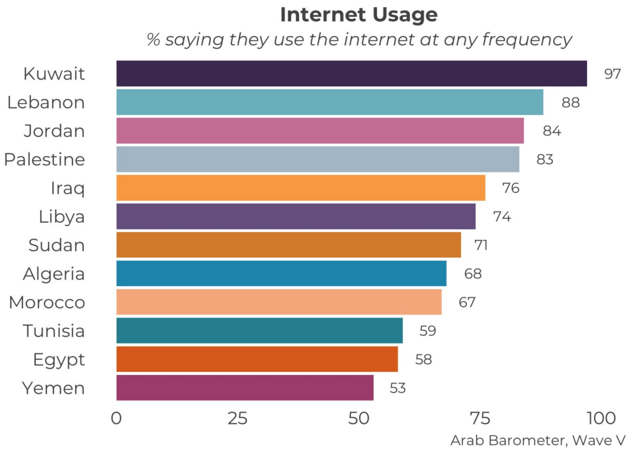
Arabic Spoken Dialect Identification (ASDI - قصدي)

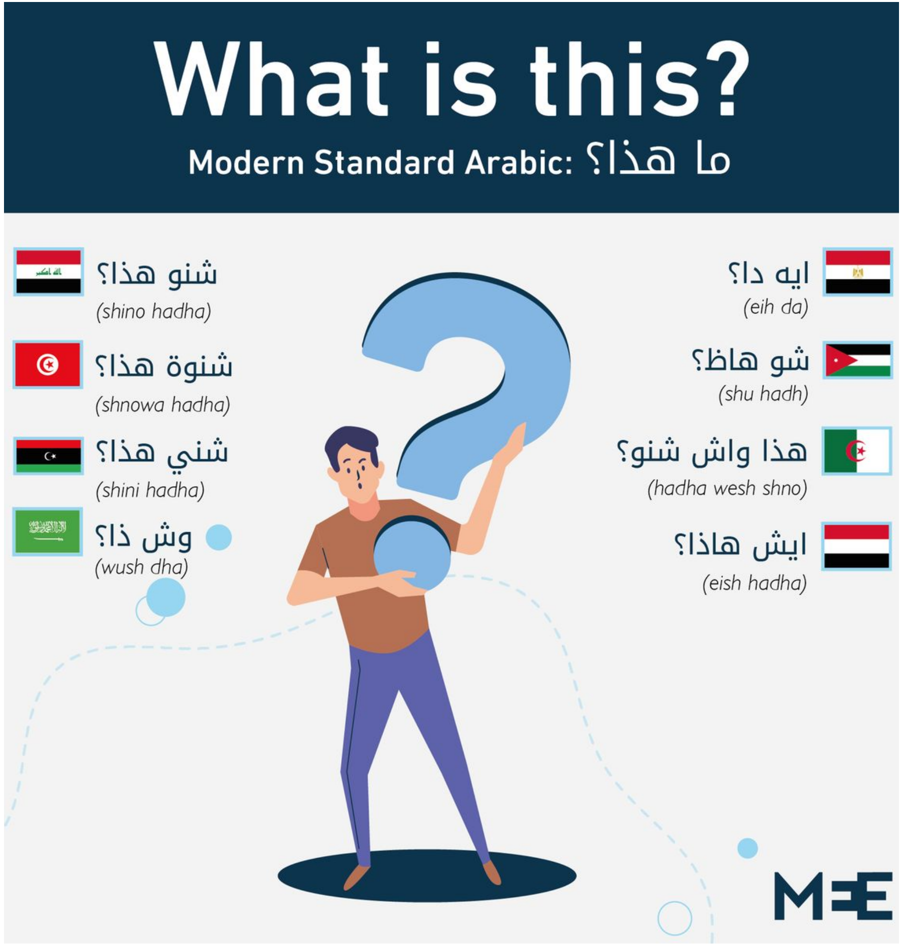
Abir Messaoudi, Mayssa Kchaou and Chayma Fourati

Mentor: Desh Raj

Motivation

- Many different dialects/sub-dialects.
- **Similarities** and **difference** between dialects.
- On digital channels, Arabic speakers express themselves better in **their own local dialect** & by their **own voices** instead of textual comments.





it is important that this large part of the population understands the transmitted content !

