Building Open-Source Speech Technology for Low-Resource Minority Languages with Sámi as an Example – Tools, Methods and Experiments

Katri Hiovain-Asikainen and Sjur Nørstebø Moshagen
UiT Norgga árktalaš universitehta

Abstract

Motivation: Describing the ongoing work of our open-source Lule and North Sámi speech technology project, where we are developing TTS and ASR in low-resource settings
Objectives: In addition to documenting the steps we have taken for our project, we discuss utilizing technologies that allow for transfer learning between neighboring languages. We also address effective and mindful use of the speech corpus, and also possibilities to use archive materials for training an ASR model for these languages.
Prerequisites: Native speakers of the language(s): linguists and voice talents. Any phonetic description of the language.
Time effort: A three-year project for a team of 4 linguists and/or engineers

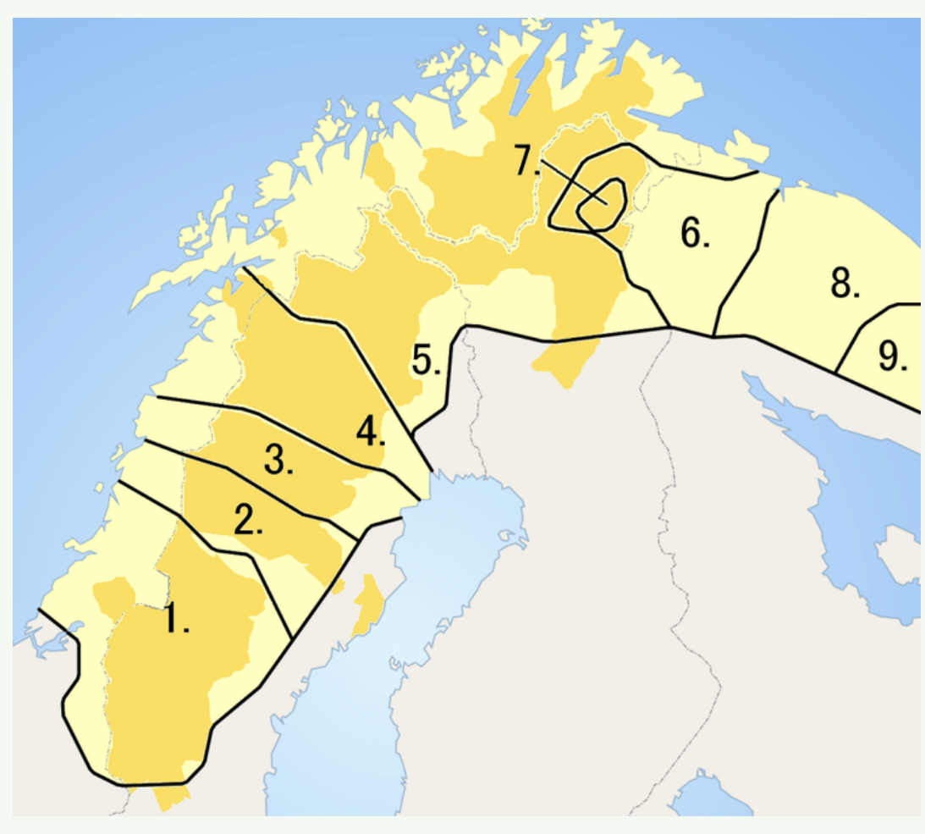
Lule and North Sámi

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 biggest Sámi language, spoken by ca. 25 000 speakers in three countries (Norway, Sweden and Finland) and is 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. Neither the framework used to develop it nor the speech corpus are publicly available.
⇒ giellalt.github.io/lang-sme/

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



	Orth.	IPA	Transl.
Q3	*oarre*	[ʔo͡ɑrːrɪɛ]	’a squirrel’ Nom.Sg
Q2	*oarre*	[ʔoɑrːɪɛ]	’a squirrel’s’ Gen.Sg ’a reason’ Nom.Sg
Q1	*oare*	[ʔoɑrɪɛ]	’a reason’s’ Gen.Sg

Table 1: Ternary length contrast of consonants in Lule Sámi, underspecified in the orthography. Abbreviations: Q3 – overlong, Q2 – long, Q1 – short. Examples originally presented in Fangel-Gustavson et al. (2014).

