An analysis of the data efficiency in Tacotron2 speech synthesis system

Georgiana SĂRACU, Adriana STAN
Communications Department
Technical University of Cluj-Napoca, Romania
georgiana.saracu@gmail.com, adriana.stan@com.utcluj.ro

***Abstract*—This paper introduces an evaluation of the amount of data required by the Tacotron2 speech synthesis model in order to achieve a good quality output synthesis. We evaluate the capabilities of the model to adapt to new speakers in very limited data scenarios. We use three Romanian speakers for which we gathered at most 5 minutes of speech, and use this data to fine tune a large pre-trained model over a few training epochs. We look at the performance of the system by evaluating the intelligibility, naturalness and speaker similarity measures, as well as performing an analysis of the trade-off between speech quality and overfitting of the network.**

The results show that the Tacotron2 network can replicate the identity of a speaker from as little as one speech sample. Also it inherently learns individual grapheme representations, such that if the training data is carefully selected to present all the common graphemes in the language, the adaptation data requirements can be significantly lowered.

***Index Terms*—speech synthesis, text-to-speech, speaker adaptation, Tacotron, Romanian, deep learning, neural models.**

I. Introduction

The quality of the text-to-speech (TTS) synthesis systems has recently advanced to a point where there is very little difference between natural and synthetic speech samples [1], [2]. This quality comes at the price having large, purposely-designed, single speaker speech recordings. This data requirement is more than often a large impediment for the expansion of language and speaker coverage in TTS systems.

When this data is not available, there are several methods which aim to alleviate this requirement. The most common of them is that of pre-training a large multi-speaker model [3], also called an `eigen voice` model, and then fine-tune it towards a small amount of a target speaker's exemplars. Other approaches refer to using speaker embeddings obtained from an external representation network and use them to condition the synthesis process [4]. Both approaches have their advantages and disadvantages, and both still rely on large amounts of high-quality speech data from multiple speakers. Yet the amount of data from each speaker is significantly reduced, and could in theory be obtained from freely available resources, such as audio podcasts, interviews, etc.

Two other important aspects of all the recently developed deep neural TTS architectures include their complexity and adaptability. The complexity would refer to the data requirements, i.e. quality and amount of data, while the adaptability would refer to their potential of factoring the speech and/or speaker characteristics within their modules. The low complexity and fast adaptability would enable a DNN TTS network to use as little data as possible from a new speaker and render high-quality synthetic speech within just a few fine tuning epochs.

In this paper we address this idea of fast adaptation within a well-known DNN TTS architecture, namely Tacotron 2 [1]. We pretrain the network with a large multi-speaker speech corpus in an attempt to obtain an eigen voice model. We then fine-tune this model with as little as a single utterance from each speaker and analyse the synthesis' objective and subjective quality, as well as its overfitting characteristics. We also look into the linguistic content of the adaptation samples and analyse if the network does indeed learn some phonetic representation of the data, which would then limit the quality of the missing phones from the output speech.

The paper is organised as follows: Section II describes the flow of our experiments, which is then evaluated and discussed in Section III. Conclusions are drawn in Section IV.

II. Methodology

A. Tacotron2 Speech Synthesis System

Tacotron2 is an end-to-end speech synthesis system capable of generating natural human-like speech given text input [1]. The system includes two components: (1) a spectrogram prediction network and (2) a neural vocoder that produces waveforms conditioned by the results of the first.

For the spectrogram prediction, Tacotron2 relies on a Mel frequency representation [5] which reduces the number of data points which need to be predicted. The Mel frequency spectrogram is a short-term power spectrum representation based on a non-linear transformation of the frequency axis. Although the Mel spectrogram does not include the phase characteristics, the phonemic aspects of the speech are invariant to the phase, and this information can be reconstructed using deterministic algorithms, such as Griffin-Lim [6], or neural-based algorithms, such as WaveGlow [7] or WaveNet [8]. The Mel scale also has the advantage of better representing the lower frequency spectrum which is more critical to speech intelligibility, as opposed to the higher frequency spectrum which pertains

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

mostly to speaker characteristics. The architecture of the spectrogram predictor relies on an encoder-decoder module with attention.

The text input sequence is passed through a stack of 3 convolutional layers, followed by batch normalization [9] and a ReLU activation layer. The hidden representation of the input sequence, is then fed into a bidirectional LSTM [10] layer and passed through the location-sensitive attention module [11]. The decoder uses the previous predictions passed through a bottleneck layer and the attention context vector to produce the current spectrogram frame. A post-net produces a residual which is added to the current frame in order to increase the prediction's quality.

