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

Intelligibility and naturalness of articulatory synthesis with VocalTractLab compared to established speech synthesis technologies

Paul Konstantin Krug, Simon Stone, Peter Birkholz

Technische Universität Dresden

paul_konstantin.krug@tu-dresden.de

Abstract

In this work, the current state-of-the-art of articulatory speech synthesis (VocalTractLab) is compared to a wide range of different text-to-speech systems that once represented or still represent the continuously evolving state-of-the-art of speech synthesis technology. The comparison systems include neural and concatenative synthesis by Google and Microsoft, as well as Hidden Markov Model-based, unit-selection and diphone synthesis developed at universities (using MaryTTS, MBROLA and DRESS). A small corpus of 15 German sentences was synthesized using the text-to-speech (and, if available, re-synthesis) functionalities of each system. The intelligibility of the synthesized utterances was evaluated in an ASR experiment. The naturalness of the utterances was evaluated in a multi-stimulus Likert test by 50 German native speakers. As an additional reference, recordings of natural speech were used in the experiments. It was found that the articulatory synthesis can achieve a performance on par with the non-commercial synthesis systems in terms of intelligibility and naturalness, while being significantly outperformed by the commercial synthesis systems.

Index Terms: text-to-speech, articulatory speech synthesis.

1. Introduction

From mechanical speech apparatuses [1], to electrical vocal tract analogues [2–6], up to sophisticated computer simulations [7–13]: articulatory speech synthesis has been a topic of research for centuries. Despite the fact that this kind of synthesis can be considered the most natural approach to speech synthesis, as it aims to directly model the speech production process that happens in a human vocal tract, it never played a significant role outside the academic world [14]. This is mainly due to (i): the difficulties that arise from modelling the time-dependent vocal tract geometries, which need to be controlled up to a very precise level. This requires a deep understanding of human speech production and knowledge of articulatory movements, which are not easily accessible experimentally. (ii) For a long time, no complete aerodynamic-acoustic simulation of the vocal tract existed. And (iii): At any given time, better sounding alternative methods were available (e.g. formant synthesis, parametric synthesis, concatenation synthesis or recently neural synthesis) that did require less or no explicit knowledge of articulatory movements. Furthermore, the generation of synthesized utterances with articulatory synthesizers generally involves a lot of manual tuning, which is usually a very time consuming process that requires expert knowledge.

Apart from very few (and outdated) exceptions such as GnuSpeech [15], no modern articulatory text-to-speech (TTS) systems were available until now. This situation has changed with the recent development of the state-of-the-art articulatory synthesizer VocalTractLab[1] [12] (VTL) version 2.3 that introduced a fully automatic phoneme-to-speech conversion for German. Using this functionality, it is possible to generate high quality re-syntheses of any given German utterance. By extending the VTL with an additional grapheme-to-phoneme conversion (G2P) and an intonation model, it is possible to setup a complete TTS pipeline [16]. Although the produced speech by VTL sounds intelligible, it is not yet known how VTL speech compares against state-of-the-art systems of well established speech synthesis technologies in terms of intelligibility and naturalness.

The current study aims to rank the VTL synthesis among widely used speech synthesis technologies such as diphone, Hidden Markov Model (HMM), unit-selection and neural synthesis. It extends the state of research on articulatory synthesis by the following contributions:

1. A full articulatory TTS system based on the open source software VocalTractLab is presented (VTL-TTS).
2. A fair comparison of articulatory synthesis (both fully-automatic TTS and manual re-synthesis, which in this case means to derive phone durations and pitch information from natural speech recordings) with eight different types of syntheses, as well as natural speech, is presented in terms of intelligibility and naturalness. Although this involves systems under active development and thus can only serve as a snapshot, it gives valuable insight into the speech synthesis landscape on the whole at this point in time.

2. Methods

A small corpus of 15 German sentences, presented in Table 1, was synthesized in a neutral speaking style using different TTS systems, namely Google Cloud TTS [17–21][2], Microsoft Azure TTS [22][2], MaryTTS [23], DRESS [24], as well as VTL-TTS. Additionally, natural speech recordings of the 15 sentences were manually re-synthesized using VTL and MBROLA [25]. The intelligibility of the syntheses was evaluated using automatic speech recognition (ASR). The naturalness of the syntheses was evaluated by 50 German native speakers in a listening experiment. Finally, a deep learning-based system for speech naturalness evaluation (NISQA) [26] was evaluated against the results from the listening experiment. All audio sample files and the data files necessary to reproduce the synthesized files are available in the supplementary materials[3].

[1] https://www.vocaltractlab.de/ (Last visited 22.04.2021).

