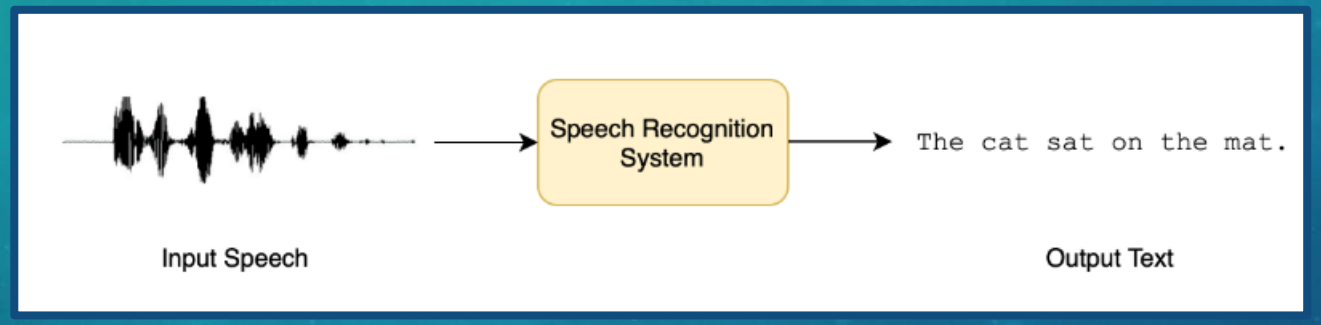
EVALUATING THE PERFORMANCE OF ASR MODELS ON OSSETIC SPEECH USING FIELDWORK AUDIO DATA

VARVARA PETROVA

ASR: GENERAL INFORMATION



- Automatic Speech Recognition (ASR, Speech-to-Text) = subfield of computer science dealing with the recognition and translation of spoken language into text by computers;
- mono- or multi-lingual models.

DATASET FOR OSSETIC ASR:
MOZILLA-FOUNDATION/COMMON_VOICE_17_0

- the only existing dataset with transcribed audios in Ossetic on HuggingFace (aside from its previous versions);
- parameters (relevant for Ossetic):

audio	sentence	upvotes	downvotes	age	gender	accent

Ossetic subset:

- 1.2 hours, 386 sentences;
- Wikipedia articles, narrated speaking style;
- train subset:
  - accent = Цæрат ‘northern’ = Standard Iron;
  - age = fourties, gender = masculine;
  - all data most likely come from one speaker.

MODEL TRAINED ON OSSETIC DATA: FACEBOOK/MMS-1B-ALL

- massive multilingual model: >1000 languages, including Ossetic;
- 1 billion parameters;
- audio sampling rate = 16,000 Hz;
- only google/fleurs (with no Ossetic subset) listed as a source of data;
- CommonVoice test dataset mentioned in a model card -> probably fine-tuned using CommonVoice.

RESEARCH QUESTION

all of the data in CommonVoice Ossetic training subset come from either one or very few middle-aged male speakers of Standard Iron ->

How good is facebook/mms-1b-all at transcribing speech of other native speakers of Standard Iron?

Is it at least somewhat successful at transcribing the speech of:

- the people whose data are used in the CommonVoice test subset?
- other middle-aged male speakers of Standard Iron?
- younger male speakers of Standard Iron?
- female speakers of Standard Iron?
- speakers of other Ossetic idioms, e. g. Digor?

OSSETIC FIELDWORK AUDIO DATA

- collected during field trips to Vladikavkaz in 2023-2025 for phonetic research;
- 22 speakers of various dialects of Ossetic (17F, 5M);
- Zoom H5 recorder, Shure WH20XLR headset microphone;
- words recorded in isolated environment and in carrier phrases (“Say [word] three times” and similar sentences), sometimes as a single recording;
- for both environments, speakers had to repeat either a word or a sentence three times with pauses;
- the consultants were asked to produce utterances in a natural speaking manner.

