Plotting Poetry
https://doi.org/10.7592/PP2025.06.koch_et_al
Popular Voices: Computational Analysis of Poetry and Song. ELM Scholarly Press: Tartu 2025.

Blowing *The Boy's Magic Horn*: Plotted and Synthesised Romanticism

Julia Koch
University of Stuttgart, Germany
julia.koch@ims.uni-stuttgart.de
0009-0003-7026-3383

Toni Bernhart
University of Stuttgart, Germany
toni.bernhart@ilw.uni-stuttgart.de
0000-0002-7255-2504

André Blessing
University of Stuttgart, Germany
andre.blessing@ims.uni-stuttgart.de

Gunilla Eschenbach
German Literature Archive, Germany
Gunilla.Eschenbach@dla-marbach.de

Markus Gärtner
University of Stuttgart, Germany
markus.gaertner@ims.uni-stuttgart.de

Kerstin Jung
University of Stuttgart, Germany
kerstin.jung@ims.uni-stuttgart.de

Nora Ketschik
University of Stuttgart, Germany
nora.ketschik@ilw.uni-stuttgart.de

Anna Kinder
German Literature Archive, Germany
anna.kinder@dla-marbach.de

Jonas Kuhn
University of Stuttgart, Germany
jonas.kuhn@ims.uni-stuttgart.de

Sandra Richter
German Literature Archive, Germany
sandra.richter@dla-marbach.de

Nadja Schauffler
University of Stuttgart, Germany
nadja.schauffler@ims.uni-stuttgart.de

Rebecca Sturm
German Literature Archive, Germany
rebecca.sturm@dla-marbach.de

Gabriel Viehhauser
University of Stuttgart, Germany
gabriel.viehhauser-mery@ilw.uni-stuttgart.de

Ngoc Thang Vu
University of Stuttgart, Germany
thang.vu@ims.uni-stuttgart.de

Abstract

At the heart of the »textklang« (Sound of Text) project is the development of a mixed-method approach to investigate the interrelation between written lyric poetry and its sonic realisation. It is an interdisciplinary collaboration between the German Literary Archive in Marbach and the University of Stuttgart that includes literary studies, digital humanities, computational linguistics, laboratory phonology and speech technology. The project's corpus is centred on the poetry of Romanticism and is based on the holdings of the German Literature Archive. In our contribution, we illustrate a multi-perspective approach to one collection within the corpus entitled *The Boy's Magic Horn* (*Des Knaben Wunderhorn*), edited in three volumes by A. v. Arnim and C. Brentano in 1806 and 1808 and including more than 700 poems. *The Boy's Magic Horn* is considered one of the most influential poetry collections in German literature because of its vivid reception in both 'low' folkloristic cultures and 'high' culture, especially in musical settings (G. Mahler, J. Brahms, F. Silcher). We share some initial outcomes of the ongoing research process considering quantitative, textual, prosodic, and sonic aspects. As part of the project's methodological and experimental toolbox, we present a speech synthesis model that has been trained on this sub-corpus and which results in a better realisation of poetic speech compared to synthesis models exclusively trained on prose data. Finally, we discuss the challenges this data pose to automatic processing tools.

1 Introduction

This paper presents work within the »textklang« (Sound of Text) project, which is based at the German Literary Archive in Marbach and at the Institute for Natural Language Processing and the Institute of Literary Studies at the University of Stuttgart. The project includes literary studies, digital humanities, computational linguistics, laboratory phonology and speech technology. It was funded by the German Federal Ministry of Education and Research and ran from 2021 to 2023.

The »textklang« project follows a mixed-method approach to investigating the interrelation between written lyric poetry and its sonic expression. Its objectives include analysing the prosody of written poetry as it is realised in recitation and musical performance, developing text-to-speech synthesis for studying prosody perception, and collecting metadata to analyse and visualise intertextuality and intermediality.

The research corpus we worked with focuses on lyric poetry from the period of German Romanticism. This poetry is particularly suitable for studying oral, aural, sonic, and prosodic aspects, as it was influenced by the concept of 'Volkslied' (popular or folk poetry) developed by Johann Gottfried Herder and the Grimm brothers (J. Grimm and W. Grimm 1815; Herder 1773). The poetry often depicts human and animal voices (especially birds), oral transmission, witness, authorship, and various styles of performing spoken words (Eschenbach and Richter 2020; Richter et al. 2023).

