Leveraging Data Collection and Unsupervised Learning for Code-switched Tunisian Arabic Automatic Speech Recognition

Ahmed Amine Ben Abdallah ⋆ , **Ata Kaboudi**, Amir Kanoun, Salah Zaiem ⋆

April 2024

⋆ First Authors

Tunisia

Background

01

Inspired from many other languages

A Mix of Arabic, French, Englisjh , Italien ,
Turkish , Amazigh

02

No spelling conventions

No Offiicial Reference Rules, Grammar to
Tunisian Arabic Dialect

03

Code-Swiched

One Sentence can contain 3 different languages

Sample

Le probleme هوا الي احنا نحكيو too many languages

Fr Ar En

Failure of Non-Specific Models

	Without Code Switching	With Code Switching
MultiLanguage Models		
MMS 1B All	104.7	102.0
Whisper Large v2	127.7	105.8
Arabic Models		
Wav2vec2.0 Ar.	89.7	96.7
Whisper Large v2 Ar	74.1	85.9

Word Error Rate For Various Models

Data Problem

Lack of Code Switched data

Current datasets Lack code switched aspect and incorporate only Arabic transcripts.

- STAC has 5% Code Switching

Data Problem

Lack of Code Switched data

Current dataset lack the Code Switched aspect and incorporate only arabic transcripts.

- STAC has 5% Code Switching

Domain Specific Data

Previous work is focus on specific domains and lack a broader richnesss of the Tunisian dialect.

- TARIC pertains Conversations from Train Stations

Data Problem

Lack Of Code-Switched Data

Current Dataset Available Lack the Code switched aspect and incorporate only Arabic transcripts.

- STAC has 5% Code Switching

Domain Specific Data

Previous work is focus on specific domains and lack a broader richnesss of the Tunisian dialect.

- TARIC pertains Conversations from Train Stations

Lack Of Open Source Models

Most previous work is proprietary and no open source model was implemented.

01

How to collect and validate Code-Switched data focusing on Tunisian dialect?

01 How to collect and validate Code-Switched data focusing on Tunisian dialect?

02 How to Implement a Code-Switched approach to transcription ?

01 How to collect and validate Code-Switched data focusing on Tunisian dialect?

02 How to Implement a Code-Switched approach to transcription ?

03 How to evaluate ASR for languages with no spelling conventions

Contributions

01

OS Datasets

Introducing TunSpeech and TunSpeech CS: Open Source Datasets for Tunisian ASR

02

OS Models

Introducing two New Speech-to-text models for Tunisian dialect: A dedicated Tunisian model and a Code-Switched model.

03

Human Evaluation

Assess noise caused by lack of spelling conventions through human evaluation

Data Collection

PreProcessing :

Text

1. Removed, diaclitics, punctuation and special characters
2. Latin characters for French and English
   Arabic characters for Arabic

PreProcessing :

Text

1. Removed Diaclitics
2. Latin Characters for French and English
   Arabic Characters for Arabic
3. Seperated Languages with Tags

Audio

1. WebRTC-VAD segmentation
2. MultiSpeaker detection and removal
3. Music detection and Removal

Dataset Overview

Public Datasets

We Leverage a range of public datasets from previous works

TARIC **IWSLT** **STAC**

Tunswitch

Our own custom collected Models

Unlabeled Audios

A Set of unlabeled audios to be used for self superviser approach

Web Scraped Text

we scraped a Tunisian textual corpus to be used in LM finetuning

Public Datasets

TARIC	Dataset of conversations in train stations	**10hours**
STAC	A radio-broadcast-based dataset with slight code-switching	**4hours**
IWSLT	A translation dataset consisting in telephonic conversations.	**160 hours**

Public Datasets

TARIC	Dataset of conversations in train stations	**10hours**
STAC	A radio-broadcast-based dataset with slight code-switching	**4hours**
IWSLT	A translation dataset consisting in telephonic conversations.	**160 hours**

Limitations :

- **Lack of Code-Swiched data**
- **Very domain specific**
- **Limited number of speakers**