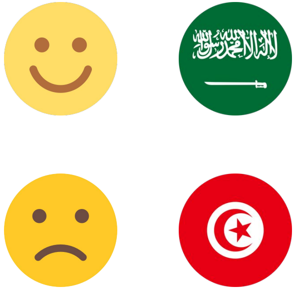
Challenges
Machine Translation

Different meanings when switching dialects!

انت كشخة

You are pretty

You are ugly



Challenges

Automatic Speech recognition

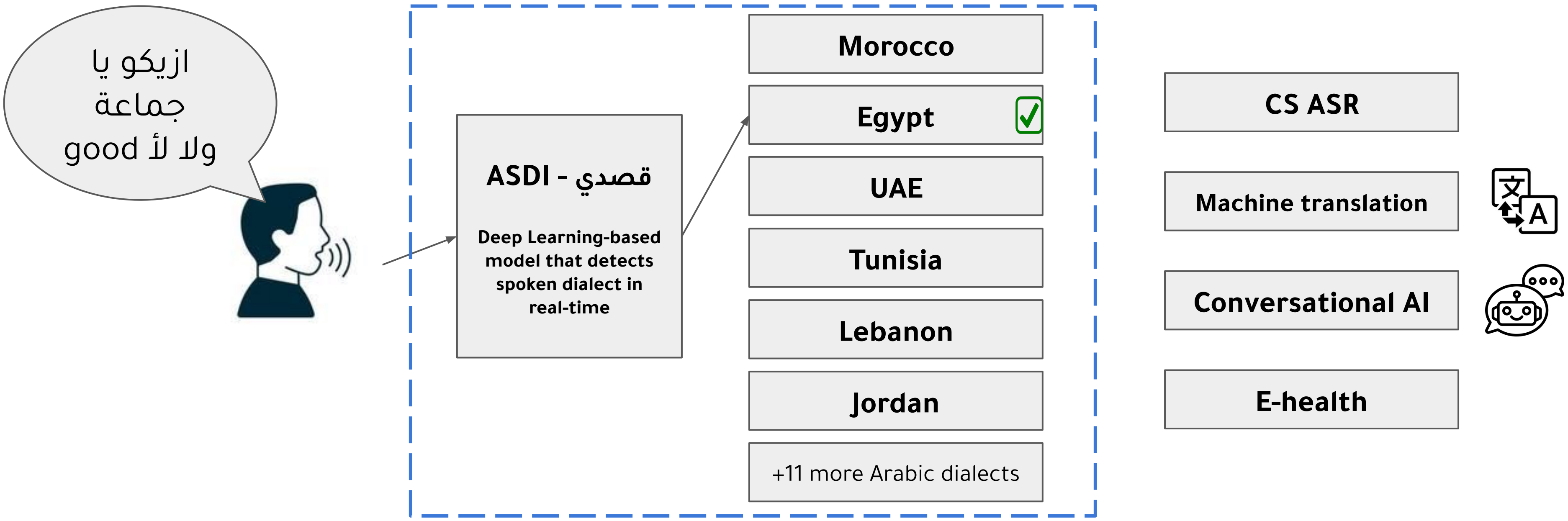
Arabic speakers tend to use **code-switching** in their daily conversation.

Existing ASR models (on Facebook, TikTok..):

- skip code-switching speech,
- Wrong transcription.

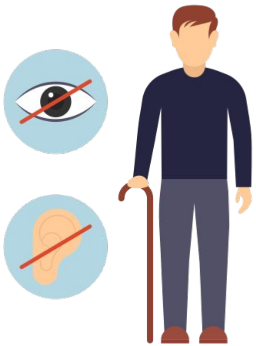
Automatically generated transcripts and subtitles can be almost error-free **if we recognize dialects**.