The research corpus was predominantly fed from holdings of the German Literature Archive, which contains round 4,250 printed texts and 2,700 recordings of recitations and musical performances. Furthermore, this corpus was increased by our own production. We recorded 700 poems from *The Boy's Magic Horn* and 180 *Children's and Household Tales* by the Grimm brothers (J. Grimm and W. Grimm 1812, 1815).

In this paper we will focus on *The Boy's Magic Horn*, a collection of German poems, and its relevance within the project. After we have 'plotted' the main characteristics of this sub-corpus, we will present our speech synthesis model trained on these data and discuss the challenges this corpus poses to automatic processing tools.

2 *The Boy's Magic Horn*

The Boy's Magic Horn. Old German songs (*Des Knaben Wunderhorn. Alte deutsche Lieder*) is a collection of poetry, written in German and edited in three volumes by Achim von Arnim and Clemens Brentano in 1806 and 1808 (Arnim and Brentano 1806, 1808a,b). It includes about 700 lyric poems that are versified and rhymed, as well as a few texts in prose (comments and an essay on poetology). Predominant genres are songs, ballads, children's songs, prayers, and short plays. Subjects and themes in *The Boy's Magic Horn* are everyday life, Christianity and Catholicism and faith, history, myths, nature, social and political experiences (especially in rural environments). *The Boy's Magic Horn* also includes period-specific misogynistic and anti-Jewish stereotypes. One of the collection's main characteristics is that it is deeply rooted in late-medieval and Early Modern literary tradition. *The Boy's Magic Horn* is considered to be one of the most influential collections of lyric poetry in German literature because of its vivid reception both in 'low' and 'high' folkloristic cultures, especially in musical settings, for instance by Gustav Mahler, Johannes Brahms, Friedrich Silcher, Robert Schumann, Karlheinz Stockhausen, and Bob Dylan (Benischek 2008).

2.1 Text and Audio Data

Data from *The Boy's Magic Horn* in the »textklang« project comprise text files and audio files that we produced, aligned, annotated and synthesised. Files for each text, in .txt format, were extracted from DTA (Deutsches Textarchiv 2023), which is a collection of digitised and scholarly edited first printings of canonical German literature. Each text file was collocated manually with the printed scholarly edition by Heinz Rölleke (Brentano 1975–1978) in order to provide highly reliable texts. These texts served as a basis both for reading during the recording sessions and for later analysis. After the audio files were edited and reviewed, some text files had to be slightly altered to match the recordings exactly, for example in German-specific variable inflectional endings or with omitted words and to establish an accurate alignment of text and audio files.

The audio files were recorded and supervised by Wolfgang Wokurek at the Phonetic Lab at the University of Stuttgart. Toni Bernhart was the speaker. In addition to the poems themselves, the recordings included paratexts such as title pages, content tables, indices, and sources on which the poems are based and which are often cited as part of the titles by the authors. Poems written in Low German and in Alemannic dialects were, however, not included mainly to be able to record everything with the same speaker. The signal

chain was composed of one Neumann U 87 Ai microphone (in omni-directional setting), audio interface Yamaha 1608-D, and mixing console Yamaha CL1, which were connected through a Dante Controller 4.1.0.5 network. Software used in recording and post production processes was Steinberg Nuendo Live 1.1.0 and Audacity 3.1.3 without a limiter, compressor, or filter plug-ins. *The Boy's Magic Horn* audio files recorded in 48 kHz 24 bit mono .wav format encompass 19 GB, which corresponds to approximately 36 hours of time.

2.2 Data and Metadata Visualisation

In order to provide data for further analysis on the interrelation of prosody and structural patterns, we set up a collection of metadata. They include the titles of each poem, the number of tokens per text, the genre of the texts, the authors and titles of the sources and pre-texts with an obvious intertextual relation to the poems as well as authors and titles of later texts which obviously allude to the poems, and the years of publication both of sources and later adaptations. We used Keshif to visualise data and metadata. Keshif is a web-based visualisation and analytics tool that allows users to explore datasets quickly (Yalçın et al. 2016). To prepare data for Keshif, data were collected, converted into a JSON file, and imported into the tool.

