Developing TTS and ASR for Lule and North Sámi languages

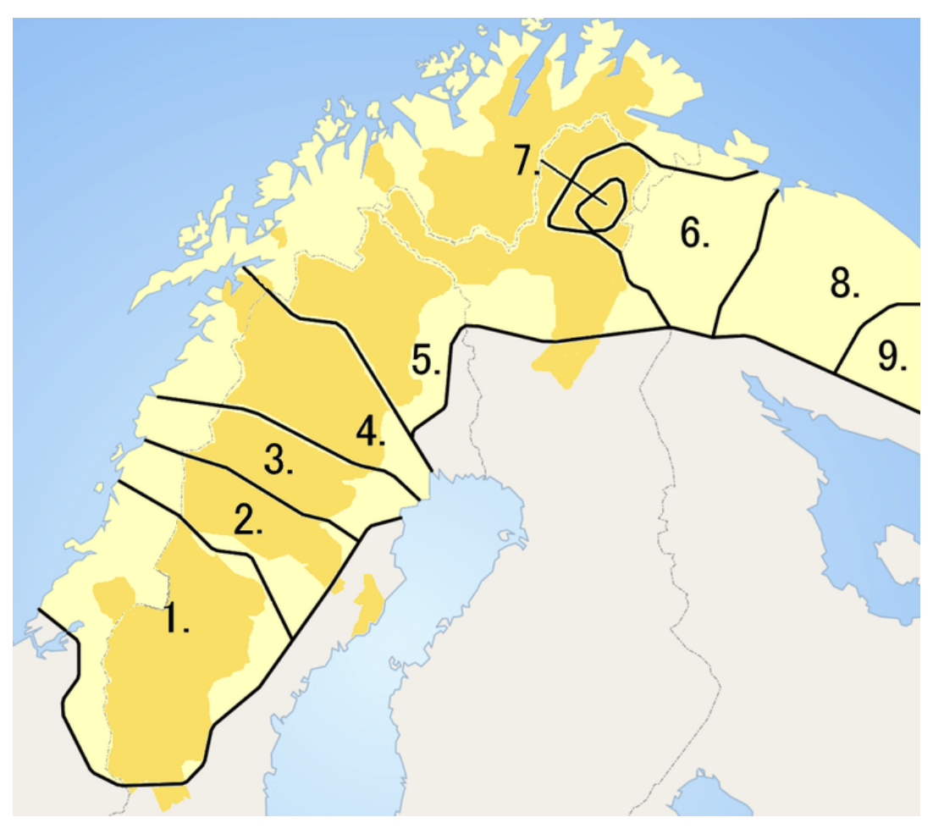
Katri Hiovain-Asikainen[1] and Javier de la Rosa[2]
katri.hiovain-asikainen@uit.no, versae@nb.no
[1]UiT Norgga árktalaš universitehta
[2]National Library of Norway

Abstract

- Recent innovations in speech technology have made high quality TTS and ASR available even for **extremely low-resource** languages.
- We present our **updated work-in-progress** on open-source speech technology for two indigenous Sámi languages (minority languages in Norway, Sweden and Finland).
- We have created text and speech corpora for training the first neural **TTS for Lule Sámi**, and updated the previous **North Sámi TTS** by collecting additional materials and by training a new model.
- We describe our first experiments with developing **ASR for North Sámi** and discuss the next steps to be taken in our project.

The Sámi Language Areas

1. South Sámi
2. Ume Sámi
3. Pite Sámi
4. **Lule Sámi**
5. **North Sámi**
6. Skolt Sámi
7. Inari Sámi
8. Kildin Sámi
9. Ter Sámi



Sámi language areas (Wikimedia Commons).

Since 2001, Divvun and Giellatekno groups (divvun.no and giellatekno.uit.no/) have developed and maintained an infrastructure of dictionaries, morphological analyzers, spell checkers and other tools.