[2] Since the companies' systems are proprietary and continuously developed, no exact descriptions of the systems are available. Hence, the references should be understood as an (incomplete) overview of important contributions to the used technologies.

[3] https://github.com/TUD-STKS/TTS_Comparison_SSW21 (Last visited 22.04.2021).

10.21437/SSW.2021-18

2.1. Articulatory synthesis and TTS pipeline

2.1.1. VocalTractLab

The articulatory synthesizer VTL provides a one-dimensional aero-acoustic simulation [27] within a model of the vocal tract that is based on magnetic resonance imaging (MRI) scans of a real human vocal tract [12]. The current version VTL 2.3 provides three different types of vocal fold models [28–30]. In this study, the geometric glottis model [28] was used, which is the VTL default.
During the time domain simulation, the articulatory dimensions of VTL are controlled by a set of time-dependent functions, a so called *gestural score* [31, 32]. A gestural score consists of several tiers, which describe the shape of the articulators, the glottis shape, the intonation and the lung pressure, respectively. While VTL allows for the direct construction and manipulation of the gestural score and thus precise control, VTL 2.3 also offers a more convenient higher level user interface (for German speech). By providing a sequence of phone labels and their respective acoustic durations, a gestural score of articulatory movements can be automatically generated, excluding the pitch contour. Therefore, only the missing intonation needs to be generated either manually or by some external means (see Section 2.2.2). The generated score can be freely edited after the automatic generation, which allows a semi-automatic workflow where an utterance is initialized automatically and then tuned manually (e.g., to match a reference utterance).

2.1.2. VTL-TTS

The used VTL-TTS pipeline consists of several stages. First, a given plain input text is converted into its SAMPA transcription, using a proprietary Web service by Aristech Gmbh [16]. The transcription also provides further annotations, such as the utterance's syllables and information on the linguistic stress of the syllables. Subsequently, a set of 70 phonetic and linguistic features is calculated. An intonation contour for the utterance is then predicted using these features fed to a deep neural network. Finally, the phone durations are predicted using empirically determined, context-dependent reference values taken from [33]. The phone sequence is then turned into a gestural score using the segment sequence interface of VTL 2.3 described above and then converted into audio.

2.2. Stimuli preparation and preprocessing

2.2.1. TTS synthesis

Six of the TTS voices were accessed via their Web clients, namely Microsoft Azure TTS[4], Google Cloud TTS[5] and MARYTTS[6]. In case of the former two services, both a neural synthesis (in the following referred to as *Azure-Neural* and *Google-Neural*), and a parametric/unit-selection[7] synthesis (in the following referred to as *Azure-Standard* and *Google-Standard*) were used to produce the desired samples. In case of the MARYTTS system, samples were synthesized via HMM-based synthesis (using the German voice *dfki-pavoque-neutral-hsmm de male hmm*, in the following referred to as *dfki-HMM*) and via unit-selection synthesis (using the German voice *dfki-pavoque-neutral de male unitselection general*, in the following referred to as *dfki-unit*) [34, 35]. The other MARYTTS parameters were set the following way (for both voices): "Input Type": *TEXT*, "Output Type": *AUDIO*, "Audio-Out": *WAVE_FILE* and "Audio-Effects": *Default (all turned off)*. For the Azure-Neural and Azure-Standard syntheses, the parameter "Voice" was set to *Conrad (Neural)* and *Stefan*, respectively. The other parameters were set the following way (for both the neural and parametric/unit-selection syntheses): "Language": *German (Germany)*, "Voice Style": *General*, "Speaking Speed": *1.00* and "Pitch": *0.00*. For the Google-Neural and Google-Standard syntheses, the parameter "Voice type" was set to *WaveNet* and *Basic*, respectively. The parameter "Voice name" was set to *de-DE-Wavenet-B* and *de-DE-Standard-B*, respectively. The other parameters were set the following way (for both voices): "Language": *Deutsch (Deutschland)*, "Speed": *1.00*, "Pitch": *0.00* and "Audio device profile": *Default*. Furthermore, samples were created using DRESS, which is a pure diphone TTS synthesis using the TD-PSOLA [36] algorithm. The male voice *Jörg* was used during the synthesis. The "Rhythmn" parameter was set to *Klatt* and the "Intonation" parameter was set to *Fujisaki (dt)*. Finally, samples were created using VTL-TTS using the previously described processing pipeline.