3 Text-to-Speech Overview

The primary focus of »textklang« is to investigate the interplay of a poem's text, its sonic realisation in recitation, and listeners' perception in an experimental setting. To conduct these experiments, we generated auditory test materials by re-synthesising and manipulating recordings of recitations using text-to-speech synthesis (TTS) (cf. Schauffler et al. 2022). Hence, high-quality speech synthesis for poetic data was crucial for our applications. In this paper, we will describe a short experiment on 'plotting' the sound of *The Boy's Magic Horn* using speech synthesis. We trained a TTS model on our recordings of this sub-corpus and examined how the model realised poetic speech and handled genre and speaker specific characteristics compared to a TTS model trained exclusively on prose.

We followed the established approach on text-to-speech synthesis to break down the challenge of generating a waveform from text into several subtasks (Lux and Vu 2022; Ren, Hu, et al. 2020; Skerry-Ryan et al. 2018). A first step was to transform the input text from a character sequence into a phoneme sequence with a grapheme-to-phoneme (G2P) conversion model, also referred to as phonemiser. While it is possible to omit this pre-processing step and train a TTS model directly on grapheme input (see e.g. Shen et al. 2018) employing a phonetic transcription is usually advantageous since a phonetic transcription represents the audio much more closely than graphemes, helping to reduce mispronunciation errors.

In the following step, the TTS model, for example Tacotron 2 (Shen et al. 2018), Fastspeech 2 (Ren, Hu, et al. 2020), or the more recent PortaSpeech (Ren, Liu, et al. 2021), generates a mel-spectrogram from the phoneme sequence.

This is a challenging task since there are infinitely many possible speech variations that can correspond to the same text: Depending on speaker identity, conveyed emotion, focus highlighting through pitch accents, or just some minor variations in prosody, the speech signal will be different. While some variations can be more or less acceptable than others, the number of adequate realisations remains infinite. This is also known as the one-to-many mapping problem in TTS (Ren, Hu, et al. 2020). Thus, learning to generate an acceptable speech representation for text with deep neural networks usually requires huge amounts of high-quality audio data and aligned text. Finally, a vocoder such as WaveNet (Oord et al. 2016) or HiFi-GAN (Kong et al. 2020) transformed the mel-spectrogram to a waveform.

4 Experimental Setup and Results

We attempted to synthesise poetic speech before in Koch et al. 2022, where we used a pretrained multilingual TTS model which we first finetuned on German prose data and then further trained on 20 poems read by a single professional speaker in a second finetuning step. The model trained on poetry was perceived as sounding by far 'more poetic' compared to the model exclusively trained on prose in a human evaluation study. In view of these promising results, we took on our PoeticTTS approach by using a robust pretrained multilingual model which we then finetuned on German prose data as our baseline. For our final model, we additionally finetuned on our recordings of *The Boy's Magic Horn*, for simplicity we call this the *poetry model*. In this experiment, we wanted to take a closer look at how certain genre specific properties can be learned from training data. In particular, we measured articulation rate and duration of pauses and further compare F_0 contours produced by the human speaker as well as the baseline and the poetry model on a small test set comprising seven poems from the collection.

4.1 Implementation Details

Our speech synthesis model was built in python with the open-source Toolkit IMS Toucan as described in Lux, Koch, et al. 2022. It was a modified implementation of FastSpeech 2 (Ren, Hu, et al. 2020) with Conformer blocks (Gulati et al. 2020) both in encoder and decoder instead of Transformer (Vaswani et al. 2017), and Fast-Pitch style (Łańcucki 2021) phone-wise averaging of F_0 and energy values instead of the original frame-wise prediction of these values. G2P conversion in IMS Toucan was performed by the *phonemizer* (Bernard and Titeux 2021) python-package using espeak-ng[1] as the backend. Further, IMS Toucan provided an integrated aligner (Lux, Koch, et al. 2023) for temporal alignment of phonemes with the corresponding audio during training. As vocoder we took the toolkit's implementation of HiFi-GAN (Kong et al. 2020). For the vocoder as well as the aligner and FastSpeech 2 models we used the

[1] https://github.com/espeak-ng/espeak-ng.

provided pretrained multilingual models of IMS Toucan release v2.2 which are trained on approximately 400 hours of data in 14 languages.