Although it obtained one of the highest subjective evaluations of a TTS system at its time, Tacotron2 has the major drawback of having a very slow inference process due to its recurrent structure. Recent studies focus on this issue and aim to perform the duration prediction of each phoneme as an individual module in the network [12], [13].

The result of the encoder-decoder architecture is just a Mel spectrogram which then needs to be converted into the final waveform. This step is commonly referred to as the vocoder step. There are numerous solutions for it, and some of the best results were obtained by WaveNet [8] and WaveGlow [7]. These, too, have the disadvantage of including recurrent units, and studies like Parallel WaveNet [14] and Clarinet [15] remove the auto-regression from the neural architecture while achieving similar qualitative results. In this work we relied on the WaveGlow vocoder to obtain the synthesised output. WaveGlow uses the normalising flows principle, which was first applied in image generation [16]. The normalising flows have the advantage of using invertible structures which can efficiently compute the distribution of the training data by projecting it into a simple known distribution, such as a spherical Gaussian.

Tacotron2 and WaveGlow are the choices for our study thanks mainly to their well tested structures and implementations, as well as their self- and peer-reported quality results in text-to-speech synthesis. The implementations we used are the ones made available by NVIDIA.[1],[2]

III. Results

A. Data

We evaluate the Tacotron2 architecture on a dataset that includes samples of audio-text pairs from 3 Romanian speakers: 2 male (**M1** and **M2**) and one female (**F1**). For each speaker, we used at most 5 minutes of speech. The data was collected from the speakers' YouTube channels, and the audio quality is somewhat lower compared to standard text-to-speech synthesis datasets, such as [17].

The audio samples' length varies from 4 to 11 seconds, and 100 ms of silence was maintained at the beginning and end of each waveform. The data was resampled at 22050 Hz so that it corresponds to the eigen voice model's sampling rate. The associated text was manually checked and normalised.

B. Eigen voice model

The eigen voice model was trained on the data from the SWARA 2.0 corpus, which is an extension of the SWARA corpus [18]. SWARA 2.0 includes an additional 29 speakers recorded in their home environments, uttering the same text as the original speakers, making it one of the largest parallel multi-speaker Romanian speech corpora. The model was not conditioned on the speaker identity and used graphemes as text representation. The training was performed over 1500 epochs at a batch size of 8.

When synthesising from the eigen voice model there is no definite or consistent voice identity. This means that, depending on the input text, the output voice is one of the speakers' identity, or it can skip from one identity to another within the utterance. This result shows that the Tacotron2 architecture does not average the speaker identities, but rather learns some dependencies or paths in the network pertaining to the different speakers. An ideal eigen voice model would in fact average the speaker characteristics, and would make it more easily adaptable to new speakers. But this question remains to be studied in our future work.

C. Model finetuning

Starting from the eigen voice model, we then used the new target speakers' data to finetune the model's weights for each speaker independently. We looked into 3 different dataset sizes:

- **S1** - a random single utterance (a few seconds);
- **S2** - 3-4 utterances selected such that they include all the common Romanian graphemes (around 12 seconds per speaker);
- **S3** - 35-37 utterances (at most 5 minutes of speech per speaker).

Due to the limited data scenario, we also looked into the trade-off between synthetic speech quality and model overfitting during the training process. For each speaker, and for each dataset size we empirically determined 3 checkpoints which would roughly correspond to: the earliest epoch with intelligible results; the best quality epoch; and the epoch at which the model starts to overfit the data. The selected epochs for each speaker and dataset size are summarised in Table I. We will refer to each of the three checkpoints as `C1`, `C2`, and `C3`, respectively.

D. Objective evaluation

For an initial evaluation of the Tacotron2 TTS system's data efficiency capabilities we rely on a ASR-based assessment. It has been recently shown that the word error rate (WER) results of a good ASR system can accurately evaluate the intelligibility of the synthesised speech [19]. We, therefore, engage a high-quality Romanian ASR system [20] which obtained a WER over our natural samples of 7.43%. The synthetic WER was computed over 10 samples from each checkpoint,

