deployments. However, the use of these systems is limited due to their lack of expressivity. So the next natural step in our development was to create a large speech dataset containing more dynamic intonation patterns. The result is the **MARA** corpus, derived from a professionally recorded audiobook, kindly provided to us by Cartea Sonoră.[1] The audiobook contains the reading of the entire Romanian novel "Mara" written by Ioan Slavici and published in 1906. It describes the struggles of a widow trader to improve the livelihood of her two children. The action is set in the Mureș county, under the Austrian-Hungarian occupation around 1850. The speech data comprises over 11 hours of recordings uttered by a professional female speaker sampled at 44kHz and 16bps.

The data received from the publisher contained only the audio delivered in chapter-length chunks. The first step was to manually recover its orthographic transcription and divide the chapters into shorter segments. In Romanian, belletristic style writings tend to include very long phrases. As long utterances are not suitable in sequence-to-sequence architectures, we relied on the speaker's phrase break pauses to perform the segmentation. As a result, we obtained 8185 utterances with an average length of 5 seconds corresponding to approximately 12 words. The next step was to annotate the text data with high-level linguistic information obtained from the RACAI Relate platform.[2] The information represents CONLLU formatted

[1]https://www.youtube.com/user/CarteaSonora

[2]http://relate.racai.ro

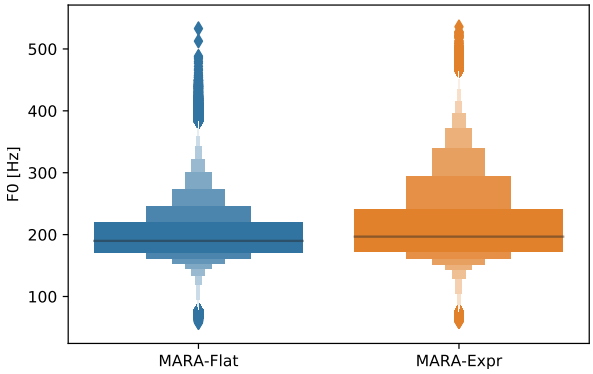

Fig. 1. Letter-value plots of the F_0 values in the MARA-Flat and MARA-Expr subsets

annotations of the utterances and includes: text normalisation, phonetic transcription, syllabification, lexical stress assignment, lemma extraction, part-of-speech tagging, chunking and dependency parsing. Using the segmentation and annotation we then trained an iterative HMM-based aligner to obtain the phone-level boundaries [17]. The accuracy of the aligner was not evaluated. The linguistic metadata is associated with each individual phoneme through the use of the extended HTS label format.[3]

Because the reading contains passages read with various degrees of expressivity, we split the data into two subsets: one pertaining mostly to the narrator's intonation pattern (**MARA-Flat**), and the other associated mostly with the dialogue lines and character voices acted by the reader (**MARA-Expr**). MARA-Flat includes the utterances which have an F_0 mean value within 100Hz around the corpus-level F_0 mean, and an F_0 standard deviation of less than 50Hz. The rest of the utterances make up the MARA-Expr subset. This division is also balanced duration-wise: MARA-Flat contains 5 hours and 44 minutes of the total 11 hours of speech data. Figure 1 shows the F_0 statistics in letter-value plot format [18] for the MARA-Flat and MARA-Expr subsets. It can be observed that the MARA-Expr subset has a much wider domain for the F_0 values.

The segmented corpus along with the annotations is available from http://speech.utcluj.ro/marasc/ and utilizes a CC-by-NC-ND 4.0 licence.[4]

III. Synthesised expressive speech data

In end-to-end synthesis systems, the individual control of F_0 or segment duration is not straightforward. One way to manipulate the prosodic features is by using the appropriate training data. However, it is usually the case that for a particular voice identity the available data does not contain the desired level of expressivity or speaking style. To overcome this issue, most approaches tend to use one-shot inference strategies. However, this is not feasible as it generally requires a similar length utterance to be available as reference.

