Developing TTS and ASR for Lule and North Sámi languages

Katri Hiovain-Asikainen[1], *Javier de la Rosa*[2]

[1]UiT The Arctic University of Norway
[2]National Library of Norway

katri.hiovain-asikainen@uit.no, versae@nb.no

Abstract

Recent innovations in speech technology have made high quality TTS and ASR available even for extremely low-resource languages. This paper presents our updated work-in-progress report of an open-source speech technology project for two indigenous Sámi languages that are minority languages in Norway, Sweden and Finland.

At this stage, we have designed and collected text and speech corpora for training the first neural text-to-speech (TTS) for Lule Sámi. We will update the previous North Sámi TTS by collecting additional materials and by training a new model using state-of-the-art technologies.

We also describe our first experiments with developing ASR for North Sámi and discuss the next steps to be taken in our project.

Index Terms: speech synthesis, low resource languages, North Sámi, Lule Sámi, automatic speech recognition, TTS, ASR

1. Introduction

Modern neural TTS technologies have made high quality speech synthesis applications feasible for any language, even those with very limited resources. The biggest challenge remains the scarcity of training data and pretrained models. In this work, we describe the process of developing neural TTS models for two Sámi languages, Lule and North Sámi, using open-source technologies. Additionally, we discuss the steps we have taken to develop ASR models for North Sámi.

The Sámi are the only indigenous people in Europe and all 9 Sámi languages are considered to be endangered to different degrees [1]. Being part of the Uralic language family, the Sámi languages are related to Finnish and Estonian. The traditional territory of the Sámi is shown in Figure 1. The Sámi can be divided into nine separate languages: South, *Ume*, *Pite*, Lule, North, Inari, Skolt, Kildin, *Ter*[1]. Neighboring Sámi languages are mutually intelligible to some extent, but full comprehension requires additional training. Importantly, the orthographies between the languages differ greatly (see, e.g., [2]). Generally, all Sámi speakers are bi- or multilingual in Sámi and in one or more majority languages. Relevantly, while Lule Sámi is spoken in Norway and Sweden, North Sámi is spoken in three countries, Norway, Sweden and Finland. This creates remarkable variation between language users from the different countries (see [3]), and it needs to be addressed when developing speech and language technology tools. Importantly, South, Lule and North Sámi have an official status in Norway which means that these languages should be present in all official contexts along Norwegian.

[1]Languages marked with italic lack a standardized orthography.

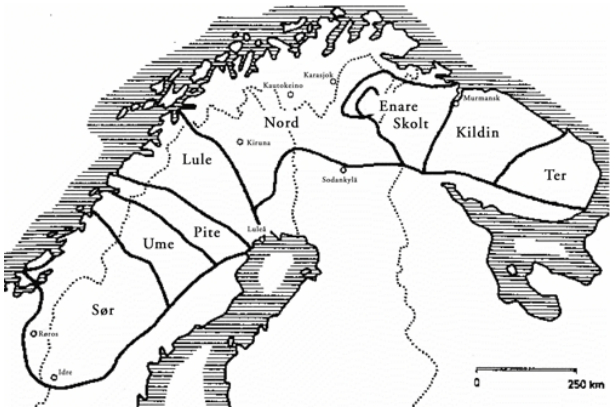

Figure 1: *A map showing the traditional speaking areas of nine Sámi languages. Akkala Sámi not shown on the map since it is considered extinct. Map: Wikimedia Commons, CC BY-SA 3.0.*

Neighboring languages Lule and North Sámi differ remarkably in terms of the amount of "language users". According to Ethnologue [4], North Sámi has by far the largest number of language users among the Sámi languages: 25,000 in all three countries where it is spoken. Lule Sámi has considerably fewer speakers: a total of 2000 in both countries it is spoken in. North Sámi, with a lesser degree of endangerment than Lule Sámi, has the highest number of language users among the Sámi languages, resulting in a greater availability of language resources and a wider variety of tools. An infrastructure of dictionaries, morphological analyzers, spell checkers and other language learning tools, have been maintained and developed since 2001 by the Divvun and Giellatekno groups (`https://divvun.no` and `https://giellatekno.uit.no/`). The current development of Sámi TTS and ASR will expand the selection of tools from text-based only into the spoken language realm as well.

