ITAcotron 2: The Power of Transfer Learning in Expressive TTS Synthesis

Anna Favaro, Licia Sbattella, Roberto Tedesco, and Vincenzo Scotti

Abstract A text-to-speech (TTS) synthesiser has to generate intelligible and natural speech while modelling linguistic and paralinguistic components characterising human voice. In this work, we present ITAcotron 2, an Italian TTS synthesiser able to generate speech in several voices. In its development, we explored the power of transfer learning by iteratively fine-tuning an English Tacotron 2 spectrogram predictor on different Italian data sets. Moreover, we introduced a conditioning strategy to enable ITAcotron 2 to generate new speech in the voice of a variety of speakers. To do so, we examined the zero-shot behaviour of a speaker encoder architecture, previously trained to accomplish a speaker verification task with English speakers, to represent Italian speakers' voiceprints. We asked 70 volunteers to evaluate intelligibility, naturalness, and similarity between synthesised voices and real speech from target speakers. Our model achieved a MOS score of 4.15 in intelligibility, 3.32 in naturalness, and 3.45 in speaker similarity. These results showed the successful adaptation of the refined system to the new language and its ability to synthesise novel speech in the voice of several speakers.

1 Introduction

The development of text-to-speech (TTS) synthesis systems is one of the oldest problems in the natural language processing (NLP) area and has a wide variety of applications [14]. Such systems are designed to output the waveform of a voice

A. Favaro (✉)
Center for Language and Speech Processing (CLSP), Johns Hopkins University, Baltimore, MD, USA
e-mail: afavaro1@jhu.edu

L. Sbattella · R. Tedesco · V. Scotti
Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano, Milano, MI, Italy
e-mail: licia.sbattella@polimi.it; roberto.tedesco@polimi.it; vincenzo.scotti@polimi.it

M. Abbas (ed.), *Analysis and Application of Natural Language and Speech Processing*, Signals and Communication Technology,
https://doi.org/10.1007/978-3-031-11035-1_1

uttering the input text string. In the last years, the introduction of approaches based on deep learning (DL), and in particular the end-to-end ones [11, 20, 22, 25], led to significant improvements.

Most of the evaluations carried out on these models are performed on languages with many available resources, like English. Thereby, it is hard to tell how good these models are and whether they are general across languages. With this work, we propose to study how these models behave with less-resourced languages, leveraging the transfer learning approach.

In particular, we evaluated the effectiveness of transfer learning on a TTS architecture, experimenting with the English and Italian languages. Thus, we started from the English TTS Tacotron 2 and fine-tuned its training on a collection of Italian corpora. Then, we extended the resulting model, with speaker conditioning; the result was an Italian TTS we named ITAcotron 2.

ITAcotron 2 was evaluated, through human assessment, on intelligibility and naturalness of the synthesised audio clips, as well as on speaker similarity between target and different voices. In the end, we obtained reasonably good results, in line with those of the original model.

The rest of this paper is divided into the following sections: In Sect. 2, we introduce the problem. In Sect. 3, we present some available solutions. In Sect. 4, we detail the aim of the paper and the experimental hypotheses we assumed. In Sect. 5, we present the corpora employed to train and test our model. In Sect. 6, we explain the structure of the synthesis pipeline we are proposing and how we adapted it to Italian from English. In Sect. 7, we describe the experimental approach we followed to assess the model quality. In Sect. 8, we comment on the results of our model. Finally, in Sect. 9, we sum up our work and suggest possible future extensions.

2 Background

Every TTS synthesiser represents an original imitation of the human reading capability, and, to be implemented, it has to cope with the technological and imaginative constraints characterising the period of its creation.

In the mid-1980s, the concomitant developments in NLP and digital signal processing (DSP) techniques broadened the applications of these systems. Their first employment was in screen reading systems for blind people, where a TTS was in charge of reading user interfaces and textual contents (e.g. websites, books, etc.), converting them into speech. Even though the early screen readers (e.g. JAWS[1]) sounded mechanical and robotic, they represented a valuable alternative for blind people to the usual braille reading.

Since the quality of TTS systems has been progressively enhanced, their adoption was later extended to other practical domains such as telecommunications services,

[1] https://www.freedomscientific.com/products/software/jaws.

language education systems, talking books and toys, and video games. In 2004, Yamaha Corporation released its first version of Vocaloid [16]. It is a voice synthesising software product, based on diphone concatenation that allows creating a virtual singer by specifying the text and the melody of a song.

In general, TTS can be conceived as a decoding problem where a highly "compressed" input sequence (text) has to be "decompressed" into audio. However, linguistic units (e.g. phonemes, characters, words) are discrete, whereas speech signals are continuous and longer than textual input sequences; this mismatch causes prediction errors to accumulate rapidly. Besides, the meaning expressed by an utterance is typically undermined by its textual counterpart since the same textual sequence can correspond to several pronunciations and speaking styles.

Nowadays, the hectic development of embodied agent technologies such as embodied conversational agents (ECA) that adopt mimics, gestures, and speech to communicate with the human user makes the modelling of human-computer dialogues a research hotspot. To endow an ECA with a human-like conversational behaviour, a TTS system cannot just synthesise understandable speech at a fast rate. Instead, it needs to account for further speech nuances in order to reproduce elements and peculiarities of human conversations [8].

Thus, to synthesise human-like speech, a TTS has to explicitly or implicitly model many factors that are not attested in the textual input. This requirement is invoked by the presence of paralinguistic components characterising human dialogic exchanges. On the whole, the synthesised speech should express the correct message (intelligibility) while sounding like human speech (naturalness) and conveying the right prosody (expressiveness) [31]. This is what makes the development of high-quality TTS systems a challenging task.

In our daily conversational exchanges, paralinguistic components (the so-called prosody) are exploited as a whole to attribute mental states and an independent mental life to our interlocutors. It follows that, if the ultimate goal is to develop an ECA that can successfully hold a conversation and be mistaken for a human, all of these components should be taken into account within the modelling pipeline [24].

Prosody is the systematic arrangement of various linguistic units into a single utterance or group of utterances, which occurs in the process of human speech production [7]. Its implementation encompasses both segmental and suprasegmental features of speech, and it aims at conveying linguistic and non-linguistic information.

