Arabic Forced Alignment: From WebMAUS to Whisper and wav2vec2

Jalal Al-Tamimi

Université Paris Cité - Laboratoire de linguistique formelle (LLF) - UMR-7110

11th RJCP (Rencontres Jeunes Chercheurs en Parole) - Workshop TAL (LLF)

5 November 2025

Overview

Overview

Motivations

- Rationale
  - Lack of open source and accessible transcribed and time-aligned multidialectal Arabic dataset
  - Lack of diacritised Arabic script
  - Inaccessibility of some romanisation/transliteration systems

Motivations

- Rationale
  - Lack of open source and accessible transcribed and time-aligned multidialectal Arabic dataset
  - Lack of diacritised Arabic script
  - Inaccessibility of some romanisation/transliteration systems
- Led to the development of the first free, open-source, accessible and GDPR compliant romanisation and forced-alignment systems: Arabic WebMAUS Al-Tamimi et al., 2022

Motivations

- Rationale
  - Lack of open source and accessible transcribed and time-aligned multidialectal Arabic dataset
  - Lack of diacritised Arabic script
  - Inaccessibility of some romanisation/transliteration systems
- Led to the development of the first free, open-source, accessible and GDPR compliant romanisation and forced-alignment systems: Arabic WebMAUS Al-Tamimi et al., 2022
- I'll present the Arabic WebMAUS system; how it was built, data used, what it can do, and issues

Motivations

- Rationale
  - Lack of open source and accessible transcribed and time-aligned multidialectal Arabic dataset
  - Lack of diacritised Arabic script
  - Inaccessibility of some romanisation/transliteration systems
- Led to the development of the first free, open-source, accessible and GDPR compliant romanisation and forced-alignment systems: Arabic WebMAUS Al-Tamimi et al., 2022
- I'll present the Arabic WebMAUS system; how it was built, data used, what it can do, and issues
- I'll end with current developments and remaining to do.

Collaborators and student support

Jalal Al-Tamimi → ATR system; Jordanian I, Moroccan, Lebanese, Levantine; Coordination; Funding

Collaborators and student support

Jalal Al-Tamimi → ATR system; Jordanian I, Moroccan, Lebanese, Levantine; Coordination; Funding

Florian Schiel:
WebMAUS
development

Collaborators and student support

Jalal Al-Tamimi $\rightarrow$ ATR system; Jordanian I, Moroccan, Lebanese, Levantine; Coordination; Funding

Florian Schiel:
WebMAUS
development

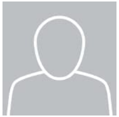
Ghada Khattab:
Lebanese/Levantine
datasets

Navdeep Sokhey:
Bahraini/Egyptian I
datasets

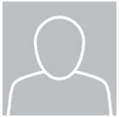
Djegdjiga Amazouz:
Algerian dataset

Abdulrahman
Dallak:
Saudi dataset I

Hajar Moussa:
Saudi dataset II



Khalid Alsubaie:
Saudi dataset III



Collaborators and student support

Jalal Al-Tamimi → ATR system; Jordanian I, Moroccan, Lebanese, Levantine; Coordination; Funding

Florian Schiel: WebMAUS development

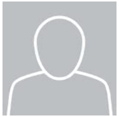
Ghada Khattab: Lebanese/Levantine datasets

Navdeep Sokhey: Bahraini/Egyptian I datasets

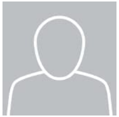
Djegdjiga Amazouz: Algerian dataset

Abdulrahman Dallak: Saudi dataset I

Hajar Moussa: Saudi dataset II



Khalid Alsubaie: Saudi dataset III



Omnia Ibrahim: Egyptian dataset II

Mohammad Abuoudeh: Jordanian dataset II

Collaborators and student support

Jalal Al-Tamimi → ATR system; Jordanian I, Moroccan, Lebanese, Levantine; Coordination; Funding

Florian Schiel: WebMAUS development

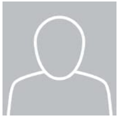
Ghada Khattab: Lebanese/Levantine datasets

Navdeep Sokhey: Bahraini/Egyptian I datasets

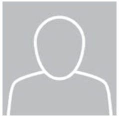
Djegdjiga Amazouz: Algerian dataset

Abdulrahman Dallak: Saudi dataset I

Hajar Moussa: Saudi dataset II



Khalid Alsubaie: Saudi dataset III

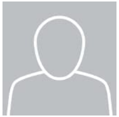


Omnia Ibrahim: Egyptian dataset II

Mohammad Abuoudeh: Jordanian dataset II

Míša Hejná: Verification alignment

Wael Almurashi: Verification alignment



Rana Almbark: Verification alignment

Ourooba Shetewi: Verification alignment

Collaborators and student support

Jalal Al-Tamimi → ATR system; Jordanian I, Moroccan, Lebanese, Levantine; Coordination; Funding