[1]https://github.com/NVIDIA/tacotron2

[2]https://github.com/NVIDIA/waveglow

Speaker	Dataset	No. of utts.	Checkpoint
M1	S1	1	C1: 60 C2: 80 C3: 100
	S2	4	C1: 100 C2: 200 C3: 300
	S3	35	C1: 150 C2: 200 C3: 500
M2	S1	1	C1: 100 C2: 150 C3: 500
	S2	3	C1: 100 C2: 200 C3: 300
	S3	37	C1: 150 C2: 200 C3: 500
F1	S1	1	C1: 60 C2: 80 C3: 100
	S2	4	C1: 100 C2: 200 C3: 300
	S3	35	C1: 400 C2: 500 C3: 700

TABLE I: Checkpoints (epochs) chosen for each of the three speakers and dataset sizes.

dataset and speaker. The results are shown in Figure 1. The subplots correspond to the three speakers: M1, M2 and F1; the hatch patterns correspond to the different dataset sizes: S1, S2, and S3; while colours pertain to the selected evaluation checkpoints. Looking the the results, it can be noticed that speaker M2 obtains the lower WERs overall, while speaker F1 is the worst performing one. One explanation for this could be that, although SWARA 2.0 contains more female voices (22) compared to male voices (15), the eigen voice model exhibits male speaking characteristics. This means that the adaptation towards a female voice would require a bit more data.

In terms of the different dataset sizes, as the number of samples increases, the WER lowers–this was to be expected. It is interesting to notice, though, that the difference between S2 and S3 is not that big. This means that the careful selection of adaptation data, in this case selecting the minimum number of utterances which contain all common graphemes in Romanian, can be more informative to the model than a large number of adaptation utterances.

With respect to the three checkpoints, as they were empirically selected, the assumption that C2 would perform better across all systems and speakers does not hold true. However, this is true when using only one sample (i.e. S1), as the model does not have enough variability in the training data, and the overfitting process is more likely to appear. Another thing to notice is that for speaker M2, the differences between using 1 sample (S1) and using 35 samples (S3) for adaptation, are not that big. It might be that speaker M2 has speaking characteristics which are closer to the ones of the

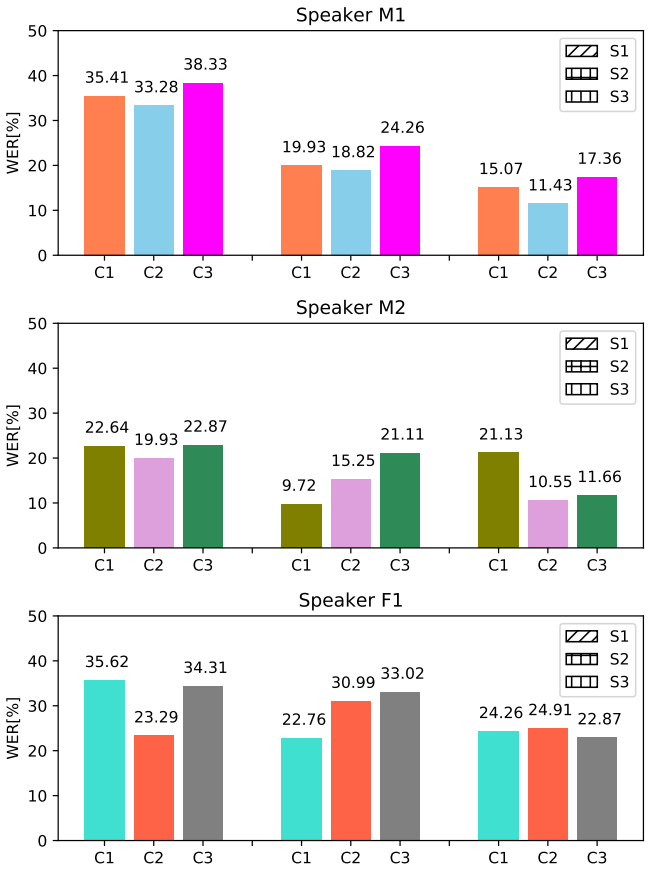

Fig. 1: WERs for all speakers and TTS systems. Each subplot corresponds to a speaker, hatching patterns pertain to the different datasets, and colours correspond to the three checkpoints.

speakers within the SWARA 2.0 dataset, making it more easily adaptable within the network's weights.