Lule and North Sámi

Lule Sámi is spoken by 800-3,000 speakers in northern Sweden and Norway, and is classified as a severely endangered language by UNESCO. A written standard of Lule Sámi was approved in 1983.
⇒ giellalt.github.io/lang-smj/

North Sámi is the largest Sámi language by number of speakers, ca. 25,000 in three countries (Norway, Sweden and Finland), also classified as endangered by UNESCO. A written standard of North Sámi was adopted in 1979. In 2015, the first TTS tool was developed for North Sámi as a closed-source project.
⇒ giellalt.github.io/lang-sme/

Introduction

- Lule and North Sámi have **official status** in Norway. As such, these languages should be present in all official contexts along with Norwegian in the digital realm. Speech technology is also essential for people with special needs: language learners, vision-impaired and dyslexic individuals.
- The current development of Sámi TTS and ASR will expand the selection of tools **from text-based only into spoken language**.
- The biggest challenge in working with extreme low-resource languages is **lack of existing data** and pre-trained models. Training datasets often need to be created from scratch or acquired from various sources.
- When working with endangered, indigenous languages, it is very important to take **ethical issues** into account: the resulting tools must answer the needs of the language community.
- In our project, only **open-source** methodologies are used: **FastPitch** for TTS and **Wav2Vec2** and **Whisper** for ASR. Our aim is also to produce open-source datasets and pre-trained models for future use.

TTS Speech Corpora

Language	Duration	Samples
Lule Sámi	Male voice: ∼8 hrs	Male voice: 7925/102
North Sámi	Male voice: ∼5.7 hrs; Female voice: ∼4.3 hrs	Male voice: 4144/51; Female voice: 3573/51

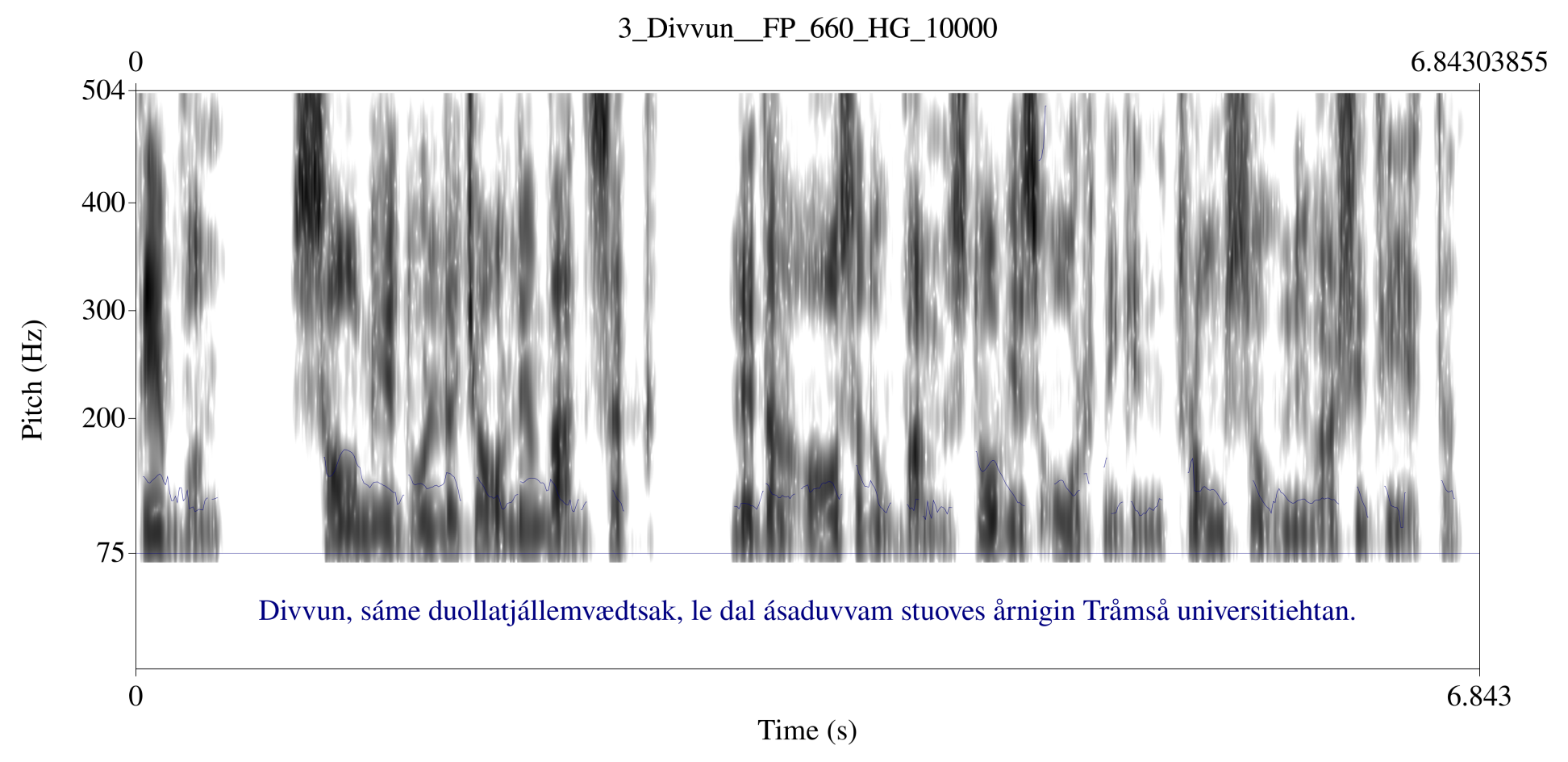
Language, corpora duration in hours, and number of samples in the training and validation sets.