TTS systems generate understandable speech from unfamiliar text in a specific language. The main goal of developing speech technology for indigenous languages is to ensure equal opportunities for language users in all communities. This which would enable the use of the Sámi languages in the same contexts than the majority languages, Norwegian, Swedish and Finnish in these cases. By facilitating accessibility for the Sámi languages in many new contexts, the development of speech technology also contributes to the preservation and revitalization of these languages. Moreover, speech technology tools are essential for individuals with specific needs, including language learners (see, e.g., [5]), people with dyslexia, visually impaired individuals, and also native language users not accustomed to

reading or writing in Sámi.

Given the needs of the language communities and our intent to support them in new ways, the goals of this paper are as follows:

1. To describe the ongoing work on developing speech technology corpora, models and tools for the Sámi languages as a part of the GiellaLT infrastructure.
2. To present the first impressions on the first ASR model for the North Sámi language.
3. To discuss evaluation methods of our models and other further steps and developments for our project.

2. Background and related work

In 2015, the first TTS tool for North Sámi was developed by Divvun and Acapela (`https://divvun.no/fi/tale/tale.html`). This tool was developed as closed-source, and neither the used framework nor the speech corpus are publicly accessible. Since support for certain operating systems has been recently discontinued, we are now working on a modern and open-source TTS system that will be openly available. Our aim is to improve the North Sámi TTS by augmenting the dataset with new material as well as leveraging state-of-the-art methodologies. The new system will be integrated into the larger GiellaLT infrastructure at `https://giellalt.github.io` and `https://github.com/divvun`, which will ensure regular maintenance and updates.

While in the early days of speech synthesis, entirely rule-based TTS frameworks such as Espeak required no corpora at all, modern data-driven systems are now often trained on big datasets like [6], which contains almost 24 hours of single-speaker data. Currently, however, state-of-the-art TTS approaches are being optimized to consume less and less data while still producing natural-sounding and intelligible voices. For instance, [7] showed that 1.3 hours of Lithuanian audio and text pairs were enough to train a Transformer based TTS model (see, e. g. [8]) that would meet the requirements for online deployment of a TTS system.

In [9], a TTS system was developed for Võro, a minority language spoken in Estonia. Although the majority language Estonian and Võro are closely related, they are not mutually intelligible and there are remarkable differences especially in the phonological systems of the languages. It was shown that it is possible to create a successful TTS model with only 1.5 hours of Võro speech data, using a Transformer based system similar to [10]. The mentioned works show that even very small datasets can be enough for the task if they are carefully designed and prepared. This direction of development in the TTS field benefits the extremely low-resource settings very well as speech data directly usable for TTS are not yet available for the Sámi languages and often suitable datasets need to be built from scratch as we did for Lule Sámi.

Previously ([11]), and before acquiring the training data for our project, we experimented with a few different older deep neural network based TTS methodologies. With one hour of low quality experimental single-speaker data, we trained a pilot version of a Lule Sámi voice using Ossian (`https://github.com/CSTR-Edinburgh/Ossian`, [12]) which was intelligible but would have not met the requirements of a modern speech technology user. By piloting the methods using small experimental data gave us better insight on the requirements for the speech corpus, i.e. the size and audio quality of the data as well as the technical requirements for training a speech model.

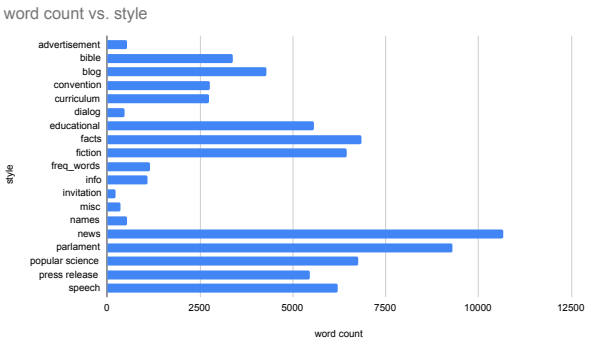

Figure 2: *The word counts per style of the Lule Sámi text corpus for TTS with altogether 74,737 words.*

In the following sections, we report the steps we have taken for the Sámi speech technology project so far using recent technologies and new datasets and explore potential avenues for future research.