4.2 Data Pre-Processing

To train our baseline model we finetuned the aligner and FastSpeech 2 models on a part of the German subset of the pretraining data for 10,000 steps. In particular, we used the Thorsten-Corpus (Müller and Kreutz 2021) as well as the HUI-Audio-Corpus- German (Puchtler et al. 2021), which contain almost exclusively prosaic data.

To train the poetry model, we used recordings of 211 poems from *The Boy's Magic Horn*. These recordings and their corresponding text transcripts were already edited and checked as described in Section 2. However, training the TTS model required short audio snippets of around five to 20 seconds. Thus, we could not use the recordings as they are but had to cut them into shorter segments. We therefore cut the texts of the poems into chunks of two verses and automatically segmented the audio file according to the durations of each chunk using the aligner model. Specifically, we calculated the durations of each phoneme in the audio by means of the aligner model, and then summed the durations of the phonemes for each two-verse chunk in the transcript to obtain the total duration of each chunk. The audio file was then split into segments such that each segment corresponded to one chunk.

However, we found that our automatic pre-processing pipeline encountered some major issues: First, our weak line-based heuristic of splitting texts into pieces of two verses did not apply well in some cases due to inconsistencies in shape and formatting of the texts in accordance with the historical first print. Our approach was especially unsuitable in instances where poems in verse form were mixed with accompanying prose text. Second, the phonemiser used in this work is designed to work on modern day German and thus struggled with archaic and inconsistent spelling in *The Boy's Magic Horn*. For example, the German preposition *bei* is often spelled *bey* in our data, which our phonemiser transcribes as [bəˈypsɪlɔn] instead of the correct [baɪ]. Moreover, the texts contained many instances of archaisms, words and wordforms in Ancient Greek and Latin, in Middle High German and regional German dialects, proper nouns, neologisms and onomatopoeia which pose a great challenge to our phonemiser. In addition, the speech data of our recordings was quite different from standard data since the speaker might for example decide to follow the meter of the poem, leading to uncommon emphasis. Further, there is much more prosodic variation in speaking rate and pitch progression, which was challenging for our aligner model trained on prose data. Together with the errors in text segmentation and phonemisation, which propagate to automatic audio segmentation, this led to a considerable number of corrupted samples. To counteract these issues, we excluded samples that interfere with the training of our TTS and aligner models leaving a remainder of 2,348 text–audio pairs.

We selected seven poems to test our models and excluded all samples corresponding to these poems from the training set. We used the remaining data to perform further finetuning of our baseline model for 5,000 steps, yielding

poem	1	2	3	4	5	6	7
human	9.39	10.24	9.14	9.42	10.07	11.90	11.87
baseline	11.75	12.34	11.82	11.88	12.57	12.67	12.92
poetry model	9.65	11.57	9.89	10.00	10.86	10.40	11.25

Table 1: Articulation rate calculated as the average number of phones per second, excluding pauses.

poem	1	2	3	4	5	6	7
human	0.30	0.19	0.19	0.26	0.23	0.23	0.14
baseline	0.11	0.10	0.08	0.10	0.11	0.11	0.10
poetry model	0.19	0.21	0.19	0.22	0.26	0.27	0.13

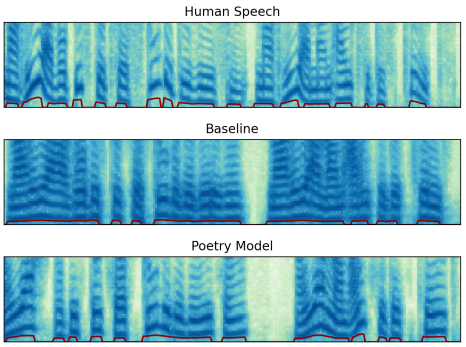
Table 2: Average duration of pauses in seconds, excluding silence at the beginning and end of audio.

our poetry model. We deliberately chose some poems for testing that contained interesting stylistic phenomena, which we will discuss in Section 5. In detail, our test set consisted of the poems [1] *Das Wunderhorn* (*The Magic Horn*), [2] *Großmutter Schlangenköchin* (*Snake Cooker Granny*), [3] *Lied des Verfolgten im Thurm* (*The Song of the Persecuted in the Tower*), [4] *Schön Dännerl* (*Cute Dännerl*), [5] *Aus dem Odenwald* (*From Odenwald*), [6] *Knecht, Magd, Ochs, Esel und alles, was mein ist* (*Servant, Maid, Ox, Donkey, and All Mine*) and [7] *Kinder-Konzert* (*Children's Concert*).