Speech prosody mainly plays the roles of [28]:

- Disambiguating the verbal component of communication (i.e. augmentative prosody) [21]
- Conveying emotions, intentions, and attitudes in communication (i.e. affective prosody) [23]

The lack of explicit control on specific speech traits characterises architectures based on DL. This usually prevents them from reproducing accurately prosodic phenomena, both locally and globally. However, the end-to-end learning approach allows for the introduction of rich conditioning on various prosodic attributes. Thus, besides generating a comprehensible synthetic product, Seq-2-Seq models enable the synthesis of speech in multiple voices, in various styles, and with different emotional nuances. In the following, we present some modern approaches to TTS, which mostly inspired our work.

3 Related Work

Modern, DL-based TTS pipelines are composed of two main blocks: a *spectrogram predictor* and a *vocoder* [14]. These components take care of, respectively, converting a string of characters into a (mel-scaled) spectral representation of the voice signal and converting the spectral representation to an actual waveform. Optionally, input text—apart from normalisation—undergoes phonemisation to present the input to the spectrogram predictor as a sequence of *phonemes* rather than *graphemes*.

Recent end-to-end solutions for spectrogram prediction are built with an *encoder-decoder* architecture [20, 22, 25, 32]. The encoder maps the input sequence to a hidden continuous space, and the decoder takes care of generating, autoregressively, the spectrogram from the hidden representation. To produce the alignment between encoder and decoder, an *attention mechanism* [2] is introduced between these two blocks.

Among the available architectures for spectrogram prediction, *Tacotron* [32], and in particular its advanced version, Tacotron 2 [25], seems to be the most flexible and re-usable.

Many works have been developed to introduce conditioning into Tacotron, obtaining a fine-grained control over different prosodic aspects. The *Global Style Token* (GST) approach enabled control over the speaking style in an unsupervised manner [33]. Another controllable aspect is the speaker voice, introduced through additional *speaker embeddings* extracted through a speaker verification network [13]. Finally, [27] proposed a methodology to control the *prominence* and *boundaries* by automatically deriving prosodic tags to augment the input character sequence. It is also possible to combine multiple techniques into a single conditioned architecture, as shown by Skerry-Ryan et al. [26].

Neural vocoders completed the DL-based TTS pipeline improving consistently the quality of synthesised voice [15, 17, 30, 34]. These vocoders substituted the Griffin-Lim algorithm [9], which was characterised by artefacts and poor audio quality, especially if compared with newer neural approaches. These components, different from the spectrogram predictors, do not strictly depend on the input language—their primary role is to invert a spectral representation into the time domain; thus, they are thought to be *language-agnostic*.

As premised, the available models are primarily trained and evaluated on English corpora, due to data availability. A general solution for data scarcity is to leverage a technique called *transfer learning* [35] that mimics what typically occurs in human learning. In most learning environments, in fact, people do not start from scratch when forming hypotheses and drawing inferences. Rather, knowledge gained from one domain is abstracted and re-used in other domains. The more related a new task is to our previous experience, the easier we can master it.

The lack of sufficiently large data sets makes DL hard to apply and yield satisfying results. The mechanism of transfer learning [29] attempts to cope with this issue by re-using the hidden representations learnt performing a task on a domain similar to the target one. The idea is to leverage the knowledge derived solving one or more source tasks and use it to improve the results in a new target task. Techniques that enable knowledge transfer represent progress towards making machine learning as efficient as human learning.

For our work, we applied a variant of transfer learning called *fine-tuning*, where we used the pre-trained weights of the network as initialisation for the actual training on the new task [35].

4 Aim and Experimental Hypotheses

Whereas recent research mainly focuses on further refining the intelligibility and naturalness of the synthesised product, speech expressiveness requires for its part explicit modelling. The lack of a similar control makes the prosody of the synthesised speech feel like an “average, anonymous voice style”, preventing it from fully displaying the range of prosody variations occurring in human speech.

In this work, we address the modelling of speech expressiveness by presenting a customised TTS architecture, *ITAcotron 2*, which is able not only to generate intelligible and natural Italian speech but also to emulate the voice of a given speaker.

A speech synthesiser of this kind can endow an ECA with a personalised voice, which contributes to increasing human engagement in the ongoing conversation [3, 19]. Furthermore, such a system can have a wider range of useful applications, such as speech-to-speech translation, navigation systems, and fairy tale reciting. It could also allow people who have lost their voices to recover their ability to communicate naturally even though they can’t provide a satisfying amount of previously recorded voice samples.

Note that, in achieving this goal, we distanced ourselves from the potential misuses accompanying the development of this technology. Impersonating someone’s voice without their consent represents a clearer example of such a drift. In developing our system, we stuck to AI Google’s principles,[2] and we hope that future users and developers will act in full compliance with these guidelines.

The experimental hypotheses underlying our research are the following: Firstly, we examined the possibility of performing language adaptation using transfer learning. To do so, we fine-tuned a customised TTS model, previously trained on

[2] https://ai.google/principles.

English data, porting it to the Italian language. Then, we conducted a listening task to collect subjective ratings expressing the intelligibility and naturalness of the synthesised Italian speech.

Secondly, we investigated the feasibility of modelling fine-grained, speaker-dependent characteristics while generating new speech from text. These features crucially contribute to the uniqueness of each utterance for the fact that every speaker is provided with a unique vocal identity. Thus, we represented speakers' *voiceprints* as fixed-dimensional embeddings (i.e. d-vectors) to condition the speech generation with the purpose of synthesising different voices from a single model.

Finally, since we employed a speaker encoder architecture that was previously trained on an English verification task to extract speakers' voiceprints, we implicitly tested the feasibility of performing language adaptation. In fact, if the speaker encoder was able to derive discriminative features for representing English speakers, it should have been able to do the same for Italian speakers. Thus, we designed this experiment to prove whether the speaker encoder was language-agnostic and applicable to a language that differs from the source one (English), without being re-trained or fine-tuned. We did so because we wanted to observe the zero-shot behaviour of this network in the new language (Italian). We apply the same strategy to the neural vocoder, by using a network previously trained with English recordings and not refined on the new language.