[3]https://wiki.inf.ed.ac.uk/twiki/pub/CSTR/F0parametrisation/hts_lab_format.pdf

[4]https://creativecommons.org/licenses/by-nc-nd/4.0/

The idea behind this study is to overcome the lack of appropriate training data by utilising synthesised speech. The synthesised data copies the F_0 contours and phone duration of an external speech resource, but uses the spectral characteristics of the original speaker. Because the spectrum is synthesised, it contains artefacts correlated with the synthesis method it uses. To average out these artefacts we mix the synthetic data with natural samples in the training stage.

TTS systems in which the control of each individual component of the speech is easily accessible are the statistical-parametric ones. To limit the influence of system specific artefacts on the results, we chose two separate training methodologies and their most representative implementations. The first methodology and the one which was ubiquitous until recent years is the one based on Hidden Markov Models, and we select the HTS implementation [19]. The second methodology uses deep learning networks with feed forward layers, but parameterises the waveform and extracts linguistic information from the text. One of its representative implementations is the Merlin toolkit [20].[5]

To mimic the lack of expressive data, the two systems were trained using only the MARA-Flat subset which amounts to approximately 5 hours of speech recordings segmented in around 4000 utterances. The text was processed by our internal front-end module to obtain the required text labels for both HTS and Merlin systems. The speech waveform was downsampled to 16kHz and parameterised by the WORLD vocoder [21] into Mel-generalised cepstral coefficients, F_0 and band aperiodicity coefficients. Phone-level alignments were obtained using an iterative HMM-based aligner.

As an additional setup, because the HTS quality is rather poor, we also analysed the use of a neural network-based post-filter to enhance the system's output [22]. This means that we synthesised the entire flat subset using the HTS system and paired the output utterances with their natural counterparts in a voice conversion-like setup. The voice conversion neural network would presumably learn to correct the errors made by the HTS system and drive the output more towards the natural data.

After all three systems were trained, we used the phone durations and F_0 contours extracted from the MARA-Expr subset, combined them with the spectral parameters generated by TTS systems for the same utterance and generated the synthetic waveforms. A schematic overview of this step is shown in Figure 2.

One problem noticed while synthesising the MARA-Expr subset with the natural F_0 contours is the fact that the speaker frequently utters breathy vowels, i.e. without a measurable F_0 value. But during the synthesis stage, the systems generate spectral coefficients corresponding to voiced phones. The result of this combination are creaky voice artefacts which propagate to the next stage of our experiments, as well, on top of the system specific artefacts (such as buzziness in HTS).

[5]https://github.com/CSTR-Edinburgh/merlin

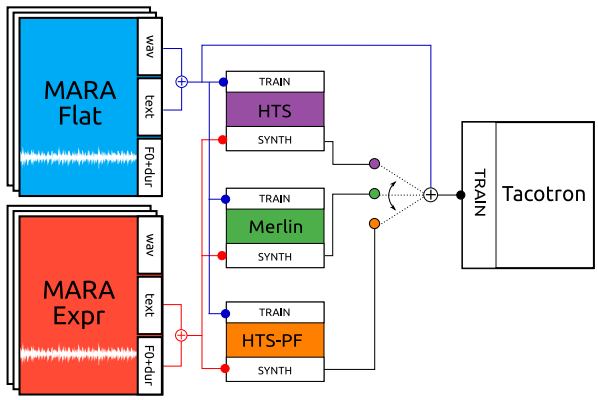

Fig. 2. Block diagram of the end-to-end systems' training process using synthesised expressive speech data

TABLE I
END-TO-END SYNTHESIS SYSTEMS' DESCRIPTION

No.	System id	Expressive data
1	**Tac:FLAT**	None
2	**Tac:ALL**	Natural data
3	**Tac:Merlin**	Merlin synthesised data
4	**Tac:HTS**	HTS synthesised data
5	**Tac:HTS-PF**	HTS with post-filter synthesised data

Correcting these errors is possible, but we did not address it at this point.

IV. END-TO-END SYNTHESIS SYSTEMS

The previous section described the process of obtaining synthesised data which encompasses the natural F_0 and phone duration with the help of statistical-parametric speech synthesisers. However, this evaluation is aimed at the end-to-end speech synthesis systems. The chosen architecture for the end-to-end TTS system is that of Tacotron [14] in the Mozilla implementation.[6]

To alleviate the influence of the synthesis artefacts, the neural network was pretrained with the MARA-Flat subset for 500 epochs. The learned weights were subsequently used to initialise all other systems. The training continued for another 500 epochs, resulting in a total of 1000 epochs. No early stopping criterion was employed because of the mismatch between the network's cost function and perceptual audio quality. In the second part of the training the synthesised expressive data was mixed with the MARA-Flat subset, as an additional guard for the low quality of the initial synthesis systems.