3. Methodology

In this section, we describe building the first neural Lule Sámi TTS from scratch. Additionally, we introduce the current situation of the North Sámi TTS which we are currently updating with more data and using a different TTS framework.

After our experiments with various other TTS methodologies, we will focus on the FastPitch (*parallel TTS with pitch prediction*) [10] framework as the main approach for our Sámi voices. In [10], Tacotron 2 and FastPitch models with WaveGlow vocoder [13] were compared by 60 participants, and the Mean Opinion Score (MOS) evaluation test result of the models was higher for the FastPitch model (MOS: 4.080) than for the Tacotron 2 model (MOS: 3.946). There are many advantages in using FastPitch instead of Tacotron 2: as the model is conditioned on fundamental frequency estimated for every input symbol, making the pitch contour of the output very natural-sounding. Importantly, in our experience, the training process for FastPitch was remarkably faster and lighter than with the Tacotron 2 approach we experimented with (see [11]).

3.1. Datasets

3.1.1. Lule Sámi

As focus of the present project is on open-source methodologies, it was important to build a collection of open-source texts with a CC-BY licence in order to make our corpus publicly available later on. To build our new Lule Sámi TTS text corpus, we reused a part of the gold corpus (`https://gtsvn.uit.no/freecorpus/goldstandard/converted/smj/`), developed in 2013 within the GiellaLT community. Furthermore, we collected additional texts of various styles we knew to be well written and proofread. The resulting corpus for TTS contains over 74,000 words and consists of various text styles as shown in Figure 2.

The question of *data efficiency* in TTS has been discussed in [14], where the authors evaluated the amount of data required by the Tacotron 2 [15] TTS to produce good quality outputs. It was shown that if the training data is carefully checked for and

constructed as to present all graphemes, essential sounds and sound combinations in a language, the data requirement can be significantly lowered. Accordingly, we checked that our text corpus covered all important phonological contrasts and sound combinations by calculating frequencies of trigrams in our corpus. Because consonant gradation is a very prominent part of Lule and North Sámi morphophonology, especially inflection patterns of most parts-of-speech in the languages, we also calculated frequencies of all consonant gradation patterns from our corpus using a grammatical description of the language [16] and filled in missing and scarce patterns. Additionally, in order to ensure that our TTS voices would also be able to correctly produce English, Swedish and Norwegian names as well as loanwords, we prepared a small amount of texts in these languages for the recordings.

Using the resulting text corpus, we recorded two voice talents during 2022 for training a male and a female voice. This resulted in approximately 8 hours of speech for the male voice and 12 hours of the female voice after cleaning and processing the data. Because we wanted to take the different Lule Sámi areal varieties into account, we chose the male speaker from Norway and the female speaker from Sweden.

We recorded the **male** voice at an office of the Norwegian broadcasting company, NRK in spring 2022. After the recordings were done, we adjusted the read texts as accurately as possible according to what was actually read by the voice talent. This included transcribing all repetitions and self-corrections in the texts. Only clear mistakes and other non-usable speech were cut out from the data, because we had a very limited amount of time for the recordings and by removing all repetitions, we would have lost a lot of usable material.

The recordings were originally made in 48000 Hz sample rate and 24 bit depth but with our chosen TTS framework, Fastpitch [10], the audio was downsampled to 22050 Hz. Additionally, we applied audio filters to our material, such as echo removal, noise gate and level normalization to guarantee the best possible audio quality. The echo removal was done using *Multiband Envelope Shaper* in Cubase software (`https://www.steinberg.net/cubase`[2]). Noise gate was applied to the material using an Audacity plugin, and for level normalization we used the *sox* command-line audio processing tool and the STL toolkit (`https://github.com/openitu/STL`).