5 Corpora

In this work, we considered three different corpora of Italian speech, containing recited utterances. We reported the main statistics about the corpora in Table 1. All clips were re-sampled at 22.050 Hz.

Mozilla Common Voice[3] (MCV) is a publicly available corpus of crowd-sourced audio recordings [1]. Contributors can either donate their voice by reading prompted sentences or validate clips by listening to others' recordings. Clips in this corpus have a native sample rate of 48.000 Hz.

Table 1 Statistics on the considered corpora for the Italian fine-tuning of the spectrogram predictor: Mozilla Common Voice (MCV), VoxForge (VF), and Ortofonico (Ort)

	Time [h]			Clips			Speakers		
Corpus	Train	Validation	Test	Train	Validation	Test	Train	Validation	Test
MCV	79.1	26.5	26.4	50,322	16,774	16,775	5151	3719	3743
VF	13.6	1.8	1.8	7176	913	918	903	584	597
Ort	2.9	0.4	0.3	1436	164	159	20	20	20

[3] https://commonvoice.mozilla.org.

VoxForge[4] (VF) is a multilingual open-source speech database that includes audio clips collected from volunteer speakers. Clips in this corpus have a native sample rate of 16.000 Hz.

Ortofonico (Ort) is a subset of the CLIPS[5] corpus of Italian speech, collected for a project funded by the Italian Ministry of Education, University and Research. Audio recordings come from radio and television programmes, map task dialogues, simulated conversations, and text excerpts read by professional speakers. Clips in this corpus subset have a native sample rate of 22.050 Hz.

Apart from the three presented corpora, we used some clips from a private collection of audiobooks in the human evaluation step. We reported further details in Sect. 7.

6 ITAcotron 2 Synthesis Pipeline

The model we propose is called ITAcotron 2. It is an entire TTS pipeline, complete with speaker conditioning, based on Tacotron 2 [13, 25]. The pipeline is composed of a phonemiser, a speaker encoder (used for the conditioning step), a spectrogram predictor, and a neural vocoder. We reported a scheme of the pipeline in Fig. 1.

The core of the model we are presenting is the spectrogram predictor. We referred to the Tacotron 2 implementation and weights provided by Mozilla[6] [10]. The

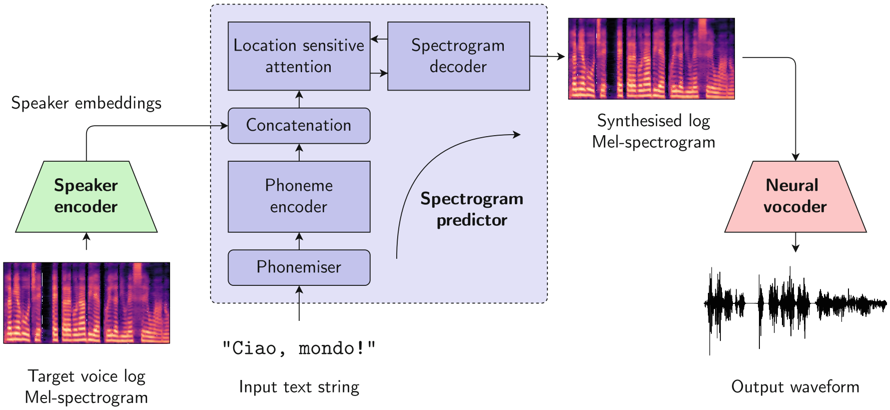


Fig. 1 ITAcotron 2 speech synthesis pipeline

[4] http://www.voxforge.org.

[5] http://www.clips.unina.it.

[6] Repository link, https://github.com/mozilla/TTS; reference commit link, https://github.com/mozilla/TTS/tree/2136433.

model uses a *phoneme encoder* to represent the input sequence to utter and an *autoregressive decoder* to generate the target spectrogram; an intermediate attention mechanism provides the input-output alignment. With respect to the original implementation, we only extended the employed phonemiser[7] to accommodate Italian's accented vowels as additional input characters. Code and pre-trained weights for speaker encoder and vocoder came from the Tacotron 2 same source.

We divided the fine-tuning process of the spectrogram predictor into two steps. In this way, we iteratively improved the output quality.

The former used only the data coming from the MCV corpus, which constituted the majority of the available data. Due to the low quality of the input audio recordings, we leveraged this step mostly to drive the network's weights towards the target language. The noisy and sometimes poorly uttered clips of this corpus resulted in synthesised clips of poor quality, which sometimes were impossible to understand. This fine-tuning was performed for 52.271 update steps (identified through validation) on mini-batches containing 64 clips each; other hyper-parameters were left unchanged from the reference implementation.

The latter fine-tuning leveraged both VF and Ort corpora. Audio clips in these corpora were of a noticeable higher quality than those of MCV in terms of audio clearness and speaker articulation. As a result, the outputs of this final stage had significantly less background noise, and the content was highly intelligible. We performed this second fine-tuning for 42.366 update steps (identified through validation) on mini-batches containing 42 clips each; other hyper-parameters were left unchanged from the reference implementation.

To achieve speaker conditioning, we concatenated the encoder representation of the spectrogram predictor with a *speaker embedding*. These embeddings were extracted from a speaker verification model [4], similar to that of the reference work by [13]. For the vocoder, instead, we adopted the more recent *full-band MelGAN* (FB-MelGAN) vocoder [34].

Notice that while we fine-tuned the spectrogram synthesis network, we did not apply the same process to the speaker embedding and neural vocoder networks. We did so because we wanted to observe the zero-shot behaviour of these networks in the new language. In this way, we could assess whether the two models are language-agnostic.

7 Evaluation Approach

Similar to [13], we divided the evaluation process of our fine-tuned model into two listening experiments:

[7] https://pypi.org/project/phonemizer/.

- Evaluation of *intelligibility and naturalness* (I&N) of the speaker-conditioned synthesised samples
- Evaluation of *speaker similarity* (SS) of the speaker-conditioned synthesised samples

For both experiments, we asked subjects to rate different aspects—in a 1 to 5 scale, with 0.5 increments [12]—of several audio clips. We divided the 70 participants into 20 experimental groups for both listening tasks. We prompted participants of each group with the same clips.