With this setup we trained 5 systems. The baseline in our expressivity experiments is the one which uses just the flat audio subset (**Tac:FLAT**). This system was trained for 1000 epochs as well, but using the same training set for the second stage. The topline is the system trained on the complete MARA corpus (**Tac:ALL**) including the natural expressive utterances. The systems which we actually evaluate are the ones based on the HTS (**Tac:HTS**), Merlin (**Tac:Merlin**), and

[6]https://github.com/mozilla/TTS

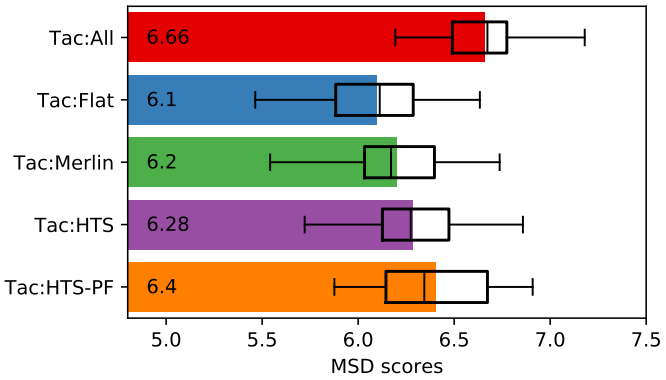

Fig. 3. MSD scores across 50 testing samples. The horizontal bars represent the mean MSD values with boxplots overlapped.

HTS plus post-filtering (**Tac:HTS-PF**) synthesised expressive data. Table I summarises the systems' ids and training datasets.

V. Evaluation

A. Objective measures

Expressivity alone in a TTS system is not sufficient, because a dynamic prosody pattern does not always translate into natural samples. So that it is important to also measure the naturalness of the synthesis. One common approach to do it is through the Mel Spectrogram Distortion (MSD) measure [23]:

$$MSD(s^t, s^r) = \frac{10\sqrt{2}}{ln10}\frac{1}{T}\sum_{t=0}^{T=1}\sqrt{\sum_{d=1}^{D}(s_d^t(t) - s_d^r(t))^2} \quad (1)$$

where s^t and s^r are the target and reference Mel spectrogram vectors, respectively; T is the total number of frames, and D is the number of Mel bins. The 0^{th} coefficient (the energy) is discarded. To align the synthesised and natural sequences, a Dynamic Time Warping (DTW) algorithm was used. Figure 3 shows the MSD scores for 50 samples not present in the training dataset of any of the systems, but which are part of the MARA-Expr subset. It can be noticed that the **Tac:All** system has a relatively higher MSD mean. This can be caused by the fact that the prosodic variation of this system is higher than the rest, and it is a result of using all the natural samples available in MARA. This means that the alignment between the natural reference and the synthetic sample is further apart. This claim is also supported by the fact that the lowest MSD score is exhibited by the **Tac:Flat** system, the one which uses only the flat prosodic utterances. However, with the exception of **Tac:All**, the rest of the systems have similar MSD mean scores, concluding that the addition of synthesised data to the training set did not affect the overall quality of the output voice.

In terms of expressivity, a method to measure it is by analysing the F_0 statistics. Figure 4 plots these statistics for the 50 test samples used in the MSD computation. It can be observed that the largest F_0 variation is contained in the

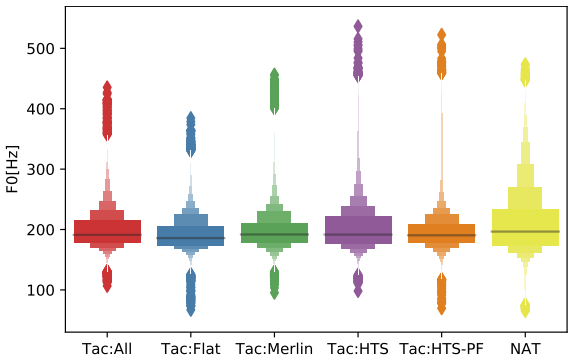

Fig. 4. F_0 statistics across 50 testing samples for all end-to-end systems.