Introduction

- The development of a TTS system as a whole requires multidisciplinary input: natural language processing (NLP), phonetics and phonology, machine learning (ML) and digital signal processing (DSP). Tasks connected to NLP are important in developing the text front-end for the TTS
- Most orthographies are underspecified with respect to the pronunciation, in the Sámi languages especially the consonant gradation phenomenon requires special effort and preferably a text2IPA conversion of the texts for better phonological accuracy of the TTS model
- In the master’s thesis by Liliia Makashova (2021), a TTS model utilizing Tacotron2 and ForwardTacotron was trained with only 3 hours of data per speaker, producing intelligible speech output for North Sámi
- The resulting TTS and ASR models can be further used to build more advanced speech technology frameworks such as dialogue systems or (spoken) language learning mobile applications that will further benefit the language communities.

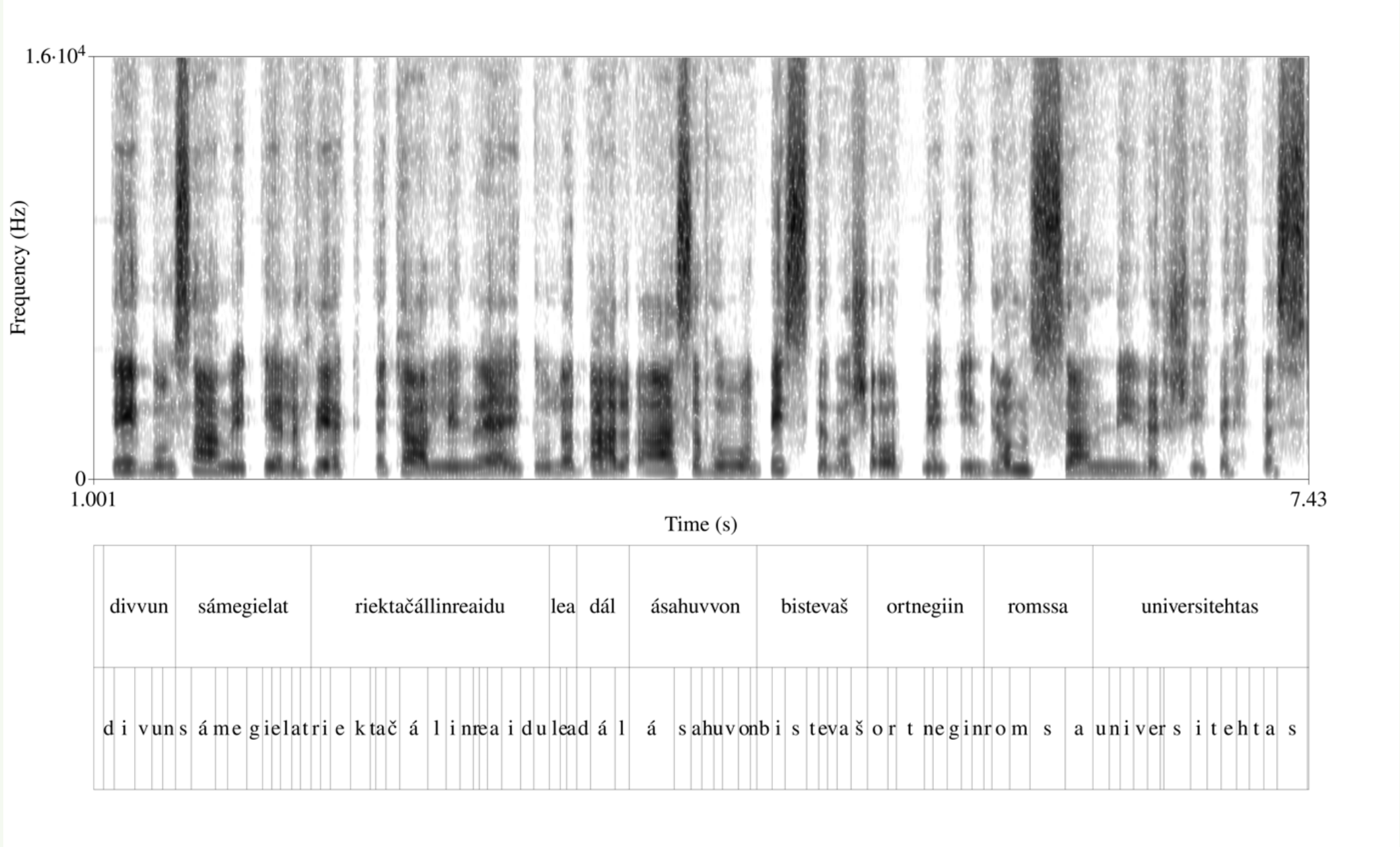
Corpus building

- **Text corpus**:
  - Collecting good quality open-source (CC-BY) texts from different domains (news, educational, fiction etc.)
  - For at least 10-12 hours of spoken material, approx. 75 000 words of text is required
- **Speech corpus**
  - A speech model in End2End systems like Tacotron2 is built directly from sentence-long audio files and the corresponding texts
  - ML-based approaches require at least 10 hours (per speaker) of CD-quality data, recorded in a professional setup

Corpus processing

- **Step 1**: After collecting the text corpus, the read speech corpus is recorded. The original texts are then edited to exactly match the recordings
- **Step 2**: The text and audio files are force-aligned (using *WebMAUS Basic*) to automatically find sentence boundaries. Before splitting the corpus to sentence-long files, the audio is cleaned from noise/artefacts, filtered and normalized
- **Step 3**: The resulting corpus is fed to the Tacotron2 (for instance) pre-processing (resampling audio etc.) and speech model training pipeline. Training the TTS model at a computing cluster (*Sigma2* in Norway), assessing the quality and accuracy of the output

Model Evaluation and results so far