RESAMPLING

- for the model to work properly on the new data, all audios need to be resampled to the sample rate of 16,000 Hz;
- we can do it using two python libraries for signal processing: soundfile and scipy.signal:

```python
import numpy as np
import soundfile as sf
from scipy.signal import resample

def resample_audio(input_file, target_sample_rate):
    data, original_sample_rate = sf.read(input_file)

    number_of_samples = round(len(data) * float(target_sample_rate) / original_sample_rate)
    resampled_data = resample(data, number_of_samples)

    return resampled_data, target_sample_rate

def process_folder(input_folder, output_folder, target_sample_rate=16000):
    if not os.path.exists(output_folder):
        os.makedirs(output_folder)

    for filename in os.listdir(input_folder):
        if filename.endswith('.WAV'):
            input_file = os.path.join(input_folder, filename)
            output_file = os.path.join(output_folder, filename)

            resampled_data, sample_rate = resample_audio(input_file, target_sample_rate)

            sf.write(output_file, resampled_data, sample_rate)
            print(f'Resampled: {filename}')

input_folder = '/content/drive/MyDrive/os-dataset/os-audios'
output_folder = '/content/drive/MyDrive/os-dataset/os-audios-resampled'

process_folder(input_folder, output_folder)
```

- examples of original and resampled audios (to ensure that ASR errors are not caused by resampling distortions):

original audio

resampled audio

TESTING THE CAPABILITIES OF FACEBOOK/MMS-1B-ALL

1. TEST SUBSET OF MOZILLA-FOUNDATION/COMMON_VOICE_17_0

Age	Gender	Transcription: expected/produced by the model
middle-aged	masc	Уæд чызг хъынцъым кæнын байдыдта, бонæй-бон æнкъарддæр кодта. / vrecisk kinsim kenin baydita bone bon elkarter kotta
middle-aged	masc	Таймураз ма бады къæсæры фарсмæ./ to imu rozno bodoe kaser fosma
middle-aged	masc	Дæ бæх дæр йæ фæллад суадзид, дæхæдæг дæр баулæфис, стæй райсом афæндараст уаис./ dbrdr iaflla cwazi dcebgdar bawl cisst rasson afandar stwais
?	fem	Ницы кæны, аныхас кæнут, бирæ нал фæдзурдзыстут иумæ. miiskano an haskonut pira nalfa zutrsto tomo
?	fem	Бæлвырд, æз радтон дзырд Тегайæн æмæ мæ дзырд хъуамæ сæххæст кодтаин. plwird ijr tonzertigan mzrt ms xesq rtain
?	fem	Цæй ми мыл æрцыди! smi mwrsd
teens	masc	Мæсыгæй куы ракаст, уæд æндæр кæрты ауыдта сылгоймаджы хуызист. masikei qrakst we dn derkert awttasilgoi maji xejist
teens	masc	Иу тар ранмæ куы бахæццæ сты, уæд Тега бæх ныуурæдта. iwtarrmqb xzs btigabarn urta
teens	masc	Рахызтысты. rah is tisti

1. TEST SUBSET OF MOZILLA-FOUNDATION/COMMON_VOICE_17_0

- little Ossetic data -> performance depends largely on the data from other languages & most of the transcriptions in a multi-language dataset use Latin script -> transcriptions are in Latin script;
- no tokenizer -> completely wrong tokenization;
- most notable phonetic issues:
  - trouble recognizing sonorants;
  - b = p in female speaker's utterances (likely due to much English data in the dataset: aspirated voiced = voiceless);
  - æ transcribed as a, o, e depending on the context (little Ossetic data in training & symbol abscent from most other scripts).

