The world's first South Sámi TTS – a revitalisation effort of an endangered language by reviving a legacy voice

Katri Hiovain-Asikainen[1,2], Thomas B. Kjærstad[1], Maja Lisa Kappfjell[1], Sjur N. Moshagen[1]
[1]UiT The Arctic University of Norway, [2]University of Helsinki, **Correspondence:** firstname.lastname@uit.no

Abstract

South Sámi (ISO 639: SMA) is a severely endangered language spoken by the South Sámi people in Norway and Sweden. Estimates of the number of speakers vary from 500 to 600. Recent advances in speech technology and the general increase in popularity of spoken language and audio content have facilitated the development of modern speech technology tools also for minority languages, such as the Sámi languages.

The current paper documents the development process of the world's first South Sámi text-to-speech (TTS) system, using only digitized archive materials from 1989–1993 as the training material. To reach an end-user suitable quality of the TTS, we have used a neural, end-to-end approach with a rule-based text processing module. The aim of our project is to contribute to the language revitalization by offering tools for language users to use spoken language in new contexts. Since the modern written standard of South Sámi was established as late as in 1978, the rise of speech technology might encourage language use even for people who are not accustomed to the written standard.

1 Introduction

The traditional speaking area of South Sámi is in the central regions in Norway and Sweden, shown in Figures 1 and 2. The estimated amount of South Sámi speakers range from 500 to 600[1]. Steps are being taken for South Sámi language revitalization in order to preserve and strengthen the speaker community and transfer the language to new generations. This has been most successful in the Snåsa municipality in Norway (see Figure 2), while an increasing number of municipalities are joining the Sámi language management area.

Even though South Sámi is established as a literary language, there are surprisingly few language revitalization initiatives. Some remarkable efforts for revitalizing the languages are currently, for example, translation of children's books, development of South Sámi as a professional language at university levels, as well as using South Sámi as a media language within NRK Sápmi, the Norwegian broadcasting company[2,3]. Despite this, the education system seems to be struggling greatly with the extensive loss of students from Sámi education as shown in the article *The Sámi leak in primary school* (Øystein Vangsnes, 2021).

Because of the long and intense contacts with neighboring Scandinavian languages, practically all adult speakers of South Sámi are bilingual in Sámi and Norwegian/Swedish. Even though South Sámi has had a written standard since 1978, not all speakers have had the opportunity to receive full education in the language and might not be comfortable reading texts in South Sámi. Due to limited environments for hearing and practicing the language, there is a strong need for tools that demonstrate and guide South Sámi pronunciation. Speech technology can help overcome these barriers by providing support for spoken language, as in pronunciation, intonation, and stress. In addition to this, there is a high demand for speech technology tools especially in (special) education, for being able to integrate the learning of spoken language into the language learning materials more easily, as well as aiding people with dyslexia or vision impairment.

The Divvun group at UiT The Arctic University of Norway develops language technology tools for indigenous and minority languages within the GiellaLT infrastructure[4] (Pirinen et al., 2023). The first Sámi TTS for North Sámi was

[1] https://snl.no/sørsamisk

[2] https://www.nord.no/aktuelt/historiske-masterstudentar-i-sorsamisk

[3] https://nynorsksenteret.no/blogg/sorsamisk-sprakutvikling-gjennom-barneboker

[4] https://giellalt.github.io/

Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages, pages 12–21

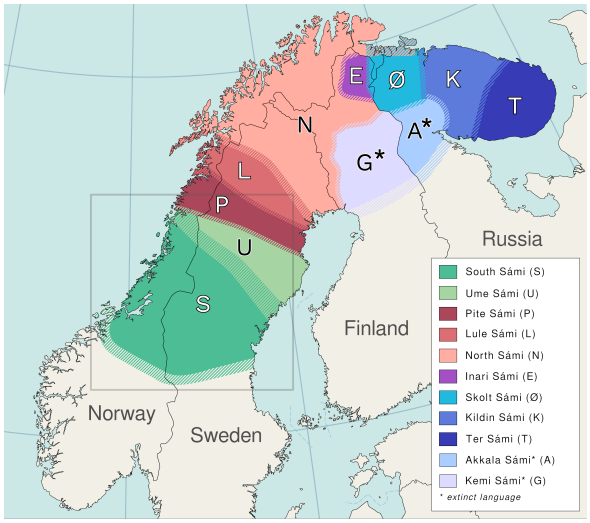

Figure 1: Map of traditional speaking areas of the Sámi languages. Adapted from https://snl.no/samisk with permission from the Great Norwegian Encyclopedia and the map author, Mikkel Berg-Nordlie. The grey box is roughly depicting the closer South Sámi area shown in Figure 2.