In the I&N experiment, we assigned each group with 4 clip pairs, for a total of 160 among all groups. Each clip pair was composed of a real clip (ground truth) coming from one of the corpora (including an additional private corpus of audiobooks) and a synthetic clip generated in the voice of the ground truth, but with different speech content (i.e. the same voice uttered a different sentence). At this step, we asked subjects to rate the intelligibility and naturalness of each clip, separately. Clips were presented in a random order (to avoid biases) and were rated right after listening.

In the SS experiment, we assigned each group with 16 clips, split into 4 subsets, for a total of 160 among all groups. We divided the SS experiment into three tasks. Each task was composed of a synthetic clip and three real clips. Subjects compared the synthetic clip to each of the other three real clips:

1. A real clip containing an utterance in the voice of the same speaker of the synthetic utterance (the *same-speaker* comparison task)
2. A real clip containing an utterance in the voice of a different speaker having the same gender of the speaker of the synthetic utterance (the *same-gender* comparison task)
3. A real clip containing an utterance in the voice of a different speaker having different gender of the speaker of the synthetic utterance (the *different-gender* comparison task)

At this step, we asked subjects to rate how similar the synthetic voice was to the one we paired it with (knowing that the fixed clip was synthetic and the other three real). Real clips were presented in a random order (to avoid biases), and subjects rated the similarity right after listening to a synthetic-real pair.

8 Results

In this section, we report results on the two experiments described in the previous section, by providing both quantitative and qualitative analyses.

We report the mean opinion score (MOS) of each task in Table 2. The overall scores were satisfying and reflected the intentions and the expectations underlying this research.

Table 2 Results of the listening tasks. MOS values are reported as $\mu \pm \sigma$. All values are computed with a support of 280 samples

Task	(Comparison) Task	Model	MOS	95% confidence interval
Intelligibility and naturalness	Intelligibility	ITAcotron 2	4.15 ± 0.78	[4.07, 4.23]
		Ground truth	4.43 ± 0.74	[4.36, 4.50]
	Naturalness	ITAcotron 2	3.32 ± 0.97	[3.22, 3.41]
		Ground truth	4.28 ± 0.86	[4.20, 4.37]
Speaker similarity	Same speaker	ITAcotron 2	3.45 ± 1.07	[3.34, 3.56]
	Same gender	ITAcotron 2	2.78 ± 1.01	[2.68, 2.89]
	Different gender	ITAcotron 2	1.99 ± 1.08	[1.91, 2.08]

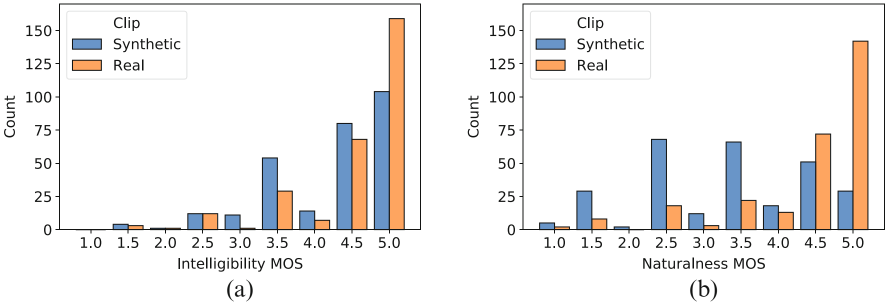
Concerning the I&N evaluation, the first thing that jumps to the eye is the high intelligibility score, very close to real clips. This high score provides clear evidence of how easy it was to understand the linguistic content of the synthetic clips. The naturalness score, however, is lower than that of intelligibility, meaning that it is still possible to distinguish between real and synthetic clips.

Concerning the SS evaluation, instead, the thing that jumps to the eye is the progressive drop in the MOS value. This reduction is precisely the expected behaviour [18]—changing the speaker should lead to lower similarity, especially when the two speakers have different gender. Thus, the MOS obtained for the same-speaker task was the best one, with a promising absolute value; in other words, the system was able to provide a good imitation of the speaker's voice. Then, changing speaker, the MOS dropped, meaning that the synthetic voice was able to express the "personality" of the speaker it was imitating, and so it was clearly distinguishable from other voices). Finally, as expected, a further drop was observed by the different-gender similarity evaluations.

The figures we obtained are close to those obtained by the reference work [13] on similar tasks, for English. However, we choose not to report a direct comparison against the work mentioned above as it focuses on English and the tasks are not perfectly comparable with ours. Nevertheless, obtaining similar scores is a hint that our approach seems sound. More detailed results on the two experiments we conducted are presented in the following.

8.1 Speech Intelligibility and Naturalness

This experiment was meant to evaluate the language adaptation hypothesis by assessing the degree of intelligibility and naturalness exhibited by the synthesised speech. MOS evaluation distributions, for real and synthesised speech, are reported as histograms in Fig. 2.



(a) (b)

Fig. 2 MOS distributions of real and synthetic clips. (**a**) Intelligibility MOS distributions. (**b**) Naturalness MOS distributions

Table 3 Results of Welch's t-test with degrees of freedom (DoF) computed as $n_1 - n_2 - 2$, where n_1 and n_2 are the sample sizes of the first and second sample. The t-value is the value used to produce the probability value (p-value) based on Student's t-distribution. All scores are computed with 558 DoF

	Sample			
Task	First	Second	t-value	p-value
Intelligibility	ITAcotron 2	Ground truth	4.375	1.448e–6
Naturalness	ITAcotron 2	Ground truth	12.414	2.237e–31
Speaker similarity	Same speaker	Same gender	7.606	1.208e–13
	Same speaker	Different gender	16.054	5.812e–48
	Same gender	Different gender	8.962	4.820e–18

Table 2 reports sample size, mean, standard deviation, and 95% confidence interval computed with empirical bootstrap [5] for real and synthetic audio MOS distributions. Results from Welch's unequal variance t-test for intelligibility are reported in Table 3.