TunSwich TO

- TO : Tunisian Only
- Sourced sentences from previous text collections Tunisiya
- Volunteers to record the spoken Sentences
- Collection of 2631 distinct phrases from 82 Speakers

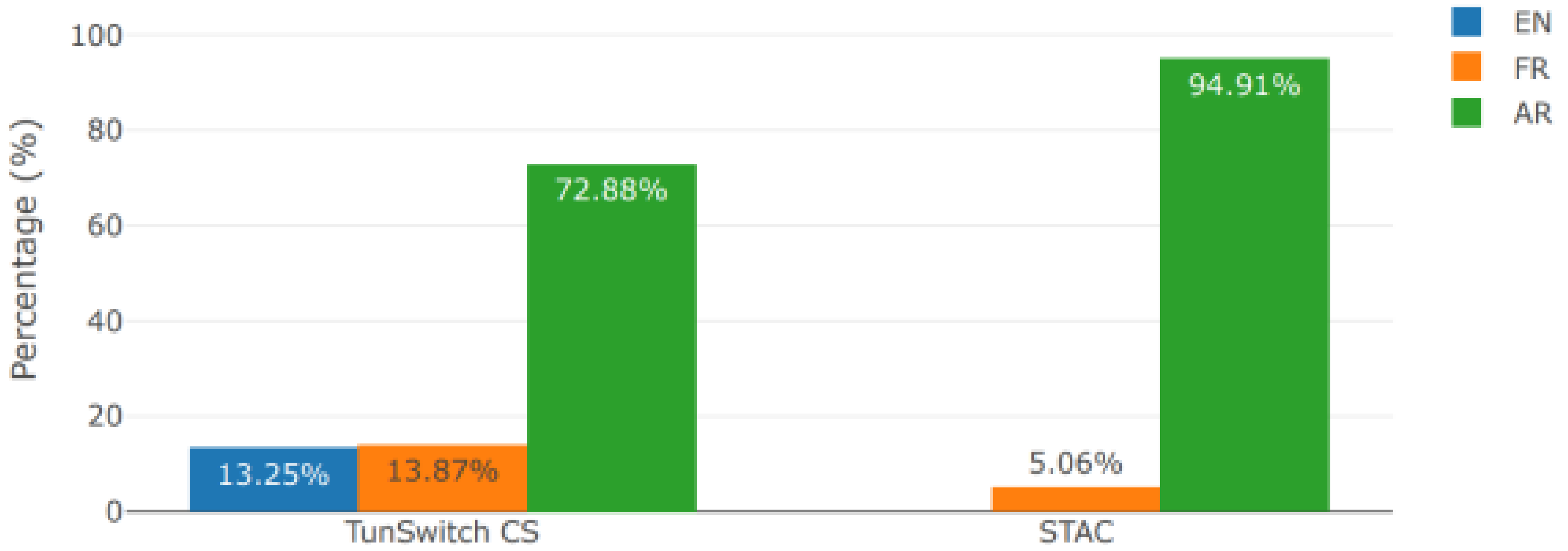
TunSwich CS

- CS: Code Switched
- From spontanous from radio broadcast

TunSwich CS

- CS: Code Switched
- From spontanous from radio broadcast
- Manual annotation with instruction set to maintain consistancy and Doccano Annotation Tool
- 9 Hours Dataset

TunSwich CS



Comparative Code Switching Percentage

Unlabelled Data

- For a Self-supervised approach
- Collection of national TV shows videos
- 153 hours after pre-processing

Scraped Text Corpus

- Scraped from various website to finetune LM

Our Data

	Dataset	PROSODY	CODE-SWITCHING	TRAIN(H)	Dev(H)	Test(H)	Labeled
	IWSLT	Spontaneous	x	151h 24m 47s	4h 55m 51s	4h 36m 28s	Y
	STAC	Spontaneous	Y	2h 29m 8s	n/A	n/a	Y
	TARIC	Spontaneous	x	9h 25m 44s	17 m 29s	12m 5s	Y
Our Contribution	**TunSwitch TO**	Read	X	2h 29m 29s	4m 25s	23m 39s	Y
Our Contribution	**TunSwitch CS**	Spontaneous	Y	8h 15m 35s	15m 43s	25m 12s	Y
Our Contribution	**SelfSwitch TO**	Spontaneous	Y	153h 18m 22s	n/a	n/a	X