Figure 2: Map of the traditional South Sámi dialects, showing also the most prominent town names and locations. (Rantanen et al., 2022).

created in 2015 in collaboration with Acapela. In recent years, Divvun has updated the North Sámi TTS and released new Lule Sámi TTS tools (Hiovain-Asikainen and Moshagen, 2022; Hiovain-Asikainen and De la Rosa, 2023; Hiovain-Asikainen and Suni, 2025). All Divvun tools are open source, freely available through *Divvun Manager*[5], and integrated into GiellaLT for continued maintenance and updates.

2 Related work

Modern users increasingly expect near-human-quality TTS, as seen in high-resource languages like English. Earlier systems depended on relatively large single-speaker datasets, such as LJ Speech (Ito and Johnson, 2017), but these lacked speaker and linguistic diversity. Recent progress has shifted toward massive multilingual and multi-speaker corpora that improve generalization and adaptability. The Emilia dataset (He et al., 2024), for example, includes 150,000 hours of speech across five major languages, while models like VALL-E 2 (Chen et al., 2024), trained on datasets such as the 50,000 hour LibriHeavy corpus (Kang et al., 2024), push synthesis quality even further. VALL-E 2 claims to reach human parity in naturalness and expressiveness, underscoring the growing role of large, diverse datasets in state-of-the-art TTS research.

While much of the recent TTS work depends on massive datasets, another line of research focuses on models that perform well with minimal data – crucial for languages like South Sámi. Large-scale approaches do not address the core low-resource challenge: the absence of sizable datasets or suitable pre-trained models. For successful adoption in very low-resource settings, synthesized speech must remain both intelligible and pleasant. We therefore selected a non-autoregressive Transformer-based model capable of high-quality output from small datasets ($\leq$10 hours). Prior work supports this choice: Xu et al. (2020) show that 1.3 hours of Lithuanian speech was enough for a Transformer-based TTS system to produce intelligible speech (Li et al., 2019), and in Võro TTS (Rätsep and Fishel, 2023), transfer learning from Estonian did not outperform training directly on 1.5 hours of Võro-only data. These studies indicate that architectures like FastPitch (Łańcucki, 2021) may require as little as 1–2 hours of speech, with limited benefit from related-language data. Because South Sámi differs substantially from other Sámi languages in both linguistics and orthography, transfer learning was not pursued here.

Due to the scarcity of South Sámi data and resources, we utilized existing but non-ideal materials, inspired by Cooper (2019), which examined various speech sources for low-resource TTS. The

[5] https://borealium.org/en/category/voices/

study found that training with a mix of high- and low-quality data, then adapting toward high-quality subsets, improved intelligibility. This approach is particularly useful for low-resource languages, as it enables combining multiple smaller datasets, as done in this work.

In summary, despite recent near-human TTS advances, we selected FastPitch (Łańcucki, 2021) for the South Sámi project due to its low data requirements and straightforward training process. State-of-the-art models like VALL-E offer impressive zero-shot capabilities, but require extensive datasets and substantial computational resources, making them unsuitable for low-resource languages. FastPitch instead provides a transparent, controllable and practical solution to develop the South Sámi TTS, similarly as we have done successfully for Lule and North Sámi before.

3 Methodology

In this section, we describe our data and methodology for developing TTS for South Sámi. We present our approach in acquiring materials and dealing with very low-quality data by performing a series of audio restoration procedures, and by taking the data quality into account in the training phase of the TTS model. Finally, we report our training setup and choice of vocoder specifically for this project.

3.1 Data description

Anna Jacobsen (Sámi name *Jaahkenelkien Aanna*), 30th October 1924 – 2nd April 2004, was a leading advocate for the South Sámi language and culture in Norway (Gaski and Kappfjell, 2006). Born into a reindeer-herding family, Jacobsen grew up speaking South Sámi and later became the first person formally examined in the language. Her work with the South Sámi langugage covered language teaching, translating, writing teaching materials, organizing language groups, being a language consultant for the Sámi Education Council and establishing the Sijti Jarnge Cultural Center and a South Sámi theater. She received honorary awards for her work.

3.1.1 Ethical considerations

Our South Sámi female voice *Aanna* is based on archival recordings of Anna Jacobsen. The recordings are used in this project with the acknowledgment and written consent of her two descendants. The descendants of the original speaker, Anna Jacobsen, were contacted by email and subsequently met through an online meeting to discuss the project and its implications. They gave their informed consent to (1) open-source publication of the recordings and transcripts, (2) their use in the development of a South Sámi TTS system, and (3) waiver of any royalties. The potential risks of voice misuse inherent in TTS technology were discussed, and the descendants expressed general awareness and written support for the project. They do not retain ownership of the resulting models, nor the ability to revoke consent after publication. This work was conducted by the Divvun group that operates as part of the Sámi community under the administration of the Sámi Parliament, and adheres to the FAIR and CARE principles as described in (Moshagen et al., 2024) for data management and ethical research involving Indigenous communities.