2.2.2. Re-synthesis

The term re-synthesis describes a synthetic reproduction of a natural speech recording that matches the original recording as precisely as possible. In case of VTL, a manual re-synthesis performed by an expert represents the highest quality that is currently achievable with the software. Hence, manual re-syntheses can give an idea of the maximum possible VTL-TTS performance, if the pre-processing (i.e. G2P, phone duration prediction and intonation prediction etc.) was ideal. For this reason, the manual VTL re-synthesis was also evaluated against the TTS systems in the experiments. In order to generate the natural utterances, necessary for the re-syntheses, a 24-year-old German native speaker was recorded at a sample rate of 44.1 kHz. Subsequently, the recordings were loaded into VTL, where the respective phoneme sequence was aligned with the natural speech so that the reproduced speech matched the original utterances as closely as possible in terms of timing. In order to match the intonation as well, the natural f_0 contour of each sentence was extracted using the software PRAAT [37]. The software TARGETOPTIMIZER [38, 39] (TO) was used in order to fit the natural contours using the TARGET-APPROXIMATION-MODEL [40,41] (TAM). This step was necessary since the pitch and articulatory gestures of VTL are based on the TAM. The obtained pitch gestures were loaded into VTL and manually fine-tuned when necessary. The audio samples were synthesized using the speaker file *JD2*, which is the default VTL speaker. There is no relation between the recorded speaker and the speaker on whose data the JD2 model is based on (apart from both persons being male). The audio samples were exported as WAV files with a sample rate of 44.1 kHz.
In order to have a second re-synthesis system to compare with VTL, an additional diphone re-synthesis was made using the open source software MBROLA[8]. The same phone durations and pitch contours as for the VTL re-syntheses were used. However, the used database for the male German speaker *de2* does

[4] https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech (last visited 09.02.2021).

[5] https://cloud.google.com/text-to-speech (last visited 22.01.2021).

[6] http://mary.dfki.de:59125/ (last visited 22.01.2021).

[7] The companies are not specific about the exact technology that is used for the standard (non-neural) voices. They state such voices are created using either parametric or unit-selection synthesis or a mixture of both.

[8] https://github.com/numediart/MBROLA (Last visited 14.02.2021).

	Utterance	IPA	Translation
1	Aber sehen will sie ihn doch.	ˈʔaːbɐ ˈzeːn vɪl ziː ʔiːn dɔx	*But she wants to see him.*
2	Er sah viele bunte Regenbogen.	eːɐ̯ zaː ˈfiːlə ˈbʊntə ˈʁeːgn̩ˌboːgn̩	*He saw many colourful rainbows.*
3	Chabos wissen wer der Babo ist.	ˈt͡ʃaːboːs ˈvɪsn̩ veːɐ̯ deːɐ̯ ˈbaːboː ʔɪst	*The boys know who the boss is.*
4	Das Telefon ist seit sieben Tagen kaputt.	das ˈteːləfoːn ʔɪst zaɪ̯t ˈziːbn̩ ˈtaːgn̩ kaˈpʊt	*The phone has been broken for seven days.*
5	Die Artikel waren wieder vorrätig.	diː ˌʔaʁˈtiːkl ˈvaːʁən ˈviːdɐ ˈfoːɐ̯ˌʁɛtɪç	*The products were in stock again.*
6	Die Soße ist viermal übergekocht.	diː ˈzoːsə ʔɪst ˈfiːɐ̯ maːl ˈʔyːbɐgəˌkɔxt	*The sauce boiled over four times.*
7	Die Straßenbahn fuhr weiter geradeaus.	diː ˈʃtʁaːsn̩ˌbaːn fuːɐ̯ ˈvaɪ̯tɐ gəʁaːdəˈʔaʊ̯s	*The tram continued straight ahead.*
8	Diese Zeitung ist bereits veraltet.	ˈdiːzə ˈt͡saɪ̯tʊŋ ʔɪst bəˈʁaɪ̯t͡s fɛɐ̯ˈʔaltət	*This newspaper is already outdated.*
9	Sie fährt keinen Ferrari, sondern einen Maserati.	ziː fɛːɐ̯t ˈkaɪ̯nən feˈʁaːʁiː ˈzɔndɐn ˈʔaɪ̯nən mazəˈʁaːtiː	*She does not drive a Ferrari, but a Maserati.*
10	Benno gefällt die orange Vase.	ˈbɛno gəˈfɛlt diː ʔoˈʁaŋʒə ˈvaːzə	*Benno likes the orange vase.*
11	Es kann hilfreich sein, wenn man weiß, wie ein Unterstand gebaut wird.	ʔɛs kan ˈhɪlfʁaɪ̯ç zaɪ̯n vɜn man vaɪ̯s viː ʔaɪ̯n ˈʔʊntɐʃtant gəˈbaʊ̯t vɪʁt	*It can be helpful to know how to build a shelter.*
12	Er schützt vor Kälte, Wind und Niederschlägen.	ʔeːɐ̯ ʃvt͡st foːɐ̯ ˈkɛltə vɪnt ʔʊnt ˈniːdɐʃlɛːgən	*It protects against cold, wind and precipitation.*
13	Conny glaubt eigentlich nicht mehr an den Osterhasen.	kɔni glaʊ̯pt ˈaɪgn̩tlɪç nɪçt meːɐ̯ ʔan deːn ˈoːstɐˌhaːzən	*Conny doesn't really believe in the Easter Bunny any more.*
14	Sie läuft schnell hin.	ziː lɔɪ̯ft ʃnɛl hɪn	*She runs there quickly.*
15	Der Petersdom ist das Wahrzeichen des Vatikans.	deːɐ̯ ˈpeːtɐsˌdoːm ɪst das ˈvaːɐ̯ˌt͡saɪ̯çn̩ dɛs vatiˈkaːns	*St Peter's Basilica is the landmark of the Vatican*