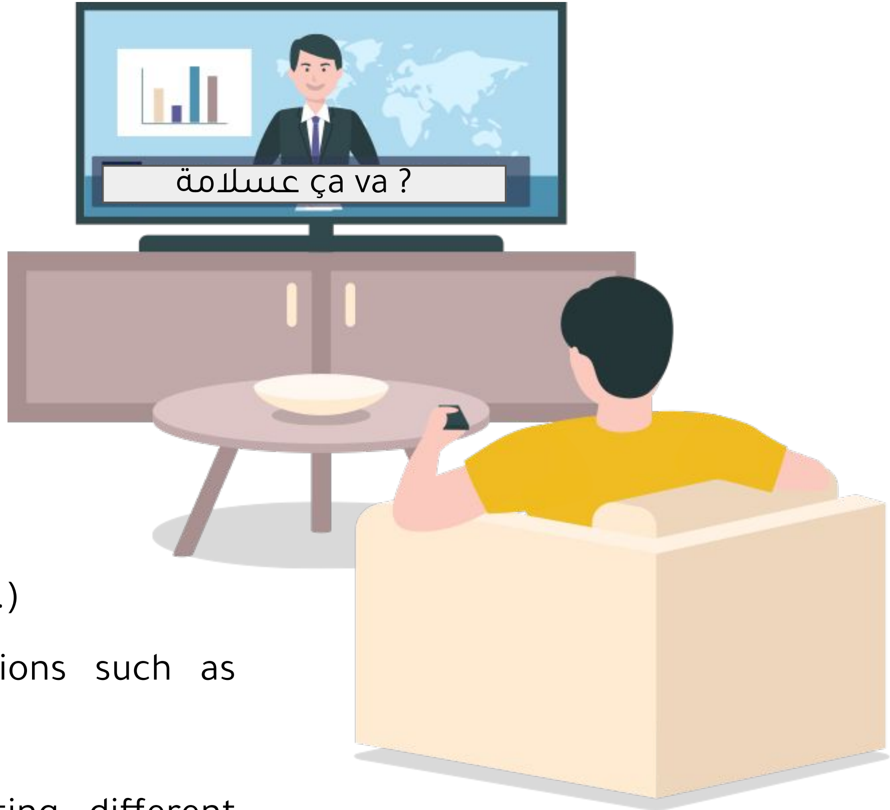
Solution



- Facilitate the speech interaction between arab nations by recognizing the speaker's dialect!
- It will be the front of multidialectal applications including: ASR, machine translation, voicebots and e-health applications!

Impact





1. Include disabled people. (deaf, blind) & illiterate.
2. Enhance customer services (banks, hotels, airports..)
3. Overcome language barriers in high-risk situations such as hospitals and courts.
4. Dialect maintenance: Identifying and documenting different dialects can help to preserve linguistic diversity and **prevent Arabic dialects from dying out!**

ASDI dataset

- **Up-to-date**:
  - Dialects are not static; they vary across space and time.
  - New generations use phrases and vocabulary that were not in existence or use in the past.

- **Code-switching:**
  - Different accents of English and French.
  - More than 30% of CS data. (Intra, inter-Sentential and intra-word.)

- **Reliable**:
  - Didn't rely on automatic annotation.
  - Annotated by Arabic native speakers!