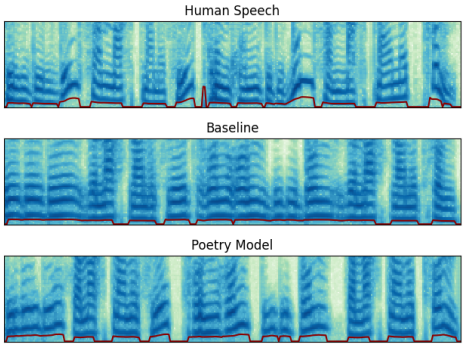
4.3 Experiments and Results

To investigate the difference in realisation of poetry between our baseline model and the poetry model, we compared prosodic features of the synthesised speech produced by both models with the human reference audio. We measured articulation rate and duration of pauses aggregated for each poem and took a closer look at pitch progression.

Table 1 presents the mean articulation rate for each poem. We calculated articulation rate by dividing the total number of phones in a poem by the sum of durations of all phones, excluding durations that correspond to pauses or silence at the beginning and end of each audio file. First, we noticed the rather slow speaking rate of the human speaker. For comparison, Trouvain et al. 2001 report a mean articulation rate of 13.06 phones per seconds in German read speech with a standard deviation of 2.03 calculated on the KielCorpus (Kohler et al. 1995) consisting of prose data. The results show that, although at the lower limit, the articulation rate of speech produced by the baseline model (which was exclusively trained on prose data) is still within the expected range for prose texts. In comparison, speech produced by our poetry model exhibits a slower articulation rate closer to the human reference, indicating that the poetry model adapted to the articulation rate of the human speaker to a large extent during the second finetuning stage.



(a) Comparison of spectrograms of the text *Maria, wo bist du zur Stube gewesen? Maria, mein einziges Kind!* (*Maria, where were you brought up? Maria, my only child!*)
(Our translation)



(b) Comparison of spectrograms of the text *Dill dill dill, so macht meine Flöt, Rum rum, bidi bum, so macht meine Trumm.* (*Dill dill dill, so goes my flute, rum rum, bidi bum, so goes my drum.*)
(Our translation)

Figure 1: Two exemplary comparisons of spectrograms produced by the baseline and poetry model, along with the human reference. The F_0 contour is shown in red.

Since pauses are an important means to control speech tempo (Trouvain et al. 2001), we further examined the average duration of pauses in seconds, shown in Table 2. Since synthetic audio was produced for each two-verse chunk separately, information about pauses between such chunks was lost. We therefore excluded all silences at the beginning and end of each audio both in synthetic and human audio samples in our calculations. We found that the values calculated in the poetry model are in close proximity to that of the human speaker, deviating in one or the other direction by only a few milliseconds. In contrast, the baseline model consistently produces shorter pauses than the reference with a difference of more than 11ms on average. These results indicate that apart from a slower articulation rate, the poetry model also adapted to the rather long pauses in our recordings.

Finally, we present an exemplary comparison of spectrograms generated by both TTS models and human reference in Figure 1. Figure 1a shows spectrograms corresponding to the first two verses of the poem *Großmutter Schlangenköchin*. Figure 1b displays spectrograms of two verses taken from the fourth stanza of *Kinder-Konzert* representing an example of onomatopoeia mimicking the sound of a flute and a drum. For both examples, we can verify our results from Table 2, which show that the poetry model tends to produce longer pauses between individual words than the baseline. Moreover, we observe that the F_0 contour indicated in red looks quite flat and monotonous in case of the baseline model. In contrast, the poetry model seems to produce a more vivid pitch that is closer to the human speaker.