Models

BASE MODEL

ENCODER, FINE-TUNABLE

WavLM Large

This model produced a better performance than wav2vec2.0 XLSR eventhought it is trained only on english

→

DECODER

MLP

Three dense layers with LeakyReLU activations, and batch normalization between layers, and is trained with Connectionist Temporal Classification (CTC) loss

→

LANGUAGE MODEL

4-gram LM

Candidate sentences are rescored using a 4-gram language model trained with the KenLM toolkit and implemented with the PyCTCDecode library

Language Modelling Options

01

Without LM

We will not use a language model to rescore output probablities

02

With inDomain LM

We utilize Languages model within the training data

03

With OutDomain LM

We leverage a language model outside of training scope

Results : Tunisian Only

	TARIC		IWSLT		Tunswtich TO	
	CER	WER	CER	WER	CER	WER
Previous Wok	-	22.6	-	41.5	-	-
	CER	WER	CER	WER	CER	WER
WithoutLM	6.44	12.84	20.28	42.74	13.34	41.45
With inDomain LM	6.23	10.81	20.27	38.8	12.5	36.1
With OutDomain LM	6.13	10.5	20.32	39.01	10.08	26.64

Self Training

The Straightforward naive approach

1

Use previously trained model to generate pseudo-labels

Unlabeled Data , 153h from national TV shows

2

Retrain model from scratch on both data and pseudo_data

Pseudolabeled Data

3

Comparative assement to evaluate Improvements

Results : Tunisian Only with Self-Training

	TARIC		IWSLT		Tunswtich TO	
	CER	WER	CER	WER	CER	WER
Previous Wok	-	22.6	-	41.5	-	-
Without Self-Training	CER	WER	CER	WER	CER	WER
WithoutLM	6.44	12.84	20.28	42.74	13.34	41.45
WithinDomainLm	6.23	10.81	20.27	38.8	12.5	36.1
With OutDomainLM	6.13	10.5	20.32	39.01	10.08	26.64
With Self-Training	CER	WER	CER	WER	CER	WER
Without LM	6.3	11.82	20.49	42.49	12.65	38.25
With InDomainLm	6.29	10.83	21.18	39.46	12.42	36.07
With OutDomainLM	6.22	10.5	21.18	39.53	9.67	25.54

Code-Switched Approach

Few Shot Code Switching

Tunisian Model

Tunswitch Model

Best Model From Previous Experiences

English Model

Speechbrain/asr-wav2vec2-commoncoive en

Acoustic model (wav2vec2.0 + CTC). A pretrained wav2vec 2.0 model is combined with two DNN layers and finetuned on CommonVoice EN. withCTC greedy decoder.