ASDI dataset

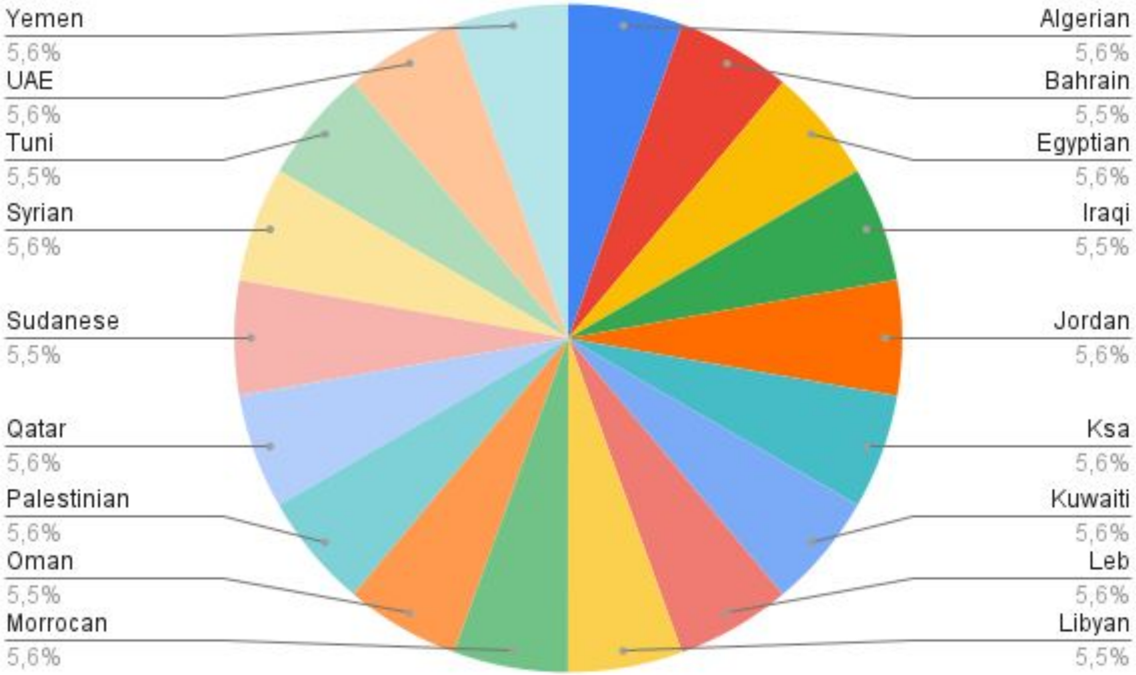
- **Sources**:
  - Podcast platforms,
  - Youtube (vlogs, podcasts..)
  - Series and movies.
  - Public Arabic Speech datasets (MGB-3, Dvoice..).

- **Domains**:
  - Politics
  - Entertainment
  - Education
  - Culture
  - Sports
  - Customer services

ASDI dataset

Total size: over **250** hours.

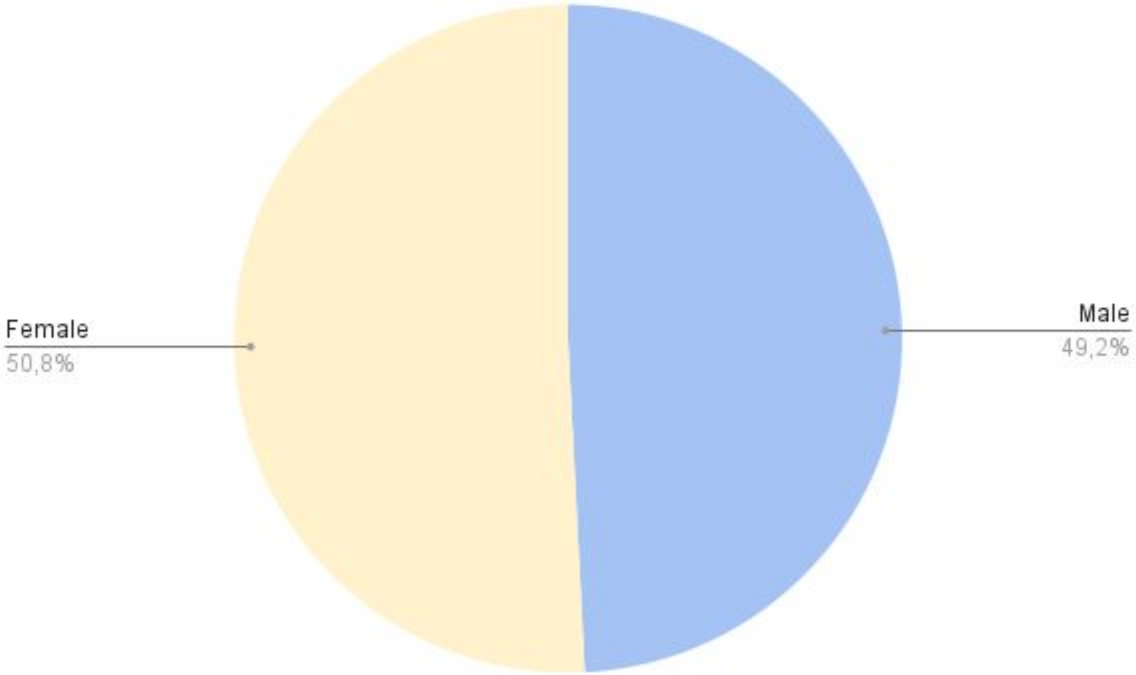
Balanced dataset with ~ 10 hours per dialect.