Table 1: *The used utterances in German, their canonical IPA transcription, and English translation.*

not contain a glottal stop. The durations of existing glottal stops in the segment sequence files used in the VTL re-synthesis were therefore split half and half between the left and right neighbouring phones. Secondary diphthongs such as /oːɐ/ were broken down into the two individual vowels, each with half the total duration. For MBROLA re-syntheses, the f_0 contours were constructed as linear interpolations between f_0 support points. On average, as many f_0 support points were used as there were phones in the utterance.

2.2.3. Natural speech

In addition to the synthetic speech samples, natural speech recordings of the 15 German sentences were also evaluated in all experiments to serve as anchor points. The speaker for the natural stimuli was different from the speaker for the re-synthesis reference recordings in order to avoid possible biases, e.g. regarding the f_0 contour. For the natural samples, a male 27-year-old non-professional German native speaker was recorded at a sample rate of 44.1 kHz. As in the previous case, there is no relation between this speaker and the VTL JD2 model. For the recordings a large diaphragm condenser microphone was used (*Microtech Gefell M930*). It was connected to a low-noise pre-amplifier (*Behringer Eurorack MX 1602*). The pre-amp was then connected to an audio interface (*MOTU 896 HD*) which was connected to a PC via FireWire. The natural speech audio samples were recorded in a sound-proofed audio studio. The speaking style was neutral.

2.2.4. Re-sampling and loudness normalization

The various synthetic and natural speech samples have different sample rates. Hence, the amount of high frequency content differs among the samples, since no frequencies can be present beyond the respective Nyquist frequencies. However, the presence or absence of high frequencies are part of the technologies that should be evaluated in this study. Hence, the samples were intentionally not downsampled to the smallest sample rate present in the data (which would implicate a high frequency cut-off for some of the samples). Instead, they were upsampled to the largest present sample rate that is $44.1\,\mathrm{kHz}$ to facilitate further processing without distorting the frequency contents. Afterwards, all samples were loudness normalized. This is very important since the various speech samples produced with the different technologies (even though peak normalized) differed widely in their loudness. However, the loudness of a sample might significantly impact the rating on a psychometric scale [42]. Hence, the audio amplitudes of all samples were first peak normalized to $-1\,\mathrm{dB\,FS}$. Subsequently, the integrated loudness according to the ITU-R BS.1770-4 recommendation (measured in dB LUFS) was calculated for each sample using the PYTHON library PYLOUDNORM. Using the same tool, all samples were then loudness normalized to the minimal loudness obtained in the previous step, which was $-25.7\,\mathrm{dB\,LUFS}$. This way all stimuli had the same loudness and the maximum peak amplitude among all samples was $-1\,\mathrm{dB\,FS}$.

2.3. Evaluation of intelligibility

Evaluating the intelligibility of the audio samples in a perception experiment with human listeners would be challenging, due to the high number of participants that would be required to obtain an adequate statistical power. Hence, automatic speech recognition was chosen as a tool to measure the intelligibility of the synthetic and natural speech samples. Four state-of-the-art commercial ASR systems, namely Google Web API, Microsoft Azure speech-to-text, IBM Watson speech-to-text and Wit.ai (owned by Facebook), were accessed via their respective API using the PYTHON libraries SPEECHRECOGNITION and IBM-WATSON. Four different systems were used in order to reduce a possible impact from the biases of the ASR systems towards certain speech styles, f_0, voice etc. The audio files were sent to each service and the speech-to-text conversion was returned as a string. The word error rate (WER) between the true text and the ASR answer was calculated using the python

library JIWER. Thereby, both the true and recognized strings were pre-processed in the following way: The punctuation was removed from the strings, all characters were converted to lower case, double or multiple white spaces were converted to a single white space, leading and trailing whitespaces were removed.