French Model

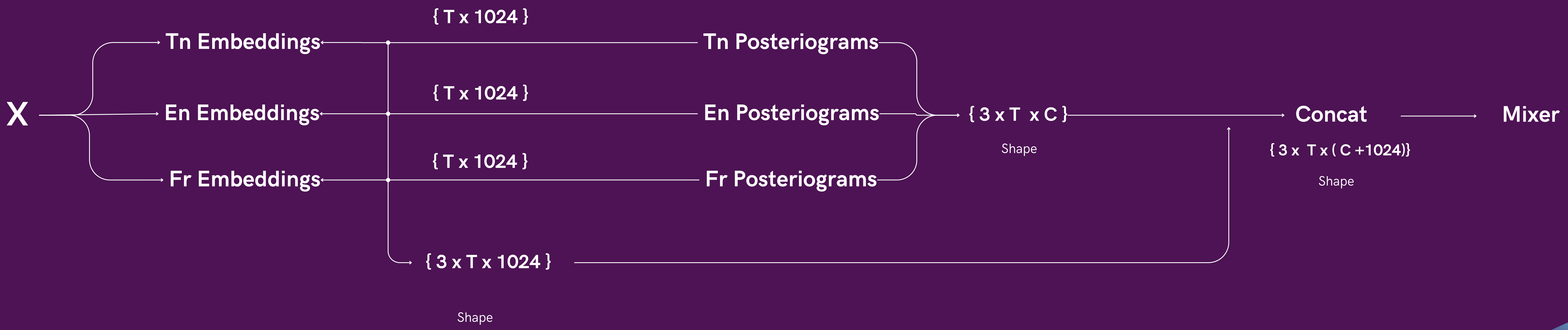
Speechbrain/asr-wav2vec2-commoncoive fr

Acoustic model (wav2vec2.0 + CTC). A pretrained wav2vec 2.0 model is combined with two DNN layers and finetuned on CommonVoice FR. withCTC greedy decoder.

Mixer

Custom BiLSTM Model

Few-Shot Code-Switching



T : Frames

C : Character Count

Language modelling options

01

Without LM

02

With TunisanOnly LM

Use a Tunisian only corpora for reference

03

With Code-Swiched LM

Use a Code-Switched , French ,English And Arabic corpora

04

With EN-FR Enriched LM

Enrich corpora with 10 K English and French monolingual sentences

Code Switching Results

	TUNSWITCH CS	
	CER	WER
Without LM	13.71	40.65
With TunisanOnly LM	17.57	47.45
With CodeSwiched LM	12.77	30.41
With EN-FR enriched LM	10.5	29.47

Evaluation

Problem

Reference	Prediction
جيبلي كاس ما	جيبلي كاس ماء
معناها فما واقع مرير لحقيقة لي يعيشوا فيه هوما	معناها فما واقع مرير الحقيقة لي يعيشوا فيه هوما

Problem

No spelling convention
Observation is consistant across many samples

Reference	Prediction
جيبلي كاس ما	جيبلي كاس ماء
معناها فما واقع مرير لحقيقة لي يعيشوا فيه هوما	معناها فما واقع مرير الحقيقة لي يعيشوا فيه هوما

Sorry to non arabic speakers

Human Evaluation

Sentence Error Rate (SER)

Annotators

25 Tunsian annotators, all fluent English and French speakers. tasked with 50 audios each.

Human Evaluation

Sentence Error Rate (SER)

Annotators

25 Tunsian annotators, all fluent English and French speakers. tasked with 50 audios each.

Evaluation

Documents detailing the process were distributed, simplifying decisions to binary for each text output. A text is valid only if approved by two annotators. Evaluate Both TO and CS.

Human Evaluation

Sentence Error Rate (SER)

	TunSwitch TO	TunSwitch CS
Automatic SER	76.45	96%
Human SER	34%	66%

Observations

Human evaluation lowers SER by 42% and 29%, with an 80% annotator agreement, highlighting WER underrates model performance."

Conclusion

01 We propose A Tunisian Only and Code Switched Dataset For Tunisian ASR

02 We propose a Tunisian only model and a code switched model for Tunisian ASR

03 We explore automatic and human evaluation for non standardized languages

Future Improvements

01 Pretrain WavLm from scratch

02 Expand Code-Switched dataset To improve results

03 Explore more advanced semi supervised techniques

References

MMS 1B All Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli, “Scaling speech technology to 1,000+ languages,” arXiv, 2023.

Wav2vec2 Ar Jonatas Grosman, “Fine-tuned XLSR-53 large model for speech recognition in Arabic,” 2021.

Whisper Large v2 Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, “Robust speech recognition via large-scale weak supervision,” 2022

Tunisiya Karen McNeil, “Tunisian arabic corpus: Creating a written corpus of an ‘unwritten’language,” Arabic corpus linguistics, vol. 30, 2018.

Open Source Resources

DataSets

Models

Paper

The Team

Salah Zaim

phD in Speech Processing and Machine Learning , Telecom Paris;
Paris

Ahmed Amine Ben Abdallah

Machine Learning Engineer ABSHORE
Tunisia

Ata Kaboudi

Masters at University of Michigan
Software Engineer at CBRE.
Founding Enginner at Memorality
Michigan

Ahmed Amine Kanoun

Chief of Projects at ABSHORE
Tunisia

Thank You

Questions ?

Scan For My Contact Details