Next, the recordings were automatically force-aligned with the adjusted texts on word and sentence levels using the WebMAUS Basic online tool [17], available at `https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic`. The advantage of using WebMAUS is that even relatively long audio files (30 min) can be processed at once without having to previously chunk the files first. There are no Sámi speech models available for WebMAUS, but using the model of a related language, Finnish, was useful for finding the sentence boundaries for splitting the data into .wav/.txt pairs.

After splitting our material to 7925 training sentences, leaving aside 102 sentences for the validation set, we ran the standard Fastpitch pre-processing script at `https://github.com/NVIDIA/DeepLearningExamples/blob/master/PyTorch/SpeechSynthesis/FastPitch` with default settings. This script prepares the dataset for training the model by extracting the mel spectrograms and pitch values for each .wav file.

Currently, we are transcribing the materials recorded from the female speaker representing the Swedish variety of Lule Sámi, and once finished, we plan to follow the same procedures as we did for the male voice.

3.1.2. North Sámi

As mentioned above, the first TTS voices (male and female) were developed for North Sámi in 2015 as a closed-source project by Divvun and Acapela. Later, in [18], the same material from the original **female** voice was used to train a neural speech model, using a setup combining Tacotron, Tacotron 2, ForwardTacotron and WaveGlow, the two latter ones from the official Nvidia repository at `https://github.com/NVIDIA/tacotron2`. The training/validation split ratio in [18] was 3591/200 and the model was trained for 600k steps. It was concluded that even though the modeling worked well, the 4-component structure of the training process could be possibly simplified without compromising the output quality.

In the present project, we "continue" this work with two tasks: 1) training a FastPitch [10] model with the exact same dataset, and 2) recording more material to complement the previous one.

Although we have not yet performed a formal evaluation for our North Sámi FastPitch model yet, we noticed overall better quality of the inference output with Fastpitch compared to Tacotron 2 from [18]. Especially, the naturalness of the prosody makes the FastPitch output more pleasant to listen to. For initial evaluation of our model, we asked a native speaker specializing in phonology to assess a few sentences we generated with the model, and a few inaccuracies and irregularities were reported in vowels particularly, possibly due to the fact that vowel length is not marked systematically in the North Sámi orthography. Thus, we decided to collect more data to fill in possible gaps in the material, and luckily, the same female voice talent who originally participated in the TTS project agreed to record more material with new texts.

In March 2023, we recorded ca. 4 more hours to add to the new, combined North Sámi TTS corpus. For the texts, we followed a similar pattern as with the Lule Sámi TTS text corpus by adding text styles such as shown in Figure 2 and even poetry. At the moment, we are pre-processing the material for training a new model for the female North Sámi voice.

3.2. Model configuration and evaluation

All models described in this section are trained on the Norwegian academic high-performance computing and storage service Sigma2 (`https://www.sigma2.no/`).

We trained a Lule Sámi FastPitch model with the male speech corpus described above (ca. 8 hours, 8100 sentences). After defining the orthographic symbol set for the language, we trained the model for 660 epochs with batch size 1 and learning rate 0.1. For inference, we used the UnivNet model [19] from NeMo collections as the vocoder.

While we have not yet conducted any formal evaluation studies of our FastPitch model, the quality of the TTS output seems very good with natural-sounding intonation. Our group member who is a native speaker of Lule Sámi did not report any remarkable mistakes in the output and stated that the voice is highly intelligible. Based on this preliminary assessment of the model, we see a lot of potential use cases for our Lule Sámi TTS. To confirm the appropriateness of the voice before releasing it for public use, we plan to conduct a thorough evaluation with it, following the methodology from [20], for example. In the planned evaluation experiment, we will measure the quality

[2]Note that Cubase is commercial software.

of the voice with different scales for articulation, speaking rate, voice pleasantness, listening effort and overall impression.

Currently, we are processing the recordings from the female Lule Sámi voice and hope to start training the second voice later this year following the same procedure as with the male voice.