In summary, we find that our poetry model trained on recordings of *The Boy's Magic Horn* adapts quite well to the speaking style of our speaker in terms of articulation rate and pitch. This is also in line with our first impression when listening to the generated audio samples.

5 Discussion

In the previous section, we described how we trained a TTS model on our data from *The Boy's Magic Horn* and investigated how prosodic properties are realised compared to a model trained on prose data. While our TTS model was able to learn some characteristics of our speaker's recitation style, other qualities cannot be produced by the model. In particular, the capacities of our model are very limited with regard to creative freedom and artistic expression. *Knecht, Magd, Ochs, Esel und alles, was mein ist* is an example: In this poem, a poor woman is given different animals, going from a flea to a cow, even servants, a husband, a child and a house, and she has to name each of them. In each stanza, her belongings are increased by one new item and its proper noun. In this way, each stanza becomes one line longer. In the fourteenth and last stanza, the account of property includes 14 lines, or rather 61 words. Our speaker took up the challenge of reciting this long sentence in one and the same breath, which he succeeded in doing. The recitation was thus fast and low in articulation. We would like to discuss the fourth stanza, where the woman is given a goat after she has already received a goose, a duck and a chicken in the stanzas before:

> Als ich ein armes Weib war,
> Zog ich über den Rhein,
> Bescheert mir Gott ein Zickelein,
> War ich ein reiches Weib,
> Gieng ich über die Wiese,
> Fragten alle Leut,
> Wie mein Zickelein hiese,
> Klipperbein heißt mein armes Zickelein,
> Wackelschwänzlein heißt mein Gänslein,
> Entequentlein heißt mein Entlein,
> Bibberlein heißt mein armes Hünelein.[2]

In the early stanzas of this poem, the human speaker pronounced each of the repeating verses clearly and at a relaxed pace. Later on, he quickened the speed of speech both to give expression to the repetitive structure of the poem and to master the sporting challenge of speaking the increasing list of belongings in one breath. In contrast, our TTS models lack information about such patterns related to context and produce each repetition at exactly the same tempo. As a result, the synthesised recitation of this poem might sound boring to human ears.

As a second example, we looked at the onomatopoetic poem *Kinder-Konzert* mentioned above. Here, the sound of various musical instruments is imitated. While the individual sound of each instrument is primarily predetermined by the textual form through the choice of vowels and consonants, the speaker conveyed the overall musical character of the poem mainly through speech rhythm. The TTS models realise the sounds of the instruments more as a kind

[2] When I was a poor woman, / I crossed the Rhine, / God gave me a goat, / I was a rich woman, / I went across the meadow, / All the people asked, / What my goat's name was, / Klipperbein is the name of my poor little goat, / Wackelschwänzlein is the name of my goose, / Entequentlein is the name of my duckling, / Bibberlein is the name of my poor little chicken. (Our translation.)

of enumeration, completely unrelated to the setting of the poem. The poetry model dealt with this challenge better than the baseline, since it grasped some metric patterns from the training data. This gave the resulting speech a more or less rhythmic structure, although unevenly matching the context.

While TTS models lack a sense of creativity, the previously discussed phenomena can be reproduced through prosody cloning, i.e. exactly replicating the original prosody of a recitation instead of predicting prosody in an unsupervised manner, or manual manipulation of the synthetic speech, as suggested in Koch et al. 2022. Apart from this, we found that the challenges of pre-processing the data (see Section 4.2) severely affected speech quality. Even after excluding a large number of samples from the training data, the resulting training set was still very unclean, i.e. it still contained a considerable number of mismatches between the phonetic transcription and the audio due to phonemiser errors or failures in automatic audio segmentation. As a result, the speech produced by the poetry model sounded blurred. Further, we encountered several cases of mispronunciation in our test data where the phonemiser failed to produce the correct phone sequence. For future work, it is crucial to develop tools that incorporate time and genre-specific linguistic knowledge and thus can deal with the challenges posed by archaic German.