FIELD DATA

```
audio_files = [
    "/content/drive/MyDrive/new_os/VS_mit_cont.WAV",
    '/content/drive/MyDrive/new_os/VS_sybyrtt_cont.WAV'
    ]

from transformers import Wav2Vec2ForCTC, AutoProcessor
import torch
import soundfile as sf

model_id = "facebook/mms-1b-all"
processor = AutoProcessor.from_pretrained(model_id)
model = Wav2Vec2ForCTC.from_pretrained(model_id)

def transcribe_audio(file_path):
    audio_input, sampling_rate = sf.read(file_path)

    inputs = processor(audio_input, sampling_rate=16000, return_tensors="pt")

    with torch.no_grad():
        outputs = model(**inputs).logits

    ids = torch.argmax(outputs, dim=-1)[0]
    transcription = processor.decode(ids)

    return transcription

for audio_file in audio_files:
    transcription = transcribe_audio(audio_file)
    print(f"Transcription for {audio_file}:\n{transcription}\n")
```

```
Transcription for /content/drive/MyDrive/new_os/VS_mit_cont.WAV:
rl  f   f rol  f   ol w fo
```

2. FIELD DATA: MIDDLE-AGED AND YOUNG MALE (STANDARD IRON)

Age	Expected transcription	Transcription produced by the model
middle-aged	Читт, читт, читт.	xb g g
middle-aged	Зæгъ фæнык æртæхатты. Зæгъ фæнык æртæхатты. Зæгъ фæнык æртæхатты.	as fon f
middle-aged	Мит, мит, мит.	b b b
middle-aged	Зæгъ сыбыртт æртæхатты. Зæгъ сыбыртт æртæхатты. Зæгъ сыбыртт æртæхатты.	r z
middle-aged	Зæгъ хæрæфырт æртæхатты. Зæгъ хæрæфырт æртæхатты. Зæгъ хæрæфырт æртæхатты.	f a f a s fa as
young	Чъепп, чъепп, чъепп. Зæгъма чъепп æртæхатты. Зæгъма чъепп æртæхатты. Зæгъма чъепп æртæхатты.	c ravmo c b f rav mo cb f ravmo c brz
young	Зæгъ сыбыртт æртæхатты. Зæгъ сыбыртт æртæхатты. Зæгъ сыбыртт æртæхатты.	rul sn bu rul s buc rul bu bc
young	Мит, мит, мит. Зæгъма мит æртæхатты. Зæгъма мит æртæхатты. Зæгъма мит æртæхатты.	rl f f rol f ol w fo

2. FIELD DATA: MIDDLE-AGED AND YOUNG MALE (STANDARD IRON)

- the output is, once again, produced in the Latin script;
- many outputs are sets of consonants (i. e. vowels are mostly omitted altogether), the quality is significantly lower compared to the test subset of CommonVoice;
- choosing the speaker from the same age group as the one whose data were used for training did not help, likely due to the insignificant role of a small Ossetic subset in the output;
- out of all consonants, model recognizes fricatives (f, s) and r with relative success and tends to parse most of labial consonants as b for both speakers;
- transcription of z appearing in the beginning of a carrier phrase (zægъ, zægъma) as s can be explained by the rarity/absence of voiced sibilant fricative in other languages of the dataset compared to its voiceless counterpart.

3. FIELD DATA:
FEMALE SPEAKERS (STANDARD IRON) AND MALE SPEAKER OF DIGOR

Speaker	Expected transcription	Transcription produced by the model
SI-F-1	Зæгъ адæмыхатт æртæхатты. Зæгъ адæмыхатт æртæхатты. Зæгъ адæмыхатт æртæхатты.	zul odo zu odo d j o
SI-F-1	Зæгъ дзæцц æртæхатты. Зæгъ дзæцц æртæхатты. Зæгъ дзæцц æртæхатты.	jo or jo or jo orofod
SI-F-2	Чъиу, чъиу, чъиу.	ili
SI-F-2	Ничи, ничи, ничи.	c wc
SI-F-3	Зæгъ бецыкк æртæхатты. Зæгъ бецыкк æртæхатты. Зæгъ бецыкк æртæхатты.	b s m b o m bc ur
SI-F-3	Зæгъма адæмыхатт æртæхатты. Зæгъма адæмыхатт æртæхатты. Зæгъма адæмыхатт æртæхатты.	zm adim sr sz zm ai o xm adim sd o s
D-M	Сес, сес, сес.	ss s ss
D-M	Зæгъа сирд æртæхатти. Зæгъа сирд æртæхатти. Зæгъа сирд æртæхатти.	cmr or z cmd or orz