As mentioned in Section 3.1.2, we trained new FastPitch models for North Sámi using mostly the same speech materials as in [18], but added some Norwegian utterances and other extra recordings with rarer sound combinations and words with exceptional pronunciation. The training material for the female voice was 4.3 hours, and 5.7 hours for the male speaker. With training/testing sentence splits of 3573/51 for the female voice and 4144/51 for the male voice, both models were trained for 1000 epochs.

Similarly with Lule Sámi, no proper evaluation test has been done to our new TTS models yet. We hope to improve the model with regards to this by doubling the size and quality of the training data with new recordings from the same female speaker and by making sure that the training data is as accurately transcribed as possible.

4. Work-in-progress for North Sámi ASR

In this section, we describe the related works and our work-in-progress of developing ASR for North Sámi. According to the feedback we get from the language communities, there is high demand for a speech-to-text tool like ASR, for example for making automatic transcriptions or subtitling videos.

To the best of our knowledge, [21] documents the development of the first ASR model for North Sámi. Using an audiobook read aloud by a single female speaker of North Sámi as training data, the resulting recognizer was intended to be deployed as part of a spoken dialogue system in the WikiTalk application ([22]).

In [18], an ASR model was trained for North Sámi in a dual transformation setup with the TTS mentioned above. The methodology for the ASR model was based on Wav2Vec2 and the training materials consisted on read speech only (the TTS material from the North Sámi speakers mentioned above). The model, available at `https://github.com/divvun/lang-sme-ml-speech`, was trained for 30k steps, reaching a word-error-rate (WER) of 41%.

To improve the model performance, we acquired additional, spontaneous speech data from various sources (mainly from the Language Banks of Norway and Finland). The size of our new speech corpus for training ASR was approx. 34 hours. We fine-tuned the model from the *facebook/wav2vec2-large-xlsr-53* pre-trained model with the new dataset for 104,750 steps and 250 epochs, reaching a WER of 29%. Generally, we noticed that the new dataset helped the model perform better also with noisy and spontaneous speech input.

A more recent multilingual model, trained in a supervised manner on 680k hours of subtitled content, showed exceedingly good results in a variety of languages [23]. The Whisper model, as the authors named it, is capable of performing well even in noisy environments and provides transcriptions in a readable format without the need for an extra inverse text normalization process. Unfortunately, the model did not include any of the Sámi languages and thus was not directly usable for our purposes. However, the model included Finnish, a closely related language to North Sámi. In the first experiment of its kind, we were able to reuse the existing weights for Finnish and fine-tune the model with North Sámi annotated speech, wiping out Finnish in the process but providing ASR for North Sámi. We achieved a WER score of 24.91% on a held-out test set randomly extracted from the training corpora described above[3]. The prototype model is available at `https://huggingface.co/NbAiLab/whisper-large-sme`. Table 1 shows examples of both ASR approaches, the target transcription, and its English translation. The Whisper model, despite being able to generate capitalization and punctuation marks, did not include any of this data in the training set for Sámi.

Table 1: *Comparing prediction outputs between Wav2Vec2 and Whisper North Sámi ASR models. Bolded parts show sections that differ from the target sentence.*

Wav2Vec2	*ja de bosui davvebiegga nu garr**osiid** go sá**h**ii muhto ma**đ**i**i** eane**s** son bosui da**đ**i**i** **čávga deappo** vánddardea**dj**i **gieasaid** jáhka **eižas** birra*
Whisper	*ja de bosui davvebiegga nu garrasit go sáhtii muhto maði eanet son bosui dađi **čávga lea eambbo go** vánddardeaddji geas**á**i **jahke** iežas birra*
Target	Ja de bosui davvebiegga nu garrasit go sáhtii, muhto maði eanet son bosui, dađi čavgadeappot vánddardeaddji g**ie**sai jáhka iežas birra.
Translation	And then the North Wind started blowing as hard as it could, but the harder the wind blew down the road, the tighter the man clung to his coat.

We are now in the process of expanding these experiments and produce a readily usable model for ASR based on the Whisper architecture, with the hope of making it multilingual in at least Lule and North Sámi.