3.1.2 The materials and transcribing process

The archival recordings of Anna Jacobsen were sourced from the Norwegian national broadcaster NRK and several audiobooks, and were digitally restored, enhanced, and transcribed by the Divvun group at UiT in collaboration with the Sámi Archives and the National Archives of Norway. Recorded between 1989 and 1993, the material spans multiple genres, including news and documentary broadcasts, biblical readings, fairy tales, and spontaneous autobiographical storytelling.

Developing a TTS dataset required text that accurately matched each recording. Many of Jacobsen's broadcasts already existed in written form, as they were later published in the anthology series Don jih daan bijre I–III (Jacobsen, 1997, 1998, 2000). Her biblical recordings were aligned with the South Sámi translation she produced together with Bierna Leine Bientie (Jacobsen and Bientie, 1993), which was scanned and OCR-processed for the project. Additional usable material came from her language-learning book Goltelidh jih soptsestidh (Jacobsen, 1993) and its accompanying audio cassettes.

Where written texts existed, the recordings were reviewed in detail and the texts were adjusted to reflect the spoken versions, which often differed slightly from the published forms. For recordings without prior transcriptions, full manual transcription was required. Three project members—two native speakers and one highly proficient non-native speaker—carried out this work during 2023–2024.

Processing roughly ten hours of audio resulted in about one hundred hours of transcription, equivalent to 2.5–3 weeks of full-time work for one per-

son. This workload is realistic for most endangered language contexts, especially since a significant portion of the material could be aligned rather than transcribed from scratch. Although labor-intensive, manual transcription remains more feasible than automated methods (such as ASR), which require large, high-quality datasets that Indigenous languages typically lack. Human transcription is therefore not a bottleneck but a practical and scalable strategy for building high-quality TTS resources in low-resource settings.

During transcription, segments with very poor audio quality or containing speakers other than Anna Jacobsen were excluded from the final TTS corpus.

3.2 Data processing

After the transcribing process, all texts were once more proof-read and all audio was cleaned of any unusable parts or noise. Then, the material was force-aligned to automatically find sentence boundaries from the audio, using a WebMAUS[6] pipeline without ASR (G2P ⇒ MAUS ⇒ Subtitle, see Kisler et al. (2017); Schiel (1999)), retaining the original text formatting and punctuation. There are no Sámi models on WebMAUS, so we used their Finnish (related language) model. The automatically aligned sentence timestamps from WebMAUS were then manually checked and used to split the data into sentences. Python scripts by the first author, utilizing the TextGridTools toolkit[7] (Buschmeier and Wlodarczak, 2013) were used to save each sentence to an individual sound file with a corresponding text transcript. After splitting the data, the net duration of the entire dataset was 10.5 hours, with 4670 individual sentences in total.

The next step in our pre-processing pipeline was to enhance the audio. Understandably, as an archive material, the audio quality of our material was not as high as normally expected from any generic text-to-speech projects. Our material was collected from different cassettes and CDs, all with varying recording conditions and probably with different digitizing equipment as well. The sound files were enhanced and de-noised using the freely available *Resemble-enhance*[8] with default settings and parameters. Resemble Enhance is an AI-powered tool that aims to improve the overall quality of speech by performing denoising and enhancement. It consists of two modules: a de-noiser, which separates speech from noisy audio, and an enhancer, which further boosts the perceptual audio quality by restoring audio distortions and extending the audio bandwidth. Running the denoiser and enhancer through the entire dataset substantially improved the audio quality. Next, we used a shell script utilizing *sox* and *svdemo* libraries to level normalize the data. Finally, the whole dataset was resampled at 22.05 kHz to be compatible with our TTS training setup.

3.3 Model configuration

Our model was trained using the FastPitch (Łańcucki, 2021) architecture with explicit duration and pitch prediction components. For our final model, we used a "multi-speaker" configuration for training by splitting the data into two subsets based on audio quality. Even though we performed audio enhancement to the entire dataset, the lower quality partition remained lower quality compared to the better quality part even though it was substantially improved in intelligibility compared to the original quality.

Out of the total 10.5 hours of data, 2 hours were manually labeled as "good quality" with speaker ID *1*, while the remaining 8.5 hours were labeled as lower quality with speaker ID *0*. This binary labeling reflects the intended use for TTS generation: ID *1* denotes recordings suitable for synthesis, whereas ID *0* denotes recordings that, while usable for training, were not intended for generation. We did not define additional quality categories or employ a continuous quality metric, as the primary goal was to distinguish between data that could or could not be directly used for synthesis.

The material was then shuffled in order and further divided into training, validation and test sets with a split of 4570/85/15, respectively. The test set (see Appendix B) was later used for an evaluation protocol. The dataset was processed using the standard FastPitch data preparation scripts to extract pitch, duration, and mel spectrograms from each utterance.