Distribution of dialects in ASDI dataset.

ASDI dataset

- 937 speakers
- 137k audios, between 2sec and 10 sec.
- Splitted into Train and validation sets.



Distribution of and Male/Female in ASDI dataset

ASDI system
Roadmap

1. Finetune ASDI dataset using an already fine-tuned Arabic wav2vec 2.0 model on ASR, **to better determine the context representations for the input audios**.

WER	CER
23.4995	**8.7133**

Results of fine-tuning on Arabic ASR

   a. Using Noisy data. (background music, noise..)
   b. Using Clean data.

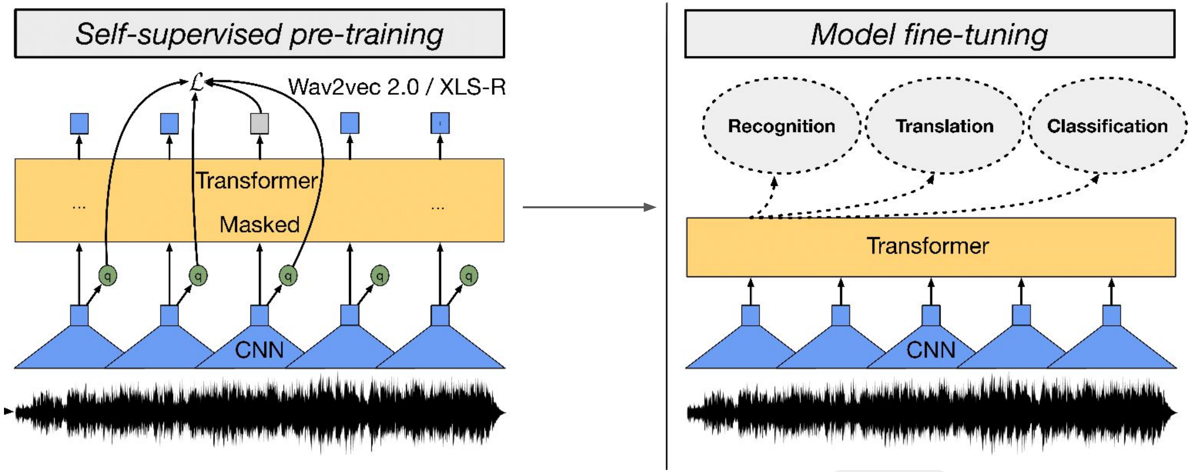
ASDI system

Roadmap

2. Pre-train a Multidialectal Acoustic model.

Finetune it on ASDI For dialect Identification

Unlabeled speech: ADI 17 + common voice



ASDI labeled data

Wav2vec 2.0 architecture [1]

[1] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. CoRR, abs/2006.11477.

Thank you.

Code switching snapshot from dataset

Algerian English

Algerian French

Lebanese English

Tunisian French