6 Conclusion

We introduced a multi-modal corpus comprising text and speech recordings of the *Boy's Magic Horn* collection by Achim von Arnim and Clemens Brentano, embedded within the »textklang« project. The speech recordings were made in cooperation with Toni Bernhart as speaker and are aligned with the corresponding texts. We enriched this corpus with various annotations of metadata for the purpose of analysis and visualisation by means of Keshif. We demonstrated using examples how this corpus can be used in the field of speech technology by training a text-to-speech system on these data and exploring which prosodic properties can be learned from it. *The Boy's Magic Horn* is a large, canonical, multi-dimensional collection of romantic (lyric) poetry that can serve as a basis for different research questions. The synopsis of text and audio files and metadata allows us to study prosodic patterns in text and sonic realisation. We believe that this corpus is a valuable resource for various applications ranging from literary studies to natural language processing and speech technology.

Acknowledgments

This research was supported by funding from the German Ministry of Education and Research (BMBF) for the »textklang« project.

References

Arnim, Achim von and Clemens Brentano (1806). *Des Knaben Wunderhorn. Alte deutsche Lieder*. Heidelberg and Frankfurt am Main: Mohr und Zimmer. URL: https://www.deutschestextarchiv.de/arnim_wunderhorn01_1806.

Arnim, Achim von and Clemens Brentano (1808a). *Des Knaben Wunderhorn. Alte deutsche Lieder. Dritter Band*. Heidelberg: Mohr und Zimmer. URL: https://www.deutschestextarchiv.de/arnim_wunderhorn03_1808.

Arnim, Achim von and Clemens Brentano (1808b). *Des Knaben Wunderhorn. Alte deutsche Lieder. Zweyter Band*. Heidelberg: Mohr und Zimmer. URL: https://www.deutschestextarchiv.de/arnim_wunderhorn02_1808.

Benischek, Caren (2008). "Liste der Vertonungen in Auswahl". In: *Von Volkston und Romantik. Des Knaben Wunderhorn in der Musik*. Ed. by Antje Tumat. Heidelberg: Winter, pp. 189–216.

Bernard, Mathieu and Hadrien Titeux (2021). "Phonemizer: Text to Phones Transcription for Multiple Languages in Python". In: *Journal of Open Source Software* 6.68, p. 3958. DOI: 10.21105/joss.03958. URL: https://doi.org/10.21105/joss.03958.

Brentano, Clemens (1975–1978). *Des Knaben Wunderhorn. Ed. by Heinz Rölleke. Frankfurter Brentano-Ausgabe, vols. 6–9*. Stuttgart: Kohlhammer.

Deutsches Textarchiv (2023). *Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache*. Ed. by Berlin-Brandenburgische Akademie der Wissenschaften. URL: https://www.deutschestextarchiv.de/.

Eschenbach, Gunilla and Sandra Richter (2020). "Sounds in contact. The American bird sounds of a German-American worker poet and new methods of comparing literary sounds". In: *Canadian Review of Comparative Literature* 47.4, pp. 449–462. DOI: 0319–051x/20/47.4/449.

Grimm, Jacob and Wilhelm Grimm (1812). *Kinder- und Haus-Märchen. Gesammelt durch die Brüder Grimm*. Berlin: Realschulbuchhandlung. URL: https://www.deutschestextarchiv.de/grimm_maerchen01_1812.

Grimm, Jacob and Wilhelm Grimm (1815). *Kinder- und Haus-Märchen. Gesammelt durch die Brüder Grimm. Zweiter Band*. Berlin: Realschulbuchhandlung. URL: https://www.deutschestextarchiv.de/grimm_maerchen02_1815.

Gulati, Anmol, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. (2020). "Conformer: Convolution-augmented Transformer for Speech Recognition". In: *Interspeech*, pp. 5036–5040.

Herder, Johann Gottfried (1773). *Von Deutscher Art und Kunst. Einige fliegende Blätter*. Hamburg: Bode. URL: https://www.deutschestextarchiv.de/herder_artundkunst_1773.

Koch, Julia, Florian Lux, Nadja Schauffler, Toni Bernhart, Felix Dieterle, Jonas Kuhn, Sandra Richter, Gabriel Viehhauser, and Ngoc Thang Vu (2022). "PoeticTTS - Controllable Poetry Reading for Literary Studies". In: *Proc. Interspeech 2022*, pp. 1223–1227. DOI: 10.21437/Interspeech.2022-10841.

