

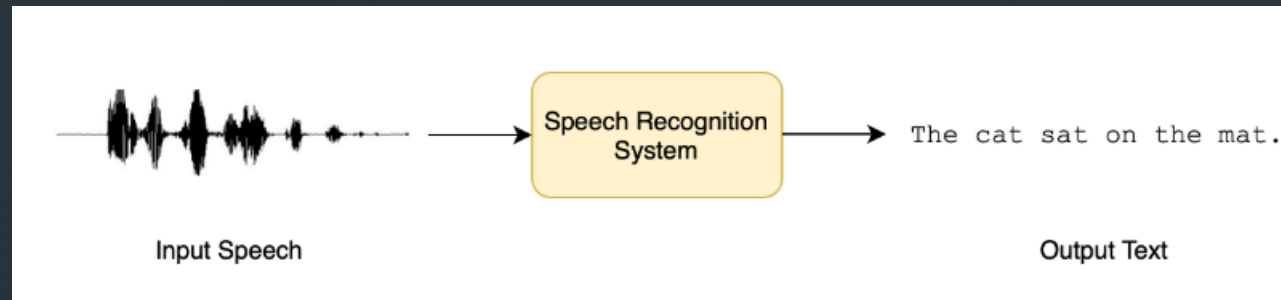
Automatic Speech Recognition for Ossetic language

Varvara Petrova

24.02.2025

General info about ASR

- Goal: transcribing a given audio to text;
- Can be implemented using mono- or multi-lingual models (with growing popularity of the latter)





Overview of existing tools



Datasets

Dataset	Train Hours	Domain	Speaking Style	Casing	Punctuation	License	Recommend Use
LibriSpeech	960	Audiobook	Narrated	✗	✗	CC-BY-4.0	Academic benchmark
Common Voice 11	3000	Wikipedia	Narrated	✓	✓	CC0-1.0	Non-native speakers
VoxPopuli	540	European Parliament	Oratory	✗	✓	CC0	Non-native speakers
TED-LIUM	450	TED talks	Oratory	✗	✗	CC-BY-NC-ND 3.0	Technical topics
GigaSpeech	10000	Audiobook, podcast, YouTube	Narrated, spontaneous	✗	✓	apache-2.0	Robustness over multiple domains
SPGISpeech	5000	Financial meetings	Oratory, spontaneous	✓	✓	User Agreement	Fully formatted transcription
Earnings-22	119	Financial meetings	Oratory, spontaneous	✓	✓	CC-BY-SA-4.0	Diversity of accents
AMI	100	Meetings	Spontaneous	✓	✓	CC-BY-4.0	Noisy speech conditions

Datasets available on HuggingFace.

Dataset example: LibriSpeech

- audiobook recordings;
- (audio, text) correspondences;
- 10 sec audio samples.

2035-147960.trans
2035-147960-0000
2035-147960-0001
2035-147960-0002
2035-147960-0003
2035-147960-0004
2035-147960-0005
2035-147960-0006
2035-147960-0007
2035-147960-0008
2035-147960-0009
2035-147960-0010
2035-147960-0011
2035-147960-0012
2035-147960-0013
2035-147960-0014
2035-147960-0015
2035-147960-0016