E. Subjective evaluation

If the WER can give a good indication of the synthetic speech's intelligibility, there are no well established measures for naturalness and speaker similarity. Therefore, we perform a subjective listening test using the MUlti Stimulus test with Hidden Reference and Anchor (MuSHRA) method[3]. Samples from all speakers, datasets and checkpoints were synthesised and compared between them, as well as to the corresponding natural sample. 15 listeners were asked to rate the samples on a scale of 0 to 100, where 100 corresponded to either the most natural sample, or the closest sample to the natural reference in terms of speaker identity. The results of the listening test are shown in Figure 2. The left column shows the naturalness scores for the three speakers and each dataset and system, while the right column shows the results for speaker similarity. As in the case of the WER, the systems using more data points achieve higher scores in the subjective evaluation (i.e. S3).

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

However, although speaker M2 had the lowest WER, it was not rated as high in naturalness and speaker similarity. The best performing speaker in this setup was M1. Although we would have expected C2 to have the best quality in terms of naturalness and speaker similarity compared to C1 and C3, this is not the case for all speakers and datasets. It might be the case that due to the very limited adaptation data, and the listeners inexperience with synthetic speech, the differences between the systems were not easily distinguishable.

Another interesting result observed from the initial listening of the synthesised samples is that if the network is not presented with any of the graphemes uttered by the target speaker, the output speech will not be intelligible for that particular grapheme. This means that the network inherently learns a representation of the graphemes, or phonemes if it is trained in this manner. This result could indicate that a careful selection of the training data can significantly reduce the number of utterances needed for the adaptation process. One other empirical observation we made was that the speech rhythm (or speed) of the target speaker is also important, meaning that faster speakers need less data duration wise. This is not surprising, but it can also raise the question of comparable corpora sizes when disregarding the prosodic characteristics of the speaker.

Furthermore, we analyzed the phonetic content of the data. We trained the model on different data scenarios. We intentionally exclude a few letters from the datasets with limited samples. We observed that, at inference, the model tends to replace the letter's pronunciation with either silence or a (median) trivial/frequent letter pronunciation. However, in most cases, the model manages to synthesize the correct utterance even though the letter is missing as a result of underlying knowledge of the pre-trained model.

To continue, the same situation as described above arose in the case of diphthongs. Diphthongs are a combination of two vowels, and they are classified as raising or falling ones according to the vowels and their position. If the data given to the trained model does not contain such pronunciation examples, the model encounters difficulties in associating whether the diphthong is rising or falling.

IV. Conclusions

In this paper we analysed the adaptation data efficiency of a well-established speech synthesis system, namely Tacotron2. The analysis focused on the ability of the architecture to extract meaningful information from very little data, as little as a single utterance. The results showed that Tacotron2's architecture can easily accommodate such training scenarios, and that if the adaptation samples is carefully chosen, the data requirements can be lowered. It is also important to notice that in the little data setup, there is an operation point at which the trade-off between speech quality and network overfitting is optimal. Although it is hard to exactly determine that operation point, this aspect should be taken into consideration.

As future work we would like to extend the results to multiple speakers, other languages, and other TTS architectures.

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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[2] R. Valle, K. J. Shih, R. Prenger, and B. Catanzaro, "Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis," *CoRR*, vol. abs/2005.05957, 2020. [Online]. Available: https://arxiv.org/abs/2005.05957

[3] I. Himawan, S. Aryal, I. Ouyang, S. Kang, P. Lanchantin, and S. King, "Speaker adaptation of a multilingual acoustic model for cross-language synthesis," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7629–7633.

[4] E. Cooper, C.-I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6184–6188.

[5] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 357–366, 12 1980.

[6] D. Griffin and Jae Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[7] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," *CoRR*, vol. abs/1811.00002, 2018. [Online]. Available: http://arxiv.org/abs/1811.00002

[8] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[9] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[10] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[11] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," vol. 28, 2015.

[12] I. Elias, H. Zen, J. Shen, Y. Zhang, J. Ye, R. J. Skerry-Ryan, and Y. Wu, "Parallel tacotron 2: A non-autoregressive neural TTS model with differentiable duration modeling," *CoRR*, vol. abs/2103.14574, 2021. [Online]. Available: https://arxiv.org/abs/2103.14574

[13] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," 2021.

[14] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. C. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis, "Parallel wavenet: Fast high-fidelity speech synthesis," 2017.

[15] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," *CoRR*, vol. abs/1807.07281, 2018. [Online]. Available: http://arxiv.org/abs/1807.07281

[16] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," 2018.

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

[18] A. Stan, F. Dinescu, C. Ţiple, Ş. Meza, B. Orza, M. Chirilă, and M. Giurgiu, "The SWARA speech corpus: A large parallel Romanian read speech dataset," in *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, 2017, pp. 1–6.