Ground-truth recordings obtained a higher intelligibility MOS ($\mu = 4.43$, $\sigma = 0.74$) than did audio clips synthesised by our model ($\mu = 4.15, \sigma = 0.78$), $t(558) = 4.37$, $p < 0.05$. Our proposed model achieved 4.1 intelligibility MOS compared to 4.34 of the ground truth. This confirmed the system ability to synthesise speech with highly intelligible content.

The speech content of the data sets used to train our synthesiser, especially in the second fine-tuning, was in most of the cases clearly and smoothly comprehensible. This could have influenced positively the output intelligibility. Moreover, using a significant amount of hours and speakers in training could have lead the model both to improve its generalisation ability and to distinguish easily useless from useful spectral information at prediction time. These aspects could have jointly increased the intelligibility of the synthesis.

In exploratory listening sessions, we noticed that the model learnt to smoothly generate out of vocabulary words, long input texts, and complex syntactic structures such as long-distance dependencies and topicalised sentences. Indeed, and surprisingly, there were cases in which its generative performance seemed to be enhanced by the presence of such complexities. This could be originated by the adoption of a double decoder that helped to reduce attention alignment problems at inference time.

However, we also observed that the model was not able to synthesise interrogative direct sentences from text. This may be caused by the fact that interrogative sentences were not present in any training sets. Otherwise, the model could have learnt to reproduce properly the suprasegmental prosody features which distinguish an assertion from a direct question.

With respect to naturalness, the MOS distributions for real and synthesised speech are reported as histograms in Fig. 2b. In addition, to visualise clearly the differences between these two score distributions, two paired plots are presented in Fig. 3b.

Sample size, mean, standard deviation, and confidence interval computed with empirical bootstrap for both real and synthetic distributions are summarised in Table 2. Results from Welch's unequal variance t-test are reported in Table 3.

Ground-truth recordings obtained a higher naturalness MOS ($\mu = 4.28$, $\sigma = 0.86$) than did audios synthesised by our model ($\mu = 3.32$, $\sigma = 0.97$), $t(558) = 12.41$, $p < 0.05$. Our proposed model achieved 3.32 naturalness MOS compared to 4.18 of the ground truth. This might have been due to an evident drawback of the ICV data set which presents a high level of background noise that the synthesiser had learnt to reproduce.

In exploratory listening sessions, we noticed that the naturalness of the synthesised voice mainly varied depending on which speaker embedding was adopted to condition the generative process. Moreover, the model learnt to pause naturally when punctuation marks, such as comma or full stop, were inserted in the input text. This was probably the consequence of not having removed punctuation in speech transcripts when training our system. Thus, the model seems to have learnt some prosodic aspects connected to the presence of punctuation marks.

In analysing the experimental results for both intelligibility and naturalness, we were also interested in discovering whether a linear correlation existed between the amount of training data associated with the voices used to synthesise the experimental stimuli (i.e. independent variable) and the quality of the speech synthesised using those voices (i.e. dependent variable). We expected that the more the system was exposed to speech data belonging to a given voice, the more it would have been able to synthesise high-quality sentences in that voice.

About intelligibility, for each of the 62 voices used in the experiment, we computed the total amount of training sentences associated with each of them and its average intelligibility MOS from the scores it received in the listening test. Then, we represented each voice as a point in a Cartesian plane, where x axis stood for its average intelligibility MOS and y axis stood for the amount of time (in minutes) that voice was seen during training.

Pearson product-moment correlation and Spearman's rank correlation coefficient were computed to detect whether a correlation could be identified between the amount of training data each voice is assigned to and the intelligibility MOS each

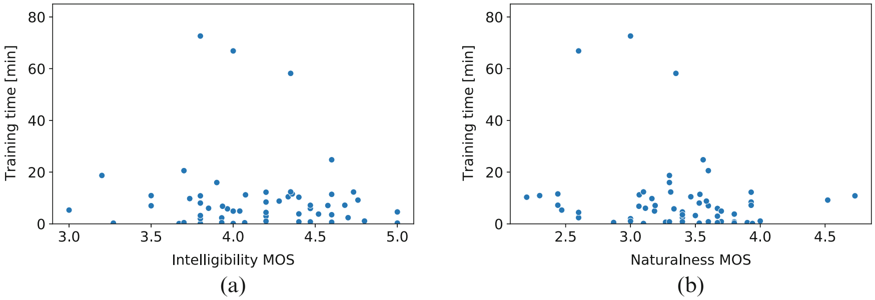


Fig. 3 I&N scores vs duration of synthetic clips. (**a**) Voice-wise average intelligibility MOS vs. duration time in training. (**b**) Voice-wise average naturalness MOS vs. duration time in training

voice received in the listening test. The p-value for the Pearson correlation between these two variables was above the significance level of 0.05, which means that the correlation coefficient was not significant ($r = -0.07, n = 62, p = 0.55$). The same occurred for Spearman's ρ correlation coefficient ($r_s = 0.03$, $n = 62$, $p = 0.83$).

The scatterplot of Fig. 3a summarises the results. Overall, no correlation was found between the amount of training data associated with a given voice and the intelligibility MOS it received in the subjective listening test.

The same correlation was investigated for speech naturalness. We represented each of the 62 voice in a Cartesian plane, where x axis stood for its average naturalness MOS and y axis stood for the amount of time (in minutes) that voice was seen in training.

We derived Pearson product-moment correlation and Spearman's rank correlation coefficient to detect whether there existed a correlation between the amount of training data each voice is presented with and the naturalness MOS each voice received when evaluated. With respect to Person correlation, the p-value between these two variables was 0.16 ($r = -0.18$, $n = 62$). Since it was greater than the significance level of 0.05, there was inconclusive evidence about the significance of the association between these two variables. The same occurred for Spearman's ρ correlation coefficient ($r_s = -0.21, n = 62, p = 0.10$).

The scatterplot in Fig. 3b summarises these results. Overall, no correlation was found between the amount of training data associated with a given voice and the naturalness MOS it received in the subjective listening test.

8.2 *Speaker Similarity*

The second experiment was meant to evaluate the effectiveness of the strategy we used to condition our TTS generative pipeline. Namely, we wanted to assess whether the model developed the capability of disentangling speaker-dependent characteristics from linguistic component when synthesising new speech.