3. FIELD DATA: FEMALE SPEAKERS (STANDARD IRON) AND MALE SPEAKER OF DIGOR

- since the output for Iron males was already of very poor quality, there is no visible decrease;]
- many transcriptions are sets of consonants, vowels are better recognized in monosyllabic words where vowels are the longest [Sokolova 1953];
- the model is, again, not as bad at recognizing sibilants, labials, and rhotics as everything else;
- the model is better at recognizing the voicedness of z in female speakers.

DISCUSSION AND PROSPECTS

- facebook/mms-1b-all trained on approx. 1,2 hours of CommonVoice Ossetic data performs poorly on CommonVoice test subset;
- …and even worse when
- solution: fine-tuning models using
  - much more Ossetic data (including other dialects and varieties);
  - models trained for ASR of languages with similar systems of phonemes and allophones;
  - models trained exclusively for ASR of languages that use Cyrillic script (e. g. Russian).

CREATING A DATASET FOR OSSETIC ASR

- dataset structure: audio in a form of either direct digital representation of sound wave (see below; also used in CommonVoice) or mfccs; transcription; sampling rate;
- dataset sources: fieldwork (approx. 6000 recordings and 42000 tokens; circa 300 recordings have already been included into a dataset); open sources: audiobooks, radio, TV, etc;
- problem: much time and computing power needed to work with audio data.

```
transcriptions_file = r'/content/drive/MyDrive/os-dataset/transcriptions.txt'
transcriptions_df = pd.read_csv(transcriptions_file, delimiter='\t', header=None, names=["full"])
transcriptions_df["filename"] = transcriptions_df['full'].apply(lambda x: re.split(" ", x)[0])
transcriptions_df["transcriptions"] = transcriptions_df['full'].apply(lambda x:' '.join(re.split(" ", x)[1:]))
transcriptions_df = transcriptions_df.drop(columns=['full'])
print(transcriptions_df)
audio_dir = r'/content/drive/MyDrive/os-dataset/os-audios'
transcriptions_df['filepath'] = transcriptions_df['filename'].apply(lambda x: os.path.join(audio_dir, x))
```

```
                    filename  \
0              ZhM_chi_is.WAV
1           ZhM_ch'iw_is.WAV
2           ZhM_ich'i_is.WAV
3   ZhM_nicherdygonaw_is.WAV
4           ZhM_nichi_is.WAV
..                       ...
66            KG_bajsyn.WAV
67              KG_balc.WAV
68              KG_c'ar.WAV
69             KG_c'ata.WAV
70              KG_c'iw.WAV

                                        transcriptions
0                                          Чи, чи, чи.
1                                     Чъиу, чъиу, чъиу.
2                                     Ичъи, ичъи, ичъи.
3           Ничердыгонау, ничердыгонау, ничердыгонау.
4                                     Ничи, ничи, ничи.
..                                                  ...
66  Байсын, байсын, байсын. Дзуры байсын æртæхатты...
67  Балц, балц, балц. Дзуры балц æртæхатты. Дзуры ...
68  Цъар, цъар, цъар. Дзуры цъар æртæхатты. Дзуры ...
69  Цъата, цъата, цъата. Дзуры цъата æртæхатты. Дз...
70  Цъиу, цъиу, цъиу. Дзуры цъиу æртæхатты. Дзуры ...
```

```
Map: 100% 71/71 [00:04<00:00, 59.42 examples/s]
Dataset({
    features: ['filename', 'transcriptions', 'filepath', 'audio', 'sampling_rate', 'audios'],
    num_rows: 71
})
```

A fragment of dataset based on fieldwork data.