- In addition to the North Sámi TTS (3 hrs of data), described in the thesis by Makashova (2021), we have experimented with a miniature 1 hr data set with Lule Sámi using a DNN/HMM-based Ossian TTS as well as tried using Tacotron2 *transfer learning* method by training the North Sámi model further with Lule Sámi data
- From these, Ossian TTS produced slightly more intelligible results with the 1 hr data set
- While the training process of the Lule Sámi voice was successful, the output showed that the data has to be improved both in terms of quality and quantity of the data as not all necessary phonemes were covered and the output was noisy
- When comparing the North and Lule Sámi TTS models, it is clear that the model output improves drastically with more data (see the spectrogram image from the North Sámi output)

Contact Info & References

Web: divvun.no,
giellatekno.uit.no, uit.no
Email: `feedback@divvun.no`
GitHub: github.com/giellalt, github.com/divvun,
github.com/divvun/lang-sme-ml-speech
Ossian-TTS: github.com/CSTR-Edinburgh/Ossian
WebMAUS Basic: clarin.phonetik.uni-muenchen.de/BASWebServices/interface

Fangel-Gustavson, N. et al. (2014): Quantity contrast in Lule Saami:A three-way system. In Proceedings of the 10th International Seminar on Speech production, pages 106–109.
Makashova, L. (2021): Speech Synthesis and Recognition for a Low-Resource Language – Connecting TTS and ASR for Mutual Benefit

Future directions and Conclusions

- In addition to TTS, we are working towards developing a tool for automatic speech recognition (ASR) for North Sámi
- In the thesis by Makashova (2021), an ASR model was trained with 6 hours of data (2 spkrs) and for 30k steps, reaching a WER (Word-Error-Rate) of 41% and 0.5 loss
- We are currently developing the ASR model by obtaining spoken North Sámi corpora from the language banks of Norway and Finland that contain spontaneous multi-speaker and multi-dialectal speech
- Such a tool will contribute to the documentation and to better usability of any untranscribed Sámi archive materials, allowing to use them for, e.g., linguistic research

Conclusions:

- We hope that the procedures described above could be applied to any (minority) language with a low-resource setting, in the task of developing speech technology applications
- As for TTS, if a speech corpus must be built from scratch, it has to be designed to prioritise quality over quantity of the corpus
- By making the speech corpus used for developing TTS openly available, future needs to collect similar corpora are reduced