Kohler, Klaus J., Matthias Pätzold, and Adrian Simpson (1995). "From scenario to segment. The controlled elicitation, transcription, segmentation and labelling of spontaneous speech". en. In: *Arbeitsberichte des Instituts für Phonetik und*

digitale Sprachverarbeitung 29. ISSN: 0172-8156. URL: https://macau.uni-kiel.de/receive/publ_mods_00002186.

Kong, Jungil, Jaehyeon Kim, and Jaekyoung Bae (2020). "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis". In: *NeurIPS* 33.

Łańcucki, Adrian (2021). "FastPitch: Parallel text-to-speech with pitch prediction". In: *ICASSP*. IEEE, pp. 6588–6592.

Lux, Florian, Julia Koch, and Ngoc Thang Vu (2022). "Low-Resource Multilingual and Zero-Shot Multispeaker TTS". In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*.

Lux, Florian, Julia Koch, and Ngoc Thang Vu (2023). "Exact Prosody Cloning in Zero-Shot Multispeaker Text-to-Speech". In: *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 962–969. DOI: 10.1109/SLT54892.2023.10022433.

Lux, Florian and Ngoc Thang Vu (2022). "Language-Agnostic Meta-Learning for Low-Resource Text-to-Speech with Articulatory Features". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 6858–6868.

Müller, Thorsten and Dominik Kreutz (2021). *Thorsten - Open German Voice (Neutral) Dataset*. https://doi.org/10.5281/zenodo.5525342.

Oord, Aäron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alexander Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu (2016). "WaveNet: A Generative Model for Raw Audio". In: *Arxiv*. URL: https://arxiv.org/abs/1609.03499.

Puchtler, Pascal, Johannes Wirth, and René Peinl (2021). "Hui-audio-corpus-german: A high quality tts dataset". In: *German Conference on Artificial Intelligence (Künstliche Intelligenz)*. Springer, pp. 204–216.

Ren, Yi, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu (2020). "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech". In: *ICLR*.

Ren, Yi, Jinglin Liu, and Zhou Zhao (2021). "PortaSpeech: Portable and High-Quality Generative Text-to-Speech". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., pp. 13963–13974. URL: https://proceedings.neurips.cc/paper/2021/file/748d6b6ed8e13f857ceaa6cfbdca14b8-Paper.pdf.

Richter, Sandra, Toni Bernhart, Felix Dieterle, Gabriel Viehhauser, Gunilla Eschenbach, Jonas Kuhn, Nadja Schauffler, André Blessing, Markus Gärtner, Kerstin Jung, Nora Ketschik, Anna Kinder, Julia Koch, Thang Vu, and Andreas Kozlik (2023). "Der Klang der Lyrik. Zur Konzeptualisierung von Sprecher und Stimme, auch für die computationelle Analyse". In: *POEMA*, pp. 39–51. DOI: https://doi.org/10.38072/2751-9821/p4.

Schauffler, Nadja, Toni Bernhart, André Blessing, Gunilla Eschenbach, Markus Gärtner, Kerstin Jung, Anna Kinder, Julia Koch, Sandra Richter, Gabriel Viehhauser, Ngoc Thang Vu, Lorenz Weseman, and Jonas Kuhn (2022). "»textklang« Towards a Multi-Modal Exploration Platform for German Poetry". In: *Proceedings of the Language Resources and Evaluation Conference*. Marseille,

France: European Language Resources Association, pp. 5345–5355. URL: https://aclanthology.org/2022.lrec-1.572.

Shen, Jonathan, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, et al. (2018). "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions". In: *ICASSP*, pp. 4779–4783.

Skerry-Ryan, RJ, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous (2018). "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron". In: *ICML*. PMLR, pp. 4693–4702.

Trouvain, Jürgen, Jacques Koreman, Attilio Erriquez, and Bettina Braun (2001). "Articulation rate measures and their relation tophone classification in spontaneous and read German speech". In: *Proc. Workshop on Adaptation Methods for Speech Recognition*, pp. 155–158.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Yalçın, M. A., N. Elmqvist, and B. B. Bederson (2016). "Keshif: Out-of-the-box visual and interactive data exploration environment". In: *Proceedings of IEEE VIS 2016 Workshop on Visualization in Practice: Open Source Visualization and Visual Analytics Software*.