5. Results and Discussion

With the dataset and settings described above, we were able to produce potentially suitable end-user TTS models for two Sámi languages. Once finished, these models will be integrated into the Divvun tool set and the GiellaLT infrastructure as well as the most common operating systems for flexible and effortless use. Developing the TTS voices for two Sámi languages works as a good case study in the very low-resource language setting.

In addition to improving the TTS quality by adding more data or using bigger datasets, we could try utilizing approaches using multilingual transfer learning or multi-speaker setups, as in [9] and [24]. In these, datasets or pre-trained models from well-resourced languages were used to improve the TTS performance, especially so if shared input representation spaces (phoneme mappings) were used. We plan to conduct an evaluation of our TTS systems to confirm that the voice is suitable and appropriate for various situations and use cases, and that is also clear and pleasant to listen to, as suggested in [25].

Our latest experimental ASR model has already shown to be useful, especially for raw-transcribing big amounts of speech materials. In the near future, we plan to develop the North Sámi ASR further and eventually make it openly available.

6. Acknowledgements

Many thanks to Antti Suni, Atte Asikainen (University of Helsinki) and Sébastien Le Maguer (ADAPTCentre / Trinity College Dublin) for their help and advice.

[3]The WER scores are not directly comparable as the test sets differ.

7. References

[1] C. Moseley, *Atlas of the World's Languages in Danger*. Unesco, 2010.

[2] P. Sammallahti, *The Saami languages: an introduction*. Davvi girji, 1998.

[3] A. Aikio, L. Arola, and N. Kunnas, "Variation in north saami," in *Globalising Sociolinguistics*. Routledge, 2015, pp. 263–275.

[4] M. P. Lewis, Ed., *Ethnologue: Languages of the World*, sixteenth ed. Dallas, TX, USA: SIL International, 2009.

[5] A. Yaneva, "Speech technologies applied to second language learning. A use case on Bulgarian. Bachelor's thesis." 2021.

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

[7] J. Xu, X. Tan, Y. Ren, T. Qin, J. Li, S. Zhao, and T.-Y. Liu, "Lrspeech: Extremely low-resource speech synthesis and recognition," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2802–2812.

[8] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.

[9] L. Rätsep and M. Fishel, "Neural text-to-speech synthesis for Võro," in *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, 2023, pp. 723–727.

[10] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6588–6592.

[11] K. Hiovain-Asikainen and S. Moshagen, "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*, 2022, pp. 169–175.

[12] A. Suni, T. Raitio, D. Gowda, R. Karhila, M. Gibson, and O. Watts, "The simple4all entry to the blizzard challenge 2014," in *Proc. Blizzard Challenge*. Citeseer, 2014.

[13] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[14] G. Săracu and A. Stan, "An analysis of the data efficiency in Tacotron2 speech synthesis system," in *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, 2021, pp. 172–176.

[15] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[16] N. E. Spiik, *Lulesamisk grammatik*. Sameskolstyrelsen, 1989.

[17] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[18] L. Makashova, "Speech synthesis and recognition for a low-resource language: Connecting TTS and ASR for mutual benefit," Master's thesis, University of Gothenburg, 2021.

[19] W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim, "Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation," *arXiv preprint arXiv:2106.07889*, 2021.

[20] I. ITU, "A method for subjective performance assessment of the quality of speech voice output devices," *International Telecommunication Union Std*, 1994.

[21] J. Leinonen *et al.*, "Automatic speech recognition for human-robot interaction using an under-resourced language," Master's thesis, 2015.

[22] G. Wilcock and K. Jokinen, "Wikitalk human-robot interactions," in *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013, pp. 73–74.

[23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[24] P. Do, M. Coler, J. Dijkstra, and E. Klabbers, "Text-to-speech for under-resourced languages: Phoneme mapping and source language selection in transfer learning," in *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, 2022, pp. 16–22.

[25] P. Wagner, J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. Eje Henter, S. Le Maguer, Z. Malisz, É. Székely, C. Tånnander *et al.*, "Speech synthesis evaluation—state-of-the-art assessment and suggestion for a novel research program," in *Proceedings of the 10th Speech Synthesis Workshop (SSW10)*, 2019.