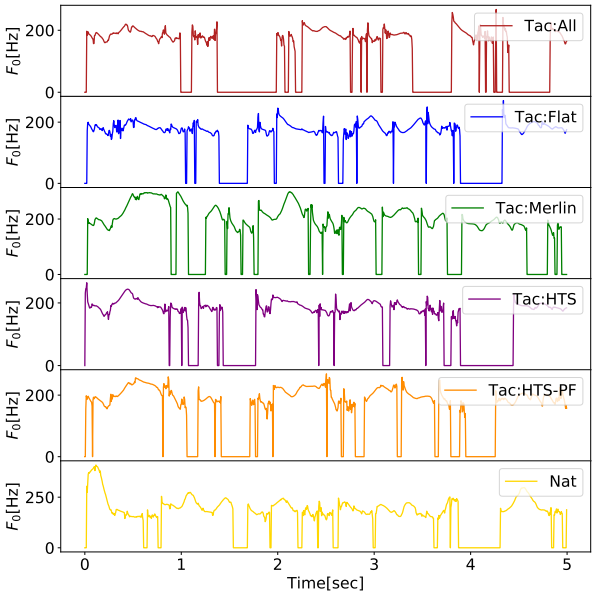

Fig. 5. A sample of an utterance's F_0 contour generated by all the end-to-end systems.

natural samples. The **Tac:HTS** system is next, followed by **Tac:All**. A sample of one utterance's F_0 contour generated by all systems is depicted in Figure 5.

B. Listening test

Even though the objective measures can indicate certain aspects regarding the quality of the synthesised speech, there is no objective measure which truly correlates with the perceptual evaluation. We conducted a listening test based on the MUlti Stimulus test with Hidden Reference and Anchor (MuSHRA) methodology.[7] The test included two sections: *naturalness* and *expressivity*. In the naturalness section, the natural sample was presented to the listeners as reference. In the expressivity section, we did not want to influence the judgement of the listener, so that the natural sample was not clearly marked as reference, but was listed among the evaluated systems. In both sections, the lower anchor was set to a sample generated

[7]ITU-R Recommendation BS.1534-1

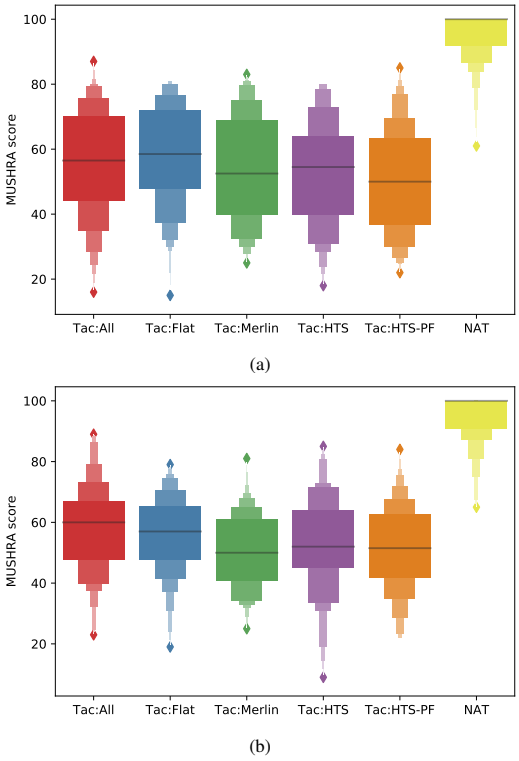

Fig. 6. Letter-value plot of MuSHRA scores for the (a) **naturalness** and (b) **expressivity** section.