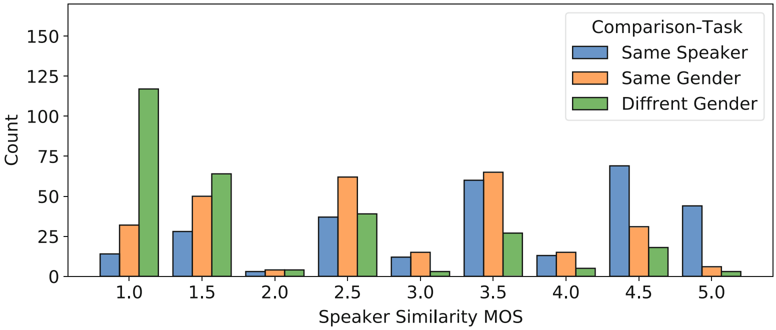


Fig. 4 SS comparison task similarity MOS distributions

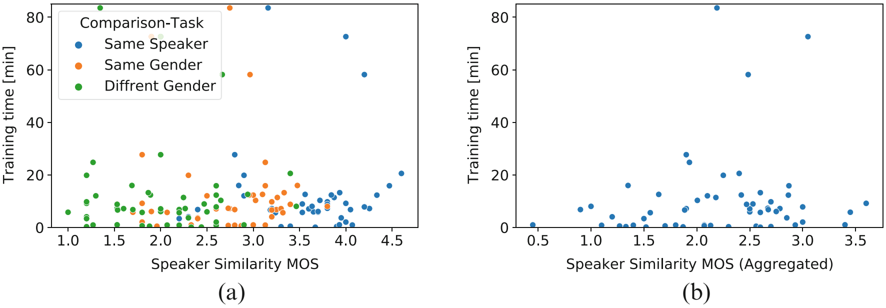


Fig. 5 Speaker similarity vs duration. (**a**) Separate. (**b**) Combined

Moreover, we were also interested in verifying whether the speaker encoder adopted to extract speaker discriminative characteristics was language-agnostic and thus applicable to Italian speech samples.

Figure 4 reports the MOS distributions of same-speaker, same-gender, and different-gender similarity judgements as histograms and box plots, respectively. In addition, to clearly visualise the differences between these three score distributions, in Fig. 5, two paired plots are presented.

We report mean, standard deviation, and confidence interval computed with empirical bootstrap for SS comparison task MOS distributions in Table 2.

Results from three Welch's unequal variance t-tests for speaker similarity are reported in Table 3. Same-speaker similarity judgements obtained a higher MOS ($\mu = 3.45$, $\sigma = 1.07$) than same-gender similarity judgements ($\mu = 2.78$, $\sigma = 1.01$), $t(558) = 7.60$, $p < 0.05$. Same-gender similarity judgements obtained a higher MOS ($\mu = 2.78$, $\sigma = 1.01$) than different-gender similarity judgements ($\mu = 1.99$, $\sigma = 1.08$), $t(558) = 8.96$, $p < 0.05$. Same-speaker similarity judgements obtained a higher MOS ($\mu = 3.45$, $\sigma = 1.07$) than different-gender similarity judgements ($\mu = 1.99$, $\sigma = 1.08$), $t(558) = 16.05$, $p < 0.05$.

On one hand, same-speaker similarity average score amounts to 3.45, which indicates a reasonable resemblance between the voices of audios synthesised in a given voice and their real counterparts. On the other hand, same-gender similarity MOS is 2.78, while different-gender similarity MOS is 1.99. It means that the raters recognised correctly the extent in which voices of speakers of the same or different genders differed from each other.

These results confirmed the overall model ability to decouple speech content from speaker-dependent characteristics and to learn high-fidelity speaker representations that can be exploited to generate speech in a desired voice. In general, our system was able to transfer effectively speakers' gender for all the 57 voices used in the experiment (Table 2).

In exploratory listening sessions, we noticed mismatches on regional accent, between synthesised and target voices. This could have influenced negatively the similarity judgements since participants were not told how to judge accents, so they could have rated poorly because of accent mismatches rather than because of low model quality. Accent mismatches could have been caused by the use of a speaker encoder trained only on English-accented speech. Since English and Italian do not match in terms of accent, this could have prevented the system from properly and systematically transferring accents from target to synthesised audios.

However, some raters reported that the accent of the synthesised speech exhibited a clear resemblance with the accent of the original recording from the same speaker. This effect was larger on IVF data set which contains more marked regional accents. It follows that, to better exert a control on the synthesised rendition, a further refinement is required to decouple speaker individual characteristics from prosody and linguistic features.

As for the previous experiment, in analysing the results, we were interested in verifying whether a correlation existed between the amount of training recordings associated with the voices used to synthesised the experimental stimuli (i.e. independent variable) and the degree of resemblance (i.e. dependent variable) that particular voice exhibited with its real counterpart. We expected that the more the system had been exposed to speech data belonging to a given voice, the more it would have been able to transfer speaker-dependent characteristics at inference time.

In assessing whether such correlation occurred, we adopted both a disaggregated and an aggregated approach to represent the similarity MOS score associated with each voice.

In the first approach, we intended the similarity score as three distinct scores. Namely, for each of the 57 voices used in the experiment, we computed comparison task average similarity MOS, separately. Additionally, we computed the amount of training data (in minutes) associated with each of these voices. It follows that each voice used in the experiment was represented in a Cartesian plane, where the x axis stood for either its comparison task average similarity score and y axis stood for the amount of time that voice was seen in training.

We derived the Pearson product-moment correlation and Spearman's rank correlation coefficient between comparison task average scores and the corresponding amount of recordings associated with each voice in training. The p-value for the Pearson correlation between the same-speaker similarity MOS and the amount of training data was above the significance level of 0.05, which indicated that the correlation coefficient was not significant ($r = 0.21$, $n = 57$, $p = 0.11$). The same occurred for the Pearson correlation computed between same-gender similarity MOS and the amount of training data synthesised ($r = -0.01$, $n = 57$, $p = 0.92$) and between different-gender similarity MOS and the amount of training data synthesised ($r = 0.03$, $n = 57$, $p = 0.79$).