After defining the orthographic symbol set (see Appendix A) for South Sámi, the model was trained for 830 epochs and altogether 14K steps on the Saga supercomputer[9] at the Norwegian computing

[6] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/Pipeline

[7] https://textgridtools.readthedocs.io/en/stable/index.html

[8] https://github.com/resemble-ai/resemble-enhance

[9] https://documentation.sigma2.no/hpc_machines/saga.html

cluster *Sigma2*. We allocated 16 GB of GPU memory at Saga for the training process, and used an effective batch size of 256 and learning rate 0.1 as training parameters.

3.4 Model inference and text processing

To achieve good and consistent quality in the synthesis output, the speaker variant for our synthesis generation script was set to speaker ID *1*, acoustically similar to the better quality subset of our dataset. For inference, we used the UnivNet model (Jang et al., 2021) from NeMo collections as the vocoder. This vocoder was chosen because it produced an audibly better quality inference output than other neural vocoders, such as the widely used HiFi-GAN (Kong et al., 2020), compared with UnivNet in Jang et al. (2021). The most important reason for better quality output for UnivNet seems to be improved generalization across diverse speakers, domains, and unseen data. This is achieved through multi-resolution spectrogram loss by comparing synthesized audio to the target audio at multiple resolutions. As our TTS model was meant for publication in a TTS application, it was also important to choose a vocoder optimized for low-latency inference, such as UnivNet, making it suitable for real-time applications.

The text processing is done using the existing tools in the GiellaLT infrastructure. That infrastructure is based on HFST (see, e.g. Lindén et al. (2013)) and VislCG3 (see e.g. Karlsson (1990), Didriksen (2010)), making it possible to build advanced text processing tools for languages that lack large corpora. In the pipeline, the raw text is tokenized and morphologically analyzed by a lexical transducer, followed by several steps to enhance and disambiguate the tokenisation. The lexical transducer has many lexicalized compounds while at the same time being able to handle dynamic compounding. In the case of TTS, we prefer the dynamic compounds, because it allows us to normalize each part of the compound independently, for example in a word like “CD-player”. We also add valency information to help with the morphological disambiguation, which is coming next. The disambiguated analysis is then used as input for normalizing digits and abbreviations, and since these tokens are already disambiguated, only the relevant normalization is applied (if the normalization itself is ambiguous, it is possible to add further disambiguation at later steps). The normalized text is finally sent to the synthesizer. A flow diagram of the whole process is shown in Figure 3.

As illustrated in Figure 3, the text-processing pipeline consists of a multi-stage linguistic analysis and normalization workflow. The example sentence:

> `<Joekoen guhkiem, jis edtjebe jaehkedh dam 35 jaepien båeries nyjsenæjjam Kloemegistie.>`
>
> (“*A very long time, if we’re to believe the 35 year old woman from Glåmos.*”)

is first segmented and tokenized into lexical units:

```
<Joekoen> <guhkiem> <,> <jis>
<edtjebe> <jaehkedh> <dam>
<35> <jaepien> <båeries>
<nyjsenæjjam> <Kloemegistie> <.>
```

Each token is then processed by finite-state–based morphological analyzers, which produce lemmas and feature bundles (e.g., part of speech, case, number) and identify compound structures and orthographic variants.

Morphological disambiguation follows, using constraint grammar rules to select context-appropriate readings and assign syntactic functions. In the normalization stage, compounds are reconstructed, numerals are lexicalized (e.g., 35 → *golmeluhkievïjhte*), and canonical lemma forms are prepared for TTS input. The resulting output is mapped to phonological representation elements, yielding the normalized sequence:

```
<joekoen guhkiem, jis edtjebe
jaehkedh dam golmeluhkievïjhte
jaepien båeries
nyjsenæjjam kloemegistie>
```

which is then passed to the TTS model for generation.

It is possible to add a fall-back normalizer for handling unknown words, and there are stubs in place to differentiate the normalization of loan words from that of native words. For example, the orthographic "y" is typically pronounced differently in various South Sámi words, depending on its phonetic context (Magga and Magga, 2012). The present synthesis model has just learned the various pronunciations based on context in the training phase, and this works quite well. However, it still makes errors, and by differentiating the various pronunciations already in the text processing (and similarly in the training material), we should be able to get an even more correct pronunciation.