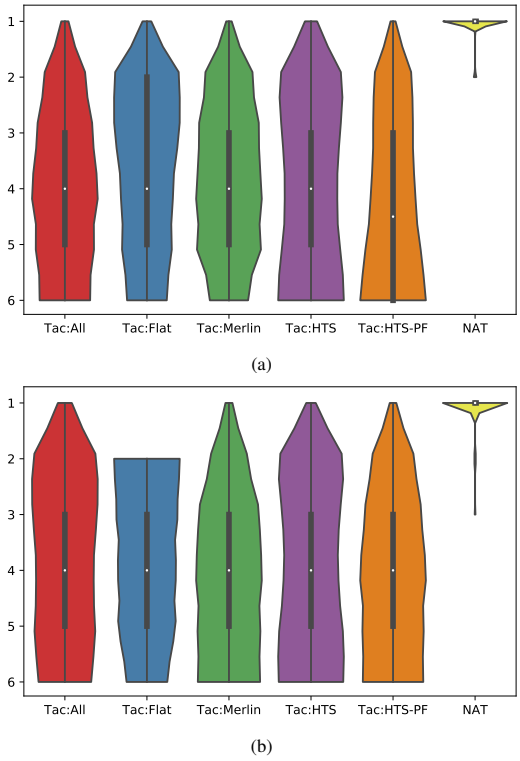

Fig. 7. Violin plot of MuSHRA rankings in the (a) **naturalness** and (b) **expressivity** sections.

by the original HTS system. 81 listeners with no self-declared hearing problems were asked to mark the perceived naturalness and expressivity of the speech samples on a scale of [0 to 100]. For each sample, 7 stimuli are presented to the listener side-by-side on the same screen, representing the 5 evaluated systems plus the natural and original HTS samples. Each listener rated 10 screens and could playback the samples as many times as they wished. The average length of the utterances is 15 seconds. Audio samples from the listening test are available here: http://speech.utcluj.ro/sped2021_mara/. Out of the 81 listeners, 15 were disqualified due having rated the reference sample below 90 MuSHRA points for more than 15% of all samples.

Letter-value plots of the naturalness and expressivity listening test results are shown in Figure 6. In the naturalness section, the best rated synthesis system was **Tac:Flat**. The fact that **Tac:Flat** is evaluated higher in naturalness can be a result of the network not having to accommodate such large F_0 variations in the data, and thus being able to better replicate the spectral characteristics of the speech. The worst performing system in naturalness evaluation was **Tac:HTS-PF**. This was expected because one side-effect of the post-filter is a metallic sounding voice quality which propagates all the way through to the end-to-end system. It is interesting to notice that the **Tac:Merlin** system is rated lower than **Tac:HTS** in both sections of the listening test, although the original systems were rated the other way around [20]. It might be the case that the accuracy of the phone-level alignment used in the training process is below the threshold at which the feed-forward network can still compensate for its effects.

In the expressivity rating, the best performance was obtained by **Tac:All**. The lowest rated system was in this case **Tac:Merlin**. This might be partially caused by the poor evaluation of this system in the naturalness section. Our listeners are not speech experts, are therefore unable to abstract the expressivity without taking into account the naturalness of the utterance.

Aside from the absolute MuSHRA scores, we also analysed the inter-system ranking performed by the listeners. Figure 7 shows the violin plots of these rankings. A lower value represents a better system. These rankings are in correspondence with conclusions and result of the absolute MuSHRA evaluation.

All of the above interpretations are based on minor differences between the systems' ratings. However, no statistically significant differences were found between these ratings. As a general conclusion of the listening test it seems that by substituting the natural samples with synthesised copies of them in the training data of an end-to-end TTS system, the neural network is able to average out the spectral artefacts of these samples. As a result, the naturalness and expressivity of the resulting voice is only marginally affected.

VI. Conclusions

This paper described a newly developed Romanian expressive speech corpus entitled MARA. It contains over 11 hours of high-quality data recorded by a professional female speaker. The data is manually segmented at short phrase level and automatically aligned at phone level. The associated text is also fully annotated by an accurate text-processing platform to include: text normalisation, phonetic transcription, syllabification, lexical stress assignment, lemma extraction, part-of-speech tagging, chunking and dependency parsing. The MARA corpus is freely available for non-commercial applications.

Starting from this new resource and based on the fact that in end-to-end speech synthesis systems controlling the prosodic parameters (e.g. F_0 and segment duration) is not straightforward, we devised a method to modify the expressivity of the system by using expressive synthesised speech as training data. Five systems were evaluated with objective and subjective measures. The results showed that using synthesised data does not affect the naturalness of the output voice, and that there are no statistically significant differences between the systems' expressivity or naturalness levels. This leaves room for improvements and potentially a faster adaptation mechanism towards a target speaker with the use of synthesised data.