2.4. Evaluation of naturalness

2.4.1. Listening experiment

In order to evaluate the naturalness of all samples, an online perception experiment was carried out using the tool webMUSHRA[9] [43]. The experiment was designed as a multi-stimulus Likert test. Thereby, participants would see a single page per sentence that contained all eleven versions of that sentence. Each version had to be played and rated in order to proceed to the next page. Participants could play an audio sample as often as desired. Each page displayed the utterance text at the top of the page. Below that each page featured the following instructions (translated to English): "*On a scale of 1 to 5 stars, how natural (i.e., how human) does each utterance sound? (1: Very unnatural, 2: Rather unnatural, 3: Neither, 4: Rather natural 5: Very natural). You have to play all versions to the end and rate all versions.*"
At the beginning of the test, participants were asked to play an example audio sample in order to adjust their listening volume to a pleasant level. Thereby, the example file was the sentence (translated to English): "*Please listen to the following sample sentence and adjust the volume so that you find it comfortable.*". It was synthesized using the IBM TTS[10] online client that was not used for other samples in the experiment. Just as all other samples, the example file was loudness normalized to -25.7 dB LUFS.
In total, 50 subjects (18 male, 32, female) aged between 18 and 50 years (median: 24.0 years, mean: 26.6 ± 6.7 years) participated in the experiment. Participants were required to be German native speakers, but due to the online nature, no additional screenings were conducted. To avoid a bias of the results, experts in (articulatory) speech synthesis technology were not encouraged to participate.

2.4.2. NISQA

As an automatic kind of speech quality assessment, the pre-trained CNN-BLSTM NISQA-TTS[11] [26] model was used in order to evaluate the naturalness of the synthesized speech samples. The predicted NISQA scores were then compared to the ratings of the human listeners to evaluate the predictive power of such an automated assessment system. This is of particular interest for articulatory synthesis, since the produced speech is not directly derived from original, human recordings, which might break the assumptions of a pre-trained assessment model.

3. Results

3.1. Evaluation of intelligibility

The word error rates across all samples are shown in Figure 1 for all four ASR systems separately. While the median of each distribution is zero, one can see that the means (Google: 0.08 ± 0.15, Azure: 0.13 ± 0.22, Wit.ai: 0.16 ± 0.24, IBM: 0.19 ± 0.27) differ due to the different amount of outliers. Based on two-sided Mann-Whitney U tests (MWU tests), the Google WER distribution of the Google ASR system is significantly different from those of Wit.ai and IBM ($p < 0.01$), but not significantly different from the distribution of the Azure system ($p > 0.01$). No significance was observed between permutations of Azure, Wit.ai and IBM ($p > 0.01$).

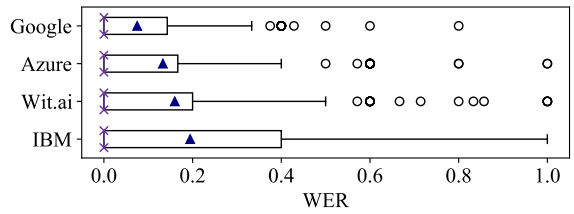

Figure 1: *Word error rates across all speech samples, separated into single distributions for the four ASR systems, shown as box plots. The position of the median of each distribution is indicated by two x-shaped markers. The position of the respective mean is indicated by a triangle.*

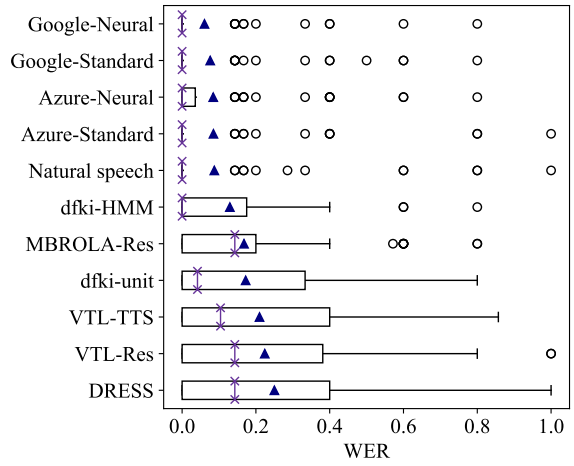

Figure 2: *Word error rates across all ASR systems, separated into single distributions for each type of synthesis, shown as box plots. Medians are indicated by two x-shaped markers. Means are indicated by a triangle.*

Figure 2 shows the WER distributions for all tested synthesis types across all ASR systems. The synthesis types are sorted by their respective mean (top: best performance, bottom: worst performance). It is observed that the first five synthesis types (Google-Neural and Standard, Azure-Neural and Standard, as well as the natural speech) achieve a median WER of 0.0 across all ASR systems, which means they are mostly identified correctly. The distributions differ slightly in their mean values, but this is mainly due to the outliers. While the median WER of the dfki-HMM syntheses is also 0.0, the distribution is still significantly broader than the distribution of the natural speech samples and those of the Google syntheses ($p < 0.01$, based on two-sided MWU tests), resulting in a higher mean. No significant difference was found among permutations of WER dis-

[9] Despite its name, the tool is not limited to MUSHRA tests, but can be used for several kinds of listening experiments. In this analysis, a multi-stimulus Likert test was performed.