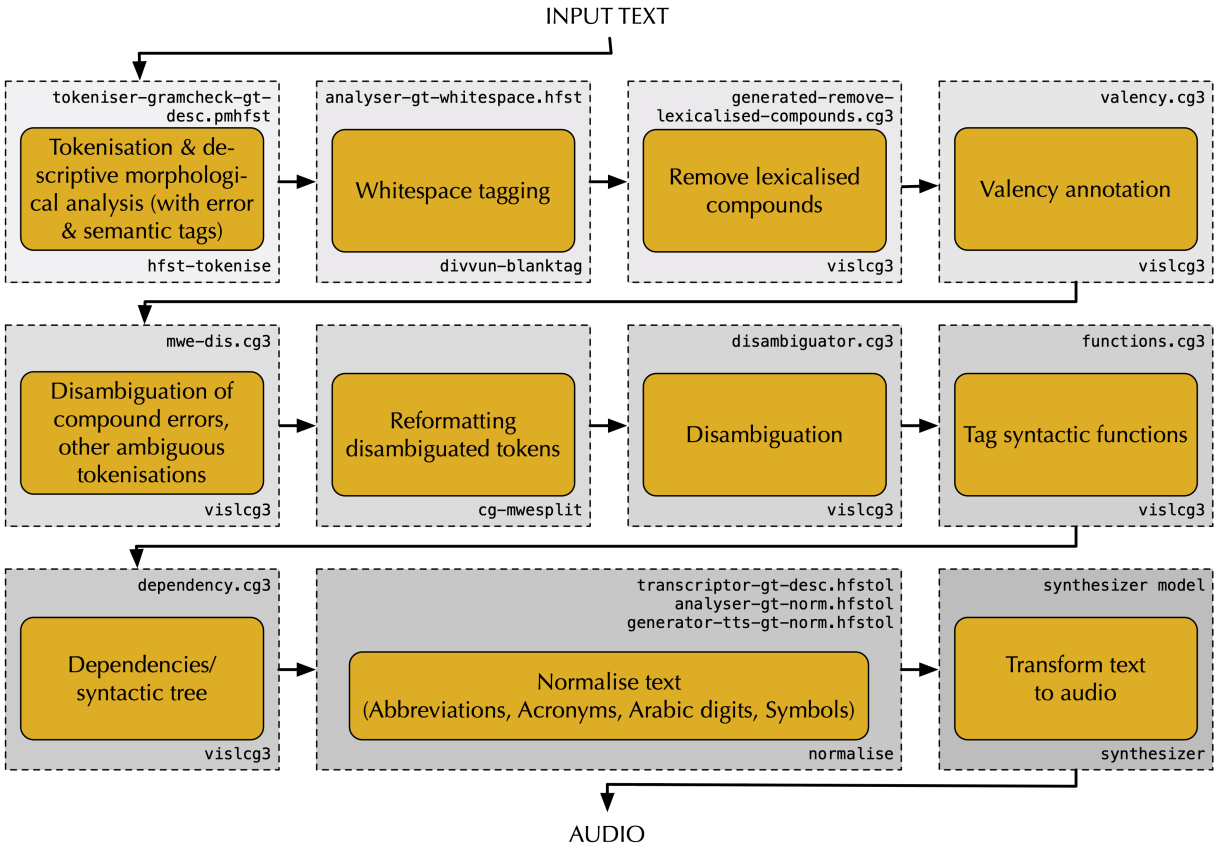

Figure 3: Flow diagram of the South Sámi TTS runtime system, including all text processing steps.

4 Results and Future Work

4.1 Technical implementation and publishing the TTS system

The resulting TTS model was compiled into TorchScript form, using the standard script from the FastPitch repository. This was done to create serializable and optimizable models from the originally trained PyTorch model. The JIT compiled model was further integrated into macOS and Windows operating systems (OSs), at the OS level, such that the voices appear as system voices within the limitations of the operating systems. In both systems, voices can be activated as screen-reader voices. In MacOS, the South Sámi voice can also be used within LibreOffice [10]. To some extent, voices can also be used for simpler read-aloud functionality for highlighted text, but there are several restrictions in both systems that limit the usability of this functionality. The OS integration is language independent and can include any voice for any language we release in the future. We plan to release support for using voices on Android and iOS in the near future.

[10] https://www.libreoffice.org/get-help/install-howto/macos/

The text processing infrastructure is both flexible and powerful, and allows us to do further syntactic and semantic analysis. This can also be used to add synthesis markup to guide, e. g., prosodic features in the synthesis, but this will have to be the target of future work. The modified FastPitch code used to create our South Sámi voice will be published on our GitHub page https://github.com/giellalt/speech-sma once the detailed documentation of the project is completed.

4.2 Initial evaluation of the TTS model

Established evaluation standards in speech technology are designed for majority languages with large speaker populations, and they cannot be readily applied to Indigenous and minority languages such as South Sámi. With few speakers, standard procedures become impractical and risk placing an undue burden on the community. This underscores a broader issue: evaluation in low-resource contexts must center community needs and capacities. While cross-linguistic comparability is valuable, methodological ideals may conflict with ethical and practical realities. In such cases, community well-being should take precedence, and transparent reporting of necessary deviations from standard

practice should be appropriate.