[19] J. Taylor and K. Richmond, "Confidence Intervals for ASR-based TTS Evaluation," in *Proceedings of Interspeech*, 2021.

[20] A.-L. Georgescu, H. Cucu, and C. Burileanu, "Kaldi-based DNN architectures for speech recognition in Romanian," in *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, 2019, pp. 1–6.

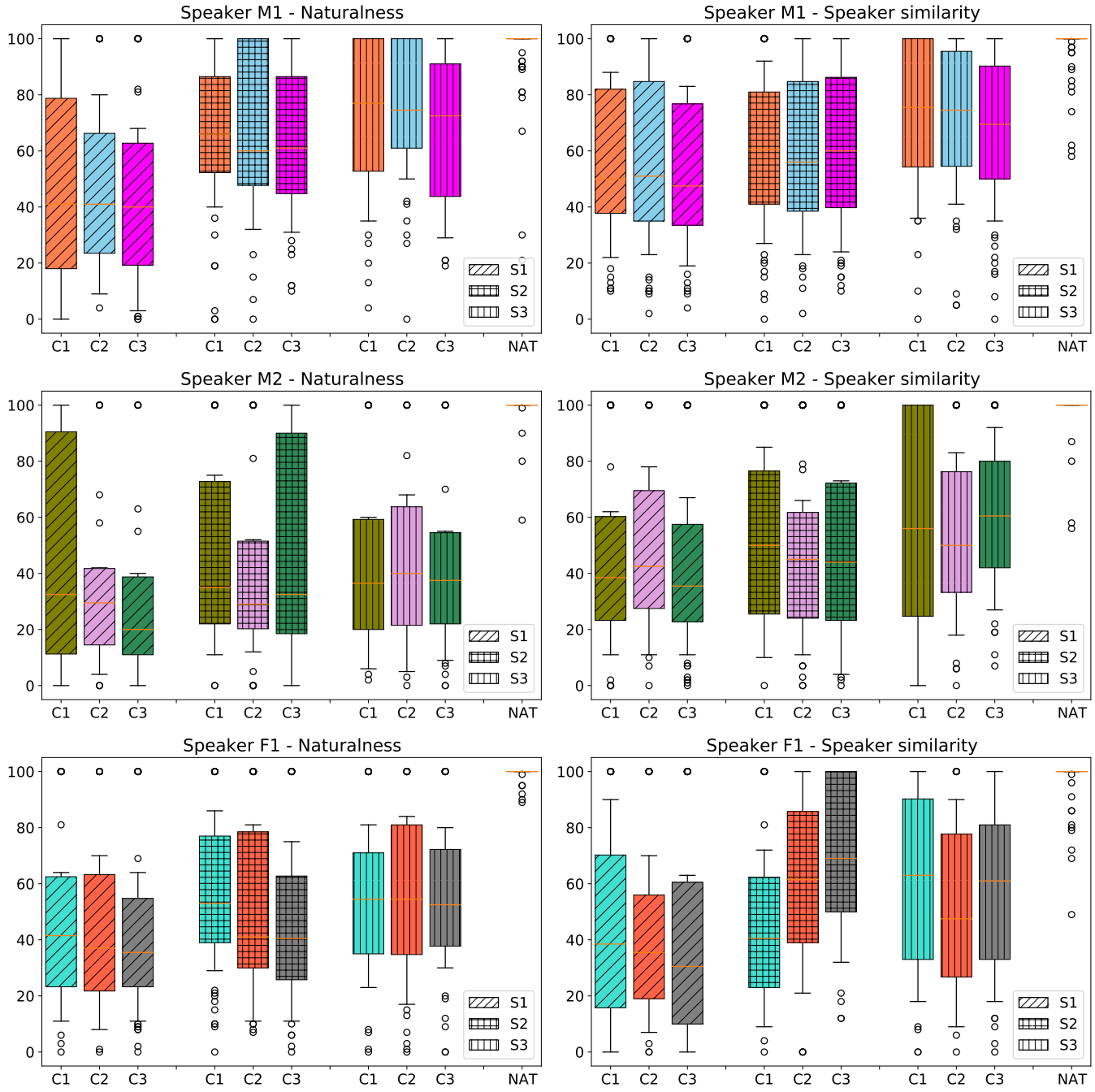

Fig. 2: Listening test results for all speakers, dataset and checkpoints. Each subplot corresponds to a speaker and a listening test category (naturalness or speaker similarity), hatching patterns pertain to the different datasets, and colours correspond to the three checkpoints.