[10] https://www.ibm.com/demos/live/tts-demo/self-service/home (Last visited 22.01.2021).

[11] https://github.com/gabrielmittag/NISQA (Last visited 14.02.2021).

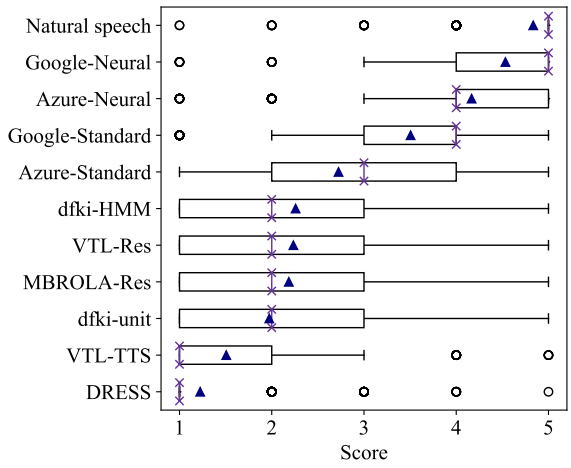

Figure 3: *Likert scores across all listeners, separated into single distributions for each type of synthesis, shown as box plots. Medians are indicated by two x-shaped markers. Means are indicated by a triangle.*

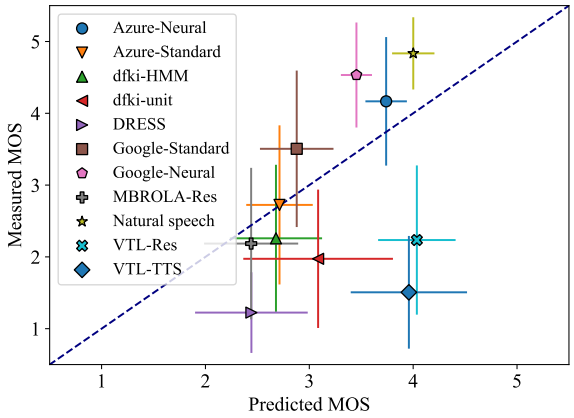

Figure 4: *Subjective MOS measured in the listening experiment plotted against predicted MOS values determined with the NISQA network. The errorbars indicate the $\pm 1\,\sigma$ interval around the mean.*

tributions of the non-commercial synthesis systems. The WER medians of the five worst performing technologies deviate from zero and range from 0.04 (dfki-unit) to 0.14 (DRESS). The mean values range from 0.17 ± 0.22 (MBROLA) to 0.25 ± 0.3 (DRESS).

3.2. Evaluation of naturalness

The results from the listening test are shown in Figure 3. The order of synthesis types decreases in performance from top to bottom (top: rated as most natural, bottom: rated as most unnatural). The constituents of all possible distribution pairs, except for permutations of dfki-HMM, VTL-Res and MBROLA-Res, are significantly different ($p < 0.01$) from each other, based on two-sided MWU tests. The natural speech performed best, with a mean rating of 4.84 ± 0.50. It is followed by the two neural syntheses (Google-Neural: 4.53 ± 0.73, Azure-Neural: 4.17 ± 0.89). The commercial parametric/unit-selection syntheses perform worse than the neural syntheses, with mean values of 3.51 ± 1.09 and 2.72 ± 1.11, respectively. The re-syntheses perform worse and similar to the dfki syntheses. VTL-TTS is rated significantly less natural (1.51 ± 0.78) and DRESS samples were rated to be the least natural sounding samples (1.22 ± 0.56).

Figure 4 shows the measured subjective scores plotted against the predicted scores from the NISQA-TTS model. It is observed that the predicted scores do not agree well with the measured data. While the performance of VTL-Res, VTL-TTS, dfki-unit and DRESS is greatly overestimated, the performance of the neural syntheses and the natural speech is underestimated. The linear correlation coefficient between the predicted and measured values is $\rho = 0.28$.

4. Discussion

A small corpus of 15 German sentences was synthesized using a wide range of different TTS systems that once represented or still represent the continuously evolving state-of-the-art both in the commercial and the academic domain of speech synthesis technology. The intelligibility and the naturalness of the syntheses was evaluated and compared against natural speech in an ASR experiment and in a listening experiment, respectively.