Research fatigue further complicates evaluation, as Sámi speakers are frequently asked to participate in studies despite the small population. Nonetheless, attempting evaluation remains important, and development of South Sámi TTS offers tangible benefits that help counterbalance concerns about extractive research. After the TTS release on 30 October 2024, an anonymous feedback survey was launched in October 2025, with initial responses indicating continued community engagement despite these constraints.

For initial testing of the model and the TTS inference, we created a set of 15 sentences, taken aside from the training data (see Appendix B), which we synthesized using the resulting model. We prepared corresponding 15 sentences both from the test set (ground-truth) and synthesized samples. After level-normalizing this evaluation set of altogether 30 samples, we created an online survey using Microsoft Forms with embedded audio files. For evaluation, we used a 5-point Likert scale (from 1 – Bad to 5 – Excellent) and asked the following questions: 1) How would you rate the general pronunciation and rhythm of the speech (**Pronunciation & Rhythm**) 2) How would you rate the pleasantness of the speech? (**Pleasantness**) 3) How would you rate the clarity of the speech (i. e. how easy it is to understand what is being said) (**Clarity**)?

Out of 15 test sentences, 12[11] paired audio items were used for the Wilcoxon signed-rank tests, with each pair consisting of a synthesized utterance and its corresponding ground-truth recording. For each item, the mean rating across evaluators was used in the comparison. The results show a statistically significant difference only for **Pleasantness**, where ground-truth recordings received higher ratings than the synthesized speech ($p \approx 0.0039$). For Clarity and Pronunciation & Rhythm, no significant differences were found between ground-truth and synthesized audio ($p \approx 0.33$ and 0.52, respectively).

In addition to the quantitative, statistical analysis, we asked native speakers to assess the quality of the South Sámi TTS in free form, providing us more qualitative, also valuable feedback on our synthetic voice. One native South Sámi speaker (outside of the authors of this paper) listened to these samples and commented any pronunciation mistakes or peculiarities in the prosody.

The evaluator provided three main comments: (1) Year numbers are sometimes omitted or read digit-by-digit on first attempt; more natural segmentation and rendering is recommended (e.g., *luhkiegaektsie-gaektsieluhkietjïjhtje* for '1887'). (2) Vowel duration could be slightly lengthened in some words, and a marginally slower speaking rate may improve naturalness and intelligibility. (3) Compound words are pronounced more accurately when written with hyphens.

These issues affect both text processing and, to a lesser degree, the speech synthesis model. Improved handling of numeric expressions, a modest reduction in speaking rate, and automatic insertion of hyphens for compounds should be straightforward adjustments. After implementing any changes, another evaluation—ideally with more participants—should be conducted to obtain more robust feedback.

4.3 First impressions and the expected impact of TTS in the South Sámi community

Overall, the development and integration of text-to-speech (TTS) technology for the South Sámi language has initially received positive reception from the community, with some media coverage as well[12]. TTS addresses a long-standing need for accessible language resources, particularly for self-study and language revitalization, by allowing South Sámi speakers to input text and hear it read aloud. This is particularly important for minority languages that lack such tools. The community expects TTS to support everyday language use, for example integration into smart home devices and public services could offer practical language support and increase accessibility for South Sámi speakers, also in Sámi administrative municipalities. TTS could also contribute to proper Universal Design implementations, making public information more inclusive.

In terms of specific use cases, the integration of TTS in digital dictionaries and language learning apps is highly anticipated. For instance, a young teacher has expressed interest in incorporating TTS into the popular flashcard program Anki[13]. This would allow learners to hear South Sámi words pronounced aloud while studying, combining visual and auditory learning modes to reinforce vocab-

[11] some evaluators did not evaluate all 15 sentences

[12] https://uit.no/nyheter/artikkel?p_document_id=864438

[13] https://apps.ankiweb.net/

ulary acquisition. This kind of integration could be a significant step forward in making language learning more efficient and accessible to a wider audience. Johan Sandberg McGuinne, a member of the community, also highlighted the importance of TTS as a teaching aid, stating, "I think it's good because I can use it as an aid in teaching and in elderly care." This sentiment underscores the diverse applications of TTS in enhancing both education and daily life for South Sámi speakers.

The integration of TTS technology into existing systems is thus expected to play a key role in the revitalization and sustainability of the South Sámi language, fostering a more inclusive linguistic environment for the community.

4.4 Future work

Our work in general includes strategies for ongoing improvement of the system as language usage evolves. The speech synthesis system separates text processing from the actual synthesis step such that the text processing is done by the existing, rule-based text processing components in the GiellaLT infrastructure, and the resulting plain-text strings are fed to the synthesis engine. The speech model in the synthesis engine does not need to be retrained or rebuilt for the whole system to improve - it is enough that improvements are made to the text processing pipeline, e.g., to improve handling of cases in numerals, or new names, words and terminology. As the text processing improves, so will the resulting generated speech. And as the source of the text processing is shared with all other tools built within the GiellaLT infrastructure, improvements in one area will automatically also improve speech synthesis.