The Lule Sámi materials were recorded within the Divvun project and the North Sámi materials were acquired from the previous North Sámi TTS project by Divvun and Acapela.

TTS Model Setup

- The Lule and North Sámi voices were trained using the official **FastPitch** framework by NVIDIA (https://fastpitch.github.io/). For both languages and all voices, the learning rate was set to 0.1 and batch size to 1.
- After defining the symbols sets for each language, the models were trained on the Norwegian academic high-performance computing and storage service **Sigma2** (sigma2.no). The Lule Sámi model was trained for 660 epochs and the North Sámi ones for 1000 epochs.
- For inference, we used the **UnivNet** model from the NVIDIA NeMo collections as the vocoder.

ASR for North Sámi



Sample Spectrogram from Lule Sámi FastPitch model inference.

- Spontaneous speech data from various sources was acquired from the Language Banks of Norway and Finland, the resulting training data for ASR was ∼34 hours.
- First, we fine-tuned a **Wav2Vec2** pre-trained model from Facebook (hf.co/facebook/wav2vec2-large-xlsr-53) with the new dataset for 104,750 steps and 250 epochs, reaching a WER of **29%**.
- Then, we trained a **Whisper** model with a newer technology for 60,000 steps, we achieved a WER score of **24.91%** on a held-out test set randomly extracted from the training corpus. The prototype model is freely available.
⇒ hf.co/NbAiLab/whisper-large-sme

North Sámi ASR Example

Wav2Vec2	*ja de bosui davvebiegga nu garr**osiid** go sá**h**ii muhto mađ**ii** eane**s** son bosui dađi**i** **čávga deappo** vánddardea**dj**i **gieasaid** jáhka **eižas** birra*
Whisper	*ja de bosui davvebiegga nu garrasit go sáhtii muhto mađi eanet son bosui dađi **čávga lea eambbo go** vánddardeaddji geas**ái** **jahke** iežas birra*
Target	Ja de bosui davvebiegga nu garrasit go sáhtii, muhto mađi eanet son bosui, dađi čavgadeappot vánddardeaddji **gie**sai jáhka iežas birra.
English	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.

Discussion

- We were able to produce potential end-user suitable TTS voices for two Sámi languages that will be integrated into the Divvun tool set and the GiellaLT infrastructure as well as into the most common operating systems for effortless use
- Next, we will adopt approaches using multilingual transfer learning and multi-speaker setups to improve the TTS performance and to conduct an evaluation test to our TTS voices to confirm their quality
- Our experimental North Sámi ASR model has already shown to be useful, especially for raw-transcribing big amounts of speech materials
- In the future, we plan to develop the North Sámi ASR model further and eventually make it openly available

More Info & References

nbailab.github.io/sigul2023_sami_tts_asr

UnivNet Vocoder: Jang, W., Lim, D., Yoon, J., Kim, B., Kim, J. (2021) UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation. Proc. Interspeech 2021, 2207-2211, doi: 10.21437/Interspeech.2021-1016