From the ASR experiment, it was observed that the WER's of the commercial TTS systems did not differ significantly from the WER of natural speech. It can be concluded that the obtained WER is rather limited by the recognition performances of the ASR systems and less by the quality of the artificial speech samples. The main reasons for the significantly worse performance of the non-commercial systems are probably the synthesis artifacts that are quite audible in case of dfki-HMM, dfki-unit and DRESS. Further, the intonation and phone durations have an impact on the performance. This is well exemplified in case of VTL-TTS and VTL-Res. Despite pitch contours and phone durations copied from natural utterances, VTL-Res performs worse than VTL-TTS with regard to WER. It seems likely that the longer and more uniform distibuted phone durations of the VTL-TTS system increase the intelligibility in this case.

As expected, the participants in the listening experiment considered the natural speech samples as the most natural sounding samples. Despite not being directly comparable due to the experimental setup of the Likert test, the obtained scores for the Google-Neural and Standard syntheses are in agreement with the MOS scores reported in [21]. In terms of naturalness, VTL-TTS performs significantly worse than VTL-Res. Hence, a more realistic modeling of intonation and phone duration could improve the articulatory TTS pipeline a lot.

To conclude, none of the TTS systems is both, as natural and as intelligible as natural speech yet, even though the commercial neural voices come very close. However, the non-commercial syntheses perform significantly worse. Within the subgroup of academic systems, semi-automatic articulatory re-synthesis proved to be very competitive in terms of naturalness and was not significantly worse than the best non-commercial system dfki-HMM. However, in order for articulatory synthesis to keep up with the modern commercial systems, the overall quality would have to improve greatly. Starting points for improving intelligibility and naturalness of VTL syntheses include an improved modeling of the noise sources inside the vocal tract, modeling tongue-loops [44], and microprosodic effects.

5. Acknowledgements

This work has been partially funded by the Leverhulme Trust Research Project Grant RPG-2019-241: "High quality simulation of early vocal learning".

6. References

[1] W. v. Kempelen, "Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine." Degen, 1791.

[2] H. Dudley *et al.*, "A synthetic speaker," *J. Franklin Inst.*, vol. 227, no. 6, pp. 739–764, 1939.

[3] H. Dudley, "Remaking speech," *J. Acoust. Soc. Am.*, vol. 11, no. 2, pp. 169–177, 1939.

[4] H. K. Dunn, "The calculation of vowel resonances, and an electrical vocal tract," *J. Acoust. Soc. Am.*, vol. 22, no. 6, pp. 740–753, 1950.

[5] K. N. Stevens *et al.*, "An electrical analog of the vocal tract," *J. Acoust. Soc. Am.*, vol. 25, no. 4, pp. 734–742, 1953.

[6] G. Rosen, "Dynamic analog speech synthesizer," *J. Acoust. Soc. Am.*, vol. 30, no. 3, pp. 201–209, 1958.

[7] P. Mermelstein, "Articulatory model for the study of speech production," *J. Acoust. Soc. Am.*, vol. 53, no. 4, pp. 1070–1082, 1973.

[8] P. Rubin *et al.*, "An articulatory synthesizer for perceptual research," *J. Acoust. Soc. Am.*, vol. 70, no. 2, pp. 321–328, 1981.

[9] S. Maeda, "A digital simulation method of the vocal-tract system," *Speech Commun.*, vol. 1, no. 3-4, pp. 199–229, 1982.

[10] P. Rubin *et al.*, "CASY and extensions to the task-dynamic model," in *1st ETRW on Speech Production Modeling: From Control Strategies to Acoustics; 4th Speech Production Seminar: Models and Data*, 1996, pp. 125–128.

[11] O. Engwall, "Combining MRI, EMA and EPG measurements in a three-dimensional tongue model," *Speech Commun.*, vol. 41, no. 2-3, pp. 303–329, 2003.

[12] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PloS ONE*, vol. 8, no. 4, p. e60603, 2013.

[13] A. Pont *et al.*, "Finite element generation of sibilants /s/ and /z/ using random distributions of Kirchhoff vortices," *Int. J. Numer. Methods Biomed. Eng.*, vol. 36, no. 2, p. e3302, 2020.

[14] C. H. Shadle and R. I. Damper, "Prospects for articulatory synthesis: A position paper," in *SSW4*, 2001.

[15] D. Hill *et al.*, "Real-time articulatory speech-synthesis-by-rules," in *Proc. AVIOS*, vol. 95, 1995, pp. 11–14.