Another aspect is that the team behind the Sámi speech synthesis projects are fully funded by the Norwegian Government, as part of long-term commitments to supporting the Sámi languages. So both financially and practically our work will continue for years and decades — we have already been doing this for twenty years — and thus maintenance and commitment to updates should be covered for the foreseeable future.

Our future work on South Sámi TTS technology involves expanding its capabilities by adding more voices, such as a male voice, and incorporating additional dialects/areal varieties of South Sámi. There is also potential to develop multilingual solutions, integrating Swedish and Norwegian material to the model for proper loan word pronunciation and to allow for code-switching in the TTS output. Augmenting the training dataset with recordings from additional native speakers would also help capture rarer words, exceptional pronunciations, loanwords, and names, further enriching the TTS system.

We also hope that in the future, our present work could inspire the revitalization of the smallest and most endangered Sámi languages like Ume and Pite Sámi, both of which have very few speakers. In the case of Ume Sámi, a significant collection of unpublished written materials exists but remains unavailable (Siegl, 2017). New language and speech technology tools could help make these resources more accessible and support the revitalization efforts. There is rising interest to still revitalize these languages, and speech technology could play an important role.

5 Conclusions

In this paper we presented a description of a novel TTS project, utilizing non-ideal, archived and digitized materials. We show that end-user suitable TTS quality is possible with limited materials and even with initially low quality audio. We suggest a way to process archive materials for TTS in an effective pipeline that is generalizable to other very low-resource languages as well. The resulting TTS voice, built from the archive audio materials by Anna Jacobsen has gotten very positive feedback from the present community, encouraging use of the language in new, spoken language contexts, contributing to the revitalization of the severely endangered language.

6 Acknowledgements

We thank Antti Suni and Sebastién Le Maguer from the University of Helsinki for their valuable advice. Our deepest gratitude goes to Ina Therés Andrea Sparrock for her exceptional contributions, particularly in transcribing the audio materials and presenting the project to the South Sámi community. We also warmly thank Halvard and Nils Johan Jacobsen, sons of Anna Jacobsen, for their cooperation and encouragement. Finally, we are grateful to Sijti Jarnge for hosting the launch of the world's first South Sámi TTS in Aarborte–Hattfjelldal during the celebration of Anna Jacobsen's 100th anniversary in October 2024.

References

Hendrik Buschmeier and Marcin Wlodarczak. 2013. Textgridtools: A textgrid processing and analysis toolkit for python. In *Tagungsband der 24. Konferenz zur elektronischen sprachsignalverarbeitung (ESSV 2013)*.

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. 2024. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*.

Erica Cooper. 2019. *Text-to-speech synthesis using found data for low-resource languages*. Columbia University.

Tino Didriksen. 2010. *Constraint Grammar Manual: 3rd version of the CG formalism variant*. GrammarSoft ApS, Denmark.

Harald Gaski and Lena Kappfjell. 2006. Saemien tjaelijh – 16 saemien tjiehpies- jïh faagelidteratuvren tjaelijh.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, and 1 others. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 885–890. IEEE.

Katri Hiovain-Asikainen and Javier De la Rosa. 2023. Developing tts and asr for lule and north sámi languages. In *Proc. 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, pages 48–52.

Katri Hiovain-Asikainen and Sjur Moshagen. 2022. Building open-source speech technology for low-resource minority languages with sámi as an example–tools, methods and experiments. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 169–175.

Katri Hiovain-Asikainen and Antti Suni. 2025. Does multilingual and multi-speaker modeling improve low-resource tts? experiments on sámi languages. In *Proc. SSW 2025*, pages 196–201.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. 2017. *URL https://keithito. com/LJ-Speech-Dataset*.

Anna Jacobsen. 1993. *Goltelidh jih soptsestidh*. Aarbortesne, Norway: Daasta berteme.

Anna Jacobsen. 1997. *Don jih daan bijre I*. Aarbortesne, Norway: Daasta berteme.

Anna Jacobsen. 1998. *Don jih daan bijre II*. Aarbortesne, Norway: Daasta berteme.

Anna Jacobsen. 2000. *Don jih daan bijre III*. Aarbortesne, Norway: Daasta berteme.

Anna Jacobsen and Bierna Leine Bientie. 1993. *Jupmelen rijhke lea gietskesne : Maarhkosen vaentjele*. Det Norske Bibelselskap.

Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. 2021. Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation. *arXiv preprint arXiv:2106.07889*.

Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. 2024. Libriheavy: A 50,000 hours asr corpus with punctuation casing and context. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10991–10995. IEEE.

Fred Karlsson. 1990. Constraint grammar as a framework for parsing unrestricted text. In *Proceedings of the 13th International Conference of Computational Linguistics*, volume 3, pages 168–173, Helsinki.

Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347.

Jinhan Kong, Jaehyeon Kim, and Jungil Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033.

Adrian Łańcucki. 2021. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592. IEEE.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. 2019. Neural speech synthesis with transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6706–6713.

Krister Lindén, Erik Axelson, Senka Drobac, Sam Hardwick, Juha Kuokkala, Jyrki Niemi, Tommi A Pirinen, and Miikka Silfverberg. 2013. Hfst—a system for creating nlp tools. In *International workshop on systems and frameworks for computational morphology*, pages 53–71. Springer.

Ole Henrik Magga and Laila Mattsson Magga. 2012. *Sørsamisk grammatikk*. Davvi Girji.

Sjur Nørstebø Moshagen, Lene Antonsen, Linda Wiechetek, and Trond Trosterud. 2024. Indigenous language technology in the age of machine learning. *Acta Borealia*, 41(2):102–116.

Flammie Pirinen, Sjur Moshagen, and Katri Hiovain-Asikainen. 2023. Giellalt—a stable infrastructure for nordic minority languages and beyond. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 643–649.

Timo Rantanen, Harri Tolvanen, Meeli Roose, Jussi Ylikoski, and Outi Vesakoski. 2022. Best practices for spatial language data harmonization, sharing and map creation—a case study of uralic. *Plos one*, 17(6):e0269648.

Liisa Rätsep and Mark Fishel. 2023. Neural text-to-speech synthesis for võro. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 723–727.

Florian Schiel. 1999. Automatic phonetic transcription of non-prompted speech. In *Proc. of the ICPhS*, pages 607–610.

Florian Siegl. 2017. Ume saami – the forgotten language. *Études finno-ougriennes*, (48).

Jin Xu, Xu Tan, Yi Ren, Tao Qin, Jian Li, Sheng Zhao, and Tie-Yan Liu. 2020. Lrspeech: Extremely low-resource speech synthesis and recognition. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2802–2812.

Øystein Vangsnes. 2021. Samiske tall forteller: Den samiske lekkasjen i grunnskolen.

A Appendix

South Sámi symbol set used in the FastPitch model training:

letters = '*AÆÅBCDEFGHIÏJKLMNOØÖPQRSTUVWXYZaæåbcdefghiïjklmnoøöpqrstuvwxyz*'

B Appendix

The set of 15 sentences for the evaluation test.

1. Nöörjen raedteste jis, Praansvaerien Novra jïh Jaahkenelkien Aanna.
2. Nåå jïh ibie leah gujht mijjieh daebpene åarjene annje man vækjele joejkedh, jïlhts aaj Frode lea mijjen båeries vuelide giehtjedamme, golteladteme jïh orrestahteme guktie nuerebh almetjh utnieh sijjide aaj sjeahta mijjen båeries vueliej mietie svihtjedh jïh aaj raakte joekestalledh.
3. Goh Jeesuse jïh dah luhkiegööktesh, jïh dah jeatjebh gïeh Jeesusinie ektesne, lin oktegh sjïdteme, dellie gihtjin dan jortesen bïjre.
4. Jïjtjh tjoeveribie jïermestalledh gåabph libie rahtjeminie dejnie mijjen Saemiedigkiebarkojne.
5. Pængstah leah aah orreme.
6. Nyjsenæjja dåerkeste juktie bælla, daajra guktie satnine sjïdti, jïh dellie båata Jeesusen uvte slienghkehte jïh gaajhkem sååarne.
7. Daan mijjen eatnemen lea stoerre aalkoealmetjefuelhkie, jïh mijjieh saemieh aaj dan fualhkan govlesovvibie.
8. Åadtjibie vaajtelidh dïhte maahta bueriedidh.
9. Eadtjohkelaakan kultuvrine barkedh.
10. Dïhte prievie båata daehtie moenehtseste maam Saemiedigkie lea tseegkeme, jïh mij edtja nænnoestimmie buektedh.
11. Såemies gaertenebuerie gujht vienth daan jaepien aaj tjoevere sïrvide joekoenlaakan bïepmedh guktie begkerellelåhkoe vaanen aerebi goh åadtjoeh leekedidh.
12. Tjaktje seenhte, naa jueskie lij gujht.
13. Nimhtie lea daan eatnemen mubpene bealesne, Orre Zeelantesne, Maorij luvhtie luvnie.
14. Guktie idtjidh dellie Jååhannesem jaehkieh?
15. Dïhte ovmurreds saernie noerhtede Saeltievaerien luvhtie båata.