Florian Schiel: WebMAUS development

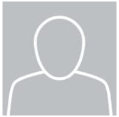
Ghada Khattab: Lebanese/Levantine datasets

Navdeep Sokhey: Bahraini/Egyptian I datasets

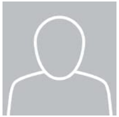
Djegdjiga Amazouz: Algerian dataset

Abdulrahman Dallak: Saudi dataset I

Hajar Moussa: Saudi dataset II



Khalid Alsubaie: Saudi dataset III

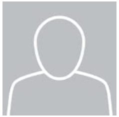


Omnia Ibrahim: Egyptian dataset II

Mohammad Abuoudeh: Jordanian dataset II

Míša Hejná: Verification alignment

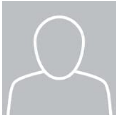
Wael Almurashi: Verification alignment



Rana Almbark: Verification alignment

Ourooba Shetewi: Verification alignment

Amina Djarfi: Automatic transcription



Younes Maatallaoui: whisper, wav2vec2, diacritization, API

Ludivine Huchin: wav2vec2, webMAUS, API

Shuhua Cao: wav2vec2, webMAUS, API

Alexandre Gallot: Data processing, wav2vec2

Overview

WebMAUS - BAS Webservices

- WebMAUS (“BAS WebServices”) ⇒ suite of webservices, free for academic users
- Comprises around speech and language processing tools (Kisler et al., 2017)
- Since its introduction in 2013, roughly 17 million media files (April 2022); likely over 20 million now!
- A powerful pipeline framework allows concatenation of several individual services
  - Automatic phonetic and syllabic segmentation
  - Labelling of a speech recording is first performed using Automatic Speech Recognition (ASR)
  - Text-to-phoneme translation, the WebMAUS engine and a Syllabification service in one processing call.
  - Etc..
- Additional tools, e.g., speaker diarisation, speech enhancement, noise reduction, automatic transcription (Google Cloud services; local installation of whisperX), etc.

WebMAUS - BAS Webservices

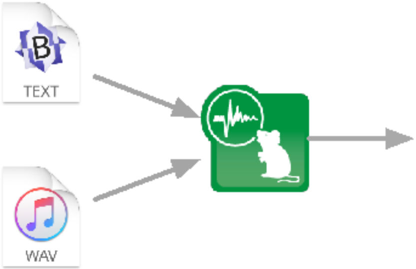
- WebMAUS ("BAS WebServices") ⇒ suite of webservices, free for academic users
- Comprises around speech and language processing tools (Kisler et al., 2017)
- Since its introduction in 2013, roughly 17 million media files (April 2022); likely over 20 million now!
- A powerful pipeline framework allows concatenation of several individual services
  - Automatic phonetic and syllabic segmentation
  - Labelling of a speech recording is first performed using Automatic Speech Recognition (ASR)
  - Text-to-phoneme translation, the WebMAUS engine and a Syllabification service in one processing call.
  - Etc..
- Additional tools, e.g., speaker diarisation, speech enhancement, noise reduction, automatic transcription (Google Cloud services; local installation of whisperX), etc.
- Arabic was not part of WebMAUS services ⇒ No accessible open-access transcribed and time-aligned datasets

Arabic WebMAUS

- To allow inclusion of Arabic to WebMAUS, various steps were required
  - A Grapheme-2-Phoneme conversion tool from text-based transcriptions
  - Various language-specific acoustic models (Arabic WebMAUS)

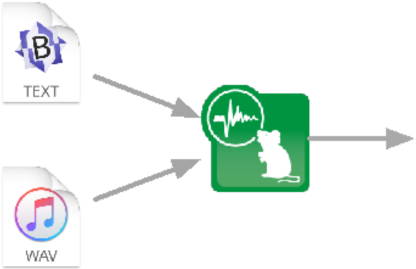
Arabic WebMAUS

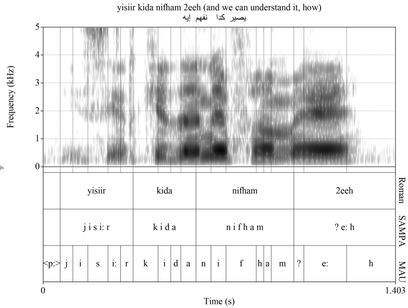
- To allow inclusion of Arabic to WebMAUS, various steps were required
  - A Grapheme-2-Phoneme conversion tool from text-based transcriptions
  - Various language-specific acoustic models (Arabic WebMAUS)



Arabic WebMAUS

- To allow inclusion of Arabic to WebMAUS, various steps were required
  - A Grapheme-2-Phoneme conversion tool from text-based transcriptions
  - Various language-specific acoustic models (Arabic WebMAUS)





- We developed both to answer this

ATR Romanisation system I

Arabic script $\Rightarrow$ transcriptions of only consonants and long vowels; Vowelisations (or diacritisation) of the short vowels is generally optional because it can be predictable based on the utterance meaning; Issues

- Forced-alignment systems require a perfect match between character and phoneme

  Solution $\rightarrow$ A dictionary with specific lexical items associated with specific phonemic transcriptions.