[16] S. Stone *et al.*, "Prospects of articulatory text-to-speech synthesis," in *Proc. ISSP*, 2020 (Accepted).

[17] X. Gonzalvo *et al.*, "Recent advances in Google real-time HMM-driven unit selection synthesizer," in *Proc. Interspeech*, 2016, pp. 2238–2242.

[18] H. Zen *et al.*, "Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices," in *Proc. Interspeech*, 2016, pp. 2273–2277.

[19] A. van den Oord *et al.*, "WaveNet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[20] Y. Wang *et al.*, "Tacotron: A fully end-to-end text-to-speech synthesis model," *CoRR*, vol. abs/1703.10135, 2017.

[21] J. Shen *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," *CoRR*, vol. abs/1712.05884, 2017.

[22] Y. Ren *et al.*, "FastSpeech: Fast, robust and controllable text to speech," in *Proc. NeurIPS*, vol. 32, 2019, pp. 3171–3180.

[23] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: A tool for research, development and teaching," *Int. J. Speech Technol.*, vol. 6, pp. 365–377, 2003.

[24] R. Hoffmann *et al.*, "Evaluation of a multilingual TTS system with respect to the prosodic quality," in *Proc. ICPhS*, vol. 3, 1999, pp. 2307–2310.

[25] T. Dutoit *et al.*, "The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes," in *Proc. ICSLP*, vol. 3, 1996, pp. 1393–1396.

[26] G. Mittag and S. Möller, "Deep learning based assessment of synthetic speech naturalness," in *Proc. Interspeech*, 2020, pp. 1748–1752.

[27] P. Birkholz, "Enhanced area functions for noise source modeling in the vocal tract," in *Proc. ISSP*, 2014, pp. 32–40.

[28] P. Birkholz *et al.*, "Perceptual optimization of an enhanced geometric vocal fold model for articulatory speech synthesis." in *Proc. Interspeech*, 2019, pp. 3765–3769.

[29] ——, "Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis," in *Proc. Interspeech*, 2011, pp. 2681–2684.

[30] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell Syst. Tech. J.*, vol. 51, no. 6, pp. 1233–1268, 1972.

[31] P. Birkholz, "Control of an articulatory speech synthesizer based on dynamic approximation of spatial articulatory targets," in *Proc. Interspeech*, 2007, pp. 2865–2868.

[32] P. Birkholz *et al.*, "Model-based reproduction of articulatory trajectories for consonant-vowel sequences," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, no. 5, pp. 1422–1433, 2010.

[33] B. Möbius and J. Von Santen, "Modeling segmental duration in German text-to-speech synthesis," in *Proc. ICSLP*, vol. 4, 1996, pp. 2395–2398.

[34] I. Steiner *et al.*, "Symbolic vs. acoustics-based style control for expressive unit selection," in *SSW7*, 2010, pp. 114–119.

[35] M. Schröder *et al.*, "Open source voice creation toolkit for the MARY TTS Platform," in *Proc. Interspeech*, 2011, pp. 3253–3256.

[36] C. Hamon *et al.*, "A diphone synthesis system based on time-domain prosodic modifications of speech," in *Proc. ICASSP*, 1989, pp. 238–241.

[37] P. Boersma and D. Weenick, "Praat: Doing phonetics by computer (version 6.0.43) [computer program]," 2005, http://www.praat.org (Last visited 15.02.2021).

[38] P. Birkholz *et al.*, "Estimation of pitch targets from speech signals by joint regularized optimization," in *Proc. EUSIPCO*, 2018, pp. 2075–2079.

[39] P. K. Krug *et al.*, "Targetoptimizer 2.0: Enhanced estimation of articulatory targets," in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2021*, 2021, pp. 145–152.

[40] Y. Xu and Q. E. Wang, "Pitch targets and their realization: Evidence from Mandarin Chinese," *Speech Commun.*, vol. 33, no. 4, pp. 319–337, 2001.

[41] S. Prom-On *et al.*, "Modeling tone and intonation in Mandarin and English as a process of target approximation," *J. Acoust. Soc. Am.*, vol. 125, no. 1, pp. 405–424, 2009.

[42] E. Vickers, "The loudness war: Background, speculation, and recommendations," in *Audio Engineering Society Convention 129*. Audio Engineering Society, 2010.

[43] M. Schoeffler *et al.*, "webMUSHRA — A comprehensive framework for web-based listening tests," *J. Open Res. Software*, vol. 6, no. 1, p. 8, 2018.

[44] H. Nam *et al.*, "Hearing tongue loops: Perceptual sensitivity to acoustic signatures of articulatory dynamics," *J. Acoust. Soc. Am.*, vol. 134, no. 5, pp. 3808–3817, 2013.