As future work, we would like to investigate other means of generating higher quality synthesised data. We would also like to test the inter-gender prosody transfer, as this might pose problems in the translation of the absolute F_0 values.

References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," *CoRR*, vol. abs/1712.05884, 2017. [Online]. Available: http://arxiv.org/abs/1712.05884

[2] H. Zen, R. Clark, R. J. Weiss, V. Dang, Y. Jia, Y. Wu, Y. Zhang, and Z. Chen, "Libritts: A corpus derived from librispeech for text-to-speech," in *Proceedings of Interspeech*, 2019.

[3] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorihms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Transactions on Audio, Speech and Language Processing*, 2008.

[4] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for dnn-based speech synthesis," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 879–883. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2015/i15_0879.html

[5] J. Lorenzo-Trueba, R. Barra-Chicote, R. San-Segundo, J. Ferreiros, J. Yamagishi, and J. M. Montero, "Emotion transplantation through adaptation in HMM-based speech synthesis," *Computer Speech and Language*, vol. 34, no. 1, pp. 292 – 307, 2015.

[6] 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.

[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] G. E. Henter, X. Wang, and J. Yamagishi, "Deep encoder-decoder models for unsupervised learning of controllable speech synthesis," *ArXiv*, vol. abs/1807.11470, 2018.

[10] Y. Zhang, S. Pan, L. He, and Z. 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)*, May 2019, pp. 6945–6949.

[11] W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen, and R. Pang, "Hierarchical Generative Modeling for Controllable Speech Synthesis," in *arXiv*, 2018. [Online]. Available: https://arxiv.org/abs/1810.07217

[12] J. Parker, Y. Stylianou, and R. Cipolla, "Adaptation of an expressive single speaker deep neural network speech synthesis system," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5309–5313.

[13] N. Prateek, M. Lajszczak, R. Barra-Chicote, T. Drugman, J. Lorenzo-Trueba, T. Merritt, S. Ronanki, and T. Wood, "In other news: a bi-style text-to-speech model for synthesizing newscaster voice with limited data," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 2 (Industry Papers)*, 2019, pp. 205–213. [Online]. Available: https://www.aclweb.org/anthology/N19-2026/

[14] 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 *Proc. Interspeech*, 2017. [Online]. Available: https://arxiv.org/abs/1703.10135

[15] A. Stan, J. Yamagishi, S. King, and M. Aylett, "The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate," *Speech Communication*, vol. 53, no. 3, pp. 442–450, 2011.

[16] A. Stan, F. Dinescu, C. Tiple, S. Meza, B. Orza, M. Chirila, and M. Giurgiu, "The SWARA Speech Corpus: A Large Parallel Romanian Read Speech Dataset," in *Proceedings of the 9th Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Bucharest, Romania, July, 6-9 2017. [Online]. Available: http://adrianastan.com/papers/2017_SPED_Swara.pdf

[17] A. Stan, Y. Mamiya, J. Yamagishi, P. Bell, O. Watts, R. Clark, and S. King, "ALISA: An automatic lightly supervised speech segmentation and alignment tool," *Computer Speech and Language*, vol. 35, pp. 116–133, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230815000650

[18] H. Heike, H. Wickham, and K. Kafadar, "Letter-value plots: Boxplots for large data," *Journal of Computational and Graphical Statistics*, vol. 26, 03 2017.

[19] K. Tokuda, H. Zen, and A. Black, "An HMM-Based Speech Synthesis System Applied To English," in *Proc. of SSW*, 10 2002, pp. 227 – 230.

[20] Z. Wu, O. Watts, and S. King, "Merlin: An Open Source Neural Network Speech Synthesis System," in *9th ISCA Speech Synthesis Workshop (2016)*, Sep. 2016, pp. 218–223.

[21] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," *IEICE Transactions*, vol. 99-D, pp. 1877–1884, 2016.

[22] M. G. Öztürk, O. Ulusoy, and C. Demiroglu, "Dnn-based speaker-adaptive postfiltering with limited adaptation data for statistical speech synthesis systems," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 7030–7034.

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