Concerning Spearman's ρ correlation coefficient, the p-value between SS similarity MOS and the amount of training recordings is about 0.40, which indicates that there was a moderate positive correlation [6] between these two variables ($p = 0.002$, $n = 57$). Differently, the p-value for Spearman's correlation between same-gender similarity MOS and the amount of training data synthesised was above the significance level of 0.05, which indicates that the correlation coefficient was not significant ($r_s = 0.22$, $n = 56$, $p = 0.08$). The same occurred between the different-gender similarity MOS and the amount of training data ($r_s = 0.14$, $n = 56$, $p = 0.30$).

A scatterplot in Fig. 5a summarises these findings. It highlights that in general no correlation exists between the amount of training data associated with a given voice and the similarity MOS evaluations it received in the subjective listening test. However, a moderate correlation can be detected between same-speaker similarity MOS and the amount of training data (green dots on the right).

In verifying the existence of such correlation, the other approach we adopted to represent the similarity MOS was to assign to each voice a single score. This score was derived by subtracting from the same-speaker similarity MOS the average between same-gender and different-gender similarity MOS. Thus, each voice used in the experiment was represented in a Cartesian plane, where the x axis stood for its aggregated similarity MOS score and y axis stood for the amount of time the spectrogram predictor processed that voice during training.

We computed the Pearson product-moment correlation and Spearman's rank correlation coefficient between the aggregated similarity MOS and the amount of time each voice was seen in training. The p-value for the Pearson correlation was above the significance level of 0.05, which indicates that the correlation coefficient was not significant ($r = 0.15$, $n = 56$, $p = 0.23$). The same occurred for Spearman's ρ correlation ($r_s = 0.21$, $n = 56$, $p = 0.10$).

A scatterplot in Fig. 5b summarises these results. Overall, no correlation was derived between the amount of training data associated with the speakers used to construct the experimental stimuli and their similarity aggregated MOS.

Since we didn't find significant correlations between intelligibility, naturalness, and speaker similarity evaluations and the amount of training data, we concluded that the model acquired a discrete ability to generalise to speakers for which a low amount of training recordings was provided or even to speakers unseen during training.

9 Conclusion and Future Work

In this paper, we showed our approach to adapt a speech synthesis pipeline from English to Italian. The procedure was language-agnostic, but spectrogram prediction network required fine-tuning data in the target language. To show how some pipeline components can be used without language adaptation, we also introduced a speaker embedding network (to achieve speaker conditioning) and a neural vocoder.

Opinion scores from a human evaluation session showed that the adaptation was successful in terms of intelligibility and naturalness. Concerning speaker conditioning, the result was not as sharp as for the first evaluation; yet, we obtained a satisfying similarity score, matching that of the reference model.

In general, the main issue arising in modelling prosody features implicated in conveying linguistic and non-linguistic information in speech is that they are fully entangled. However, some of these characteristics are speaker-dependent, such as accent and idiolect, while others are speaker-independent such as prosodic phrasing, lexical stress, and tune variation. Thus, if on one side controlling a given prosody phenomenon using a unique latent embedding space would allow a complete control over all linguistic and non-linguistic components, disentangling speaker-dependent and speaker-independent characteristics enables simplified models and better decoupling, from a human perspective, of the controlled aspects.

Traditionally, prosody modelling relied on labelling prosodic phenomena and developing rule-based systems or statistical models from speech data. These strategies allow a high control on speech products, but they require to derive hand-crafted features, which is difficult and time-consuming in the presence of large data sets. In contrast, end-to-end neural TTS systems permit to generate high-fidelity speech with a simplified pipeline and to learn prosody as an inherent part of the model. Even though these unsupervised methods are extensively used, they still miss exerting an accurate and clear control over the output prosody.

Thus, future research will focus on different prosody phenomena, to identify strategies to model micro and macro prosody patterns. To do so, a possibility will be to leverage independent representations, in the form of GST or latent vectors from a variational auto-encoder (VAE), for each of the speech traits of interest. In addition, a multi-head attention mechanism can increase the system parallelisation capability and help to cope with the hardness arising with RNN application in modelling human speech long-distance dependencies.

On the whole, to improve the model ability to regulate its generation in accordance to the mental and emotional status of its interlocutor, we could augment the input with a multi-modal stream of information, encoding the features of the previous conversational turn. For instance, the user's prompt could be represented by feeding the TTS with a visual input encoding the user's facial expression, an acoustic input encoding prosodic information characterising her/his speech, and a linguistic embedding encoding information related to the meaning of her/his words. The TTS should then be trained to align its generation in terms of prosody and linguistic content, in accordance with the previous conversational turn—thing that we, as humans, do in every conversational exchange.

Acknowledgments This work was partially supported by the European Union's Horizon 2020 project *WorkingAge* (grant agreement no. 826232).

Appendix

The source code developed during this project is available at the following link: https://github.com/vincenzo-scotti/ITAcotron_2. Inside the repository, we also provide the links to download the weights of the fine-tuned model ITAcotron 2, for Italian speech synthesis. We remind that the original source code we forked, and the weights of the speaker encoder and neural vocoder, was taken from the reference open-source project developed by Mozilla[8]

References

1. Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F.M., Weber, G.: Common voice: A massively-multilingual speech corpus. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11–16, 2020, pages 4218–4222. European Language Resources Association (2020)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings (2015)
3. Cassell, J.: Embodied conversational agents: Representation and intelligence in user interfaces. AI Mag. **22**(4), 67–84 (2001)
4. Chung, J.S., Huh, J., Mun, S., Lee, M., Heo, H.-S., Choe, S., Ham, C., Jung, S., Lee, B.-J., Han, I.: In defence of metric learning for speaker recognition. In: Meng, H., Xu, B., Zheng, T.F. (eds.) Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25–29 October 2020, pages 2977–2981. ISCA (2020)
5. Dekking, F.M., Kraaikamp, C., Lopuhaä, H.P., Meester, L.E.: A Modern Introduction to Probability and Statistics: Understanding why and how. Springer Science & Business Media, Berlin (2005)
6. Fowler, J., Cohen, L., Jarvis, P.: Practical Statistics for Field Biology. Wiley, Hoboken (2013)
7. Fujisaki, H.: Prosody, models, and spontaneous speech. In: Sagisaka, Y., Campbell, N., Higuchi, N. (eds.) Computing Prosody, Computational Models for Processing Spontaneous Speech, pp. 27–42. Springer, Berlin (1997)

[8] Repository link, https://github.com/mozilla/TTS; reference commit link, https://github.com/mozilla/TTS/tree/2136433.