2035-147960-0000 SHE WAS FOUR YEARS OLDER THAN I TO BE SURE AND HAD SEEN MORE OF THE WORLD BUT I WAS A BOY AND SHE WAS A GIRL AND I RESENTED HER PROTECTING MANNER
2035-147960-0001 THIS CHANGE CAME ABOUT FROM AN ADVENTURE WE HAD TOGETHER
2035-147960-0002 ONE DAY WHEN I RODE OVER TO THE SHIMERDAS I FOUND ANTONIA STARTING OFF ON FOOT FOR RUSSIAN PETER'S HOUSE TO BORROW A SPADE AMBROSCH NEEDED
2035-147960-0003 THERE HAD BEEN ANOTHER BLACK FROST THE NIGHT BEFORE AND THE AIR WAS CLEAR AND HEADY AS WINE
2035-147960-0004 IT WAS ON ONE OF THESE GRAVEL BEDS THAT I MET MY ADVENTURE
2035-147960-0005 I WHIRLED ROUND AND THERE ON ONE OF THOSE DRY GRAVEL BEDS WAS THE BIGGEST SNAKE I HAD EVER SEEN
2035-147960-0006 I KNOW I AM JUST AWFUL JIM I WAS SO SCARED
2035-147960-0007 I NEVER KNOW YOU WAS SO BRAVE JIM SHE WENT ON COMFORTINGLY
2035-147960-0008 A FAINT FETID SMELL CAME FROM HIM AND A THREAD OF GREEN LIQUID OOOZED FROM HIS CRUSHED HEAD
2035-147960-0009 LOOK TONY THAT'S HIS POISON I SAID
2035-147960-0010 I EXPLAINED TO ANTONIA HOW THIS MEANT THAT HE WAS TWENTY FOUR YEARS OLD THAT HE MUST HAVE BEEN THERE WHEN WHITE MEN FIRST CAME LEFT ON FROM BUFFALO AND INDIAN TIMES
2035-147960-0011 WE DECIDED THAT ANTONIA SHOULD RIDE DUDE HOME AND I WOULD WALK
2035-147960-0012 I FOLLOWED WITH THE SPADE OVER MY SHOULDER DRAGGING MY SNAKE
2035-147960-0013 OTTO FUCHS WAS THE FIRST ONE WE MET
2035-147960-0014 HE COULD STAND RIGHT UP AND TALK TO YOU HE COULD DID HE FIGHT HARD
2035-147960-0015 OTTO WINKED AT ME
2035-147960-0016 A SNAKE OF HIS SIZE IN FIGHTING TRIM WOULD BE MORE THAN ANY BOY COULD HANDLE




Common Voice for Ossetic

- Wikipedia articles;
- Narrated speaking style;
- 1.2 hours;
- 1237 sentences.

Models

	Connectionist Temporal Classification models	Seq2Seq models
Principle	<ol style="list-style-type: none">1. Reading the audio waveform2. Mapping the input into a sequence of hidden states (downsampling to one hidden-state vector/20 ms)3. Getting class label predictions using blank tokens	<p>outputting full words rather than a sequence of individual characters -> no one-to-one correspondence of input and output</p> <p>decoder as a language model -> more powerful approach than CTC</p>
Examples	<p>Wav2Vec: raw audio waveforms as input</p> <p>M-CTC-T: mel spectrograms as input</p> <p>HuBERT: trained to predict “discrete speech units”</p>	<p>Speech2Text</p> <p>Whisper</p>



ASR for Ossetic: Data and research prospect

Ossetic data: General info

- >15,000 tokens (either isolated words or carrier phrase sentences);
- ~6000 audio recordings;
- 17 speakers (9 Iron, 8 Digor), 2 dialects, 4 varieties (Standard Iron, Kudar Iron, Digora Digor, Chikola Digor);
- originally collected for phonetic research in 2023-2024

Ossetic data: Advantages and restrictions

Advantages	Restrictions
+ high-quality audio recordings (Zoom recorder and headset microphone)	- high-quality audio recordings -> poor performance in noisy environments expected
+ representative of inter- and intra-speaker variability	- high (sub)dialectal variation: SI зул [ʒul] 'crooked', дзул [zul] 'bread' KI зул [zul] 'crooked', дзул [ʒul] 'bread'
+ realizations of all phonemes in most positions -> possibly more representative than narrative sample of comparable	- mostly isolated words or words in uniform carrier phrase templates (-> unfit for TTS)
(?) short recordings -> no distortions caused by artificial	- predominantly dictionary forms

Aims and hypotheses

- Fieldwork-based Ossetic audio dataset = main product;
- Fine-tuning foundational speech models.

Hypotheses:

- better results in recognizing segments and sequences with 1) the highest representation 2) no interdialectal variation;
- better results for audios with no background noise;
- no tokenizers trained on datasets of size comparable to English/Russian/etc. -> lesser advantage of decoder models