8. Gilmartin, E., Collery, M., Su, K., Huang, Y., Elias, C., Cowan, B.R., Campbell, N.. Social talk: making conversation with people and machine. In Chaminade, T., Nguyen, N., Ochs, M., Lefèvre, F. (eds.) Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents, ISIAA@ICMI 2017, Glasgow, United Kingdom, November 13, 2017, pp. 31–32. ACM, New Yrok (2017)
9. Griffin, D.W., Lim, J.S.: Signal estimation from modified short-time fourier transform. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '83, Boston, Massachusetts, USA, April 14–16, 1983, pp. 804–807. IEEE, Piscataway (1983)
10. Gölge, E.: Solving Attention Problems of TTS Models with Double Decoder Consistency (2020)
11. Hsu, P.-C., Wang, C.-H., Liu, A.T., Lee, H.-Y.: Towards robust neural vocoding for speech generation: A survey. CoRR **abs/1912.02461** (2019)
12. ITU-T Recommendation: P.910: Subjective video quality assessment methods for multimedia applications (1999)
13. Jia, Y., Zhang, Y., Weiss, R.J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Lopez-Moreno, I., Wu, Y.: Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada, pp. 4485–4495 (2018)
14. Jurafsky, D., Martin, J.H.: Speech and Language Processing, 2nd edn. Prentice-Hall, Hoboken (2009)
15. Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., van den Oord, A., Dieleman, S., Kavukcuoglu, K.: Efficient neural audio synthesis. In: Dy, J.G., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018. Proceedings of Machine Learning Research, , vol. 80, pp. 2415–2424. PMLR (2018)
16. Kenmochi, H.: Vocaloid and Hatsune Miku phenomenon in Japan. In: Interdisciplinary Workshop on Singing Voice (2010)
17. Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W.Z., Sotelo, J., de Brébisson,A., Bengio, Y., Courville, A.C.: Melgan: Generative adversarial networks for conditional waveform synthesis. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, pp. 14881–14892 (2019)
18. Leung, Y., Oates, J., Chan, S.P.: Voice, articulation, and prosody contribute to listener perceptions of speaker gender: A systematic review and meta-analysis. J. Speech Language Hearing Res. **61**(2), 266–297 (2018)
19. Moridis, C.N., Economides, A.A.: Affective learning: Empathetic agents with emotional facial and tone of voice expressions. IEEE Trans. Affect. Comput. **3**(3), 260–272 (2012)
20. Ping, W., Peng, K., Gibiansky, A., Arik, S.Ö., Kannan, A., Narang, S., Raiman, J., Miller, J.: Deep voice 3: Scaling text-to-speech with convolutional sequence learning. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, April 30–May 3, 2018, Conference Track Proceedings. OpenReview.net (2018)
21. Prieto, P., Borràs-Comes, J., Roseano, P.: Interactive Atlas of Romance Intonation (2010)
22. Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.-Y.: Fastspeech: Fast, robust and controllable text to speech. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, pp. 3165–3174 (2019)
23. Schuller, B., Batliner, A.: Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing. Wiley, Hoboken (2013)
24. Schuller, D., Schuller, B.W.: The age of artificial emotional intelligence. Computer **51**(9), 38–46 (2018)

25. Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R.J., Saurous, R.A., Agiomyrgiannakis, Y., Wu, Y.: Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, April 15–20, 2018, pp. 4779–4783. IEEE, Piscataway (2018)
26. Skerry-Ryan, R.J., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R.J., Clark, R., Saurous, R.A.: Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In: Dy, J.G., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, July 10–15, 2018. Proceedings of Machine Learning Research, vol. 80, pp. 4700–4709. PMLR (2018)
27. Suni, A., Kakouros, S., Vainio, M., Simko, J.: Prosodic prominence and boundaries in sequence-to-sequence speech synthesis. CoRR **abs/2006.15967** (2020)
28. Taylor, P.: Text-to-Speech Synthesis. Cambridge University Press, Cambridge (2009)
29. Torrey, L., Shavlik, J.: Transfer learning. In: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, pp. 242–264. IGI Global, Pennsylvania (2010)
30. van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A.W., Kavukcuoglu, K.: Wavenet: A generative model for raw audio. In: The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, 13–15 September 2016, p. 125. ISCA (2016)
31. Wang, Y., Skerry-Ryan, R.J., Stanton, D., Wu, D., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q.V., Agiomyrgiannakis, Y., Clark, R., Saurous, R.A.: Tacotron: A fully end-to-end text-to-speech synthesis model. CoRR **abs/1703.10135** (2017)
32. Wang, Y., Skerry-Ryan, R.J., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q.V., Agiomyrgiannakis, Y., Clark, R., Saurous, R.A.: Tacotron: Towards end-to-end speech synthesis. In: Lacerda, F. (ed.) Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, August 20–24, 2017, pp. 4006–4010. ISCA (2017)
33. Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R.J., Battenberg, E., Shor, J., Xiao, Y., Jia, Y., Ren, F., Saurous, R.A.: Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In: Dy, J.G., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, July 10–15, 2018. Proceedings of Machine Learning Research, vol. 80, pp. 5167–5176. PMLR (2018)
34. Yang, G., Yang, S., Liu, K., Fang, P., Chen, W., Xie, L.: Multi-band melgan: Faster waveform generation for high-quality text-to-speech. In: IEEE Spoken Language Technology Workshop, SLT 2021, Shenzhen, January 19–22, 2021, pp. 492–498. IEEE, Piscataway (2021)
35. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8–13 2014, Montreal, Quebec, pp. 3320–3328 (2014)