  $\rightarrow$ Standard Arabic lacks a common and a standardized Romanisation system

ATR Romanisation system I

Arabic script ⇒ transcriptions of only consonants and long vowels; Vowelisations (or diacritisation) of the short vowels is generally optional because it can be predictable based on the utterance meaning; Issues

- Forced-alignment systems require a perfect match between character and phoneme

  Solution → A dictionary with specific lexical items associated with specific phonemic transcriptions.

  → Standard Arabic lacks a common and a standardized Romanisation system
- Dialectal Arabic ⇒ spoken-only varieties with no specific written script available (except in some cases)

  No specific standardised written system; no diacritisation (even when using Google Cloud services)

  Multiple spelling variations accepted for a particular word item → even a langauge model would fail to account for all intricacies

ATR Romanisation system I

Arabic script $\Rightarrow$ transcriptions of only consonants and long vowels; Vowelisations (or diacritisation) of the short vowels is generally optional because it can be predictable based on the utterance meaning; Issues

- Forced-alignment systems require a perfect match between character and phoneme

  Solution $\rightarrow$ A dictionary with specific lexical items associated with specific phonemic transcriptions.

  $\rightarrow$ Standard Arabic lacks a common and a standardized Romanisation system
- Dialectal Arabic $\Rightarrow$ spoken-only varieties with no specific written script available (except in some cases)

  No specific standardised written system; no diacritisation (even when using Google Cloud services)

  Multiple spelling variations accepted for a particular word item $\rightarrow$ even a langauge model would fail to account for all intricacies
- Current systems not adequate:

  Arabizi $\Rightarrow$ generic; multiple sounds = same symbol $\rightarrow$ e.g., symbols '2' and 'a' for the letter hamza (for IPA /ʔ/; X-SAMPA '?')

  Buckwalter Arabic translator of Arabic script to romanized symbols $\Rightarrow$ although it allows for vowelisation of short vowels, these are unfortunately rarely transcribed in the orthographic transcriptions

ATR Romanisation system II

Our solution

- Develop a phonetically-based orthographic transcription of spoken speech
- Transparent and direct match between sounds and orthography with a 1-to-1 match between a produced sound and a symbol to transcribe it, using ASCII characters.
- ATR contains 98 phonemes $\Rightarrow$ covers all possible sounds present in the various varieties including standard

ATR Romanisation system II

Our solution

- Develop a phonetically-based orthographic transcription of spoken speech
- Transparent and direct match between sounds and orthography with a 1-to-1 match between a produced sound and a symbol to transcribe it, using ASCII characters.
- ATR contains 98 phonemes $\Rightarrow$ covers all possible sounds present in the various varieties including standard
- Many phonetic variants included, e.g.,
  - All MSA Cs (singleton and geminates) and Vs (short and long)
  - Phonemes /zˤ/ and /dˤ ðˤ/, /lˤ/
  - Variants of /x χ/ or /ɣ ʁ/
  - Phonemes /q g/
  - Phonemes /ʧ ʤ/
  - 12 long + 12 short vowels
  - Etc..

ATR Romanisation system III

49 phonemes ⇒ Geminates + long vowels table (*2 for singleton and short vowels).

IPA	ATR System	X-SAMPA
ʔʔ	22	??
bb	bb	bb
tt	tt	tt
θθ	t\t\	TT
ʒʒ	jj	ZZ
ħħ	HH	X\
xx	xx	xx
χχ	XX	XX
dd	dd	dd
ðð	d\d\	DD
rr	rr	rr
zz	zz	zz
ss	ss	ss
ʃʃ	s\s\	SS
sˤsˤ	SS	s_?\s_?\
dˤdˤ	DD	d_?\d_?\
tˤtˤ	TT	t_?\t_?\
ðˤðˤ	D\D\	D_?\D_?\
zˤzˤ	ZZ	z_?\z_?\

IPA	ATR System	X-SAMPA
lˤlˤ	LL	l_?\l_?\
ʕʕ	33	?\
ɣɣ	GG	GG
ʁʁ	G\G\	G\G\
ff	ff	ff
qq	qq	qq
gg	gg	gg
kk	kk	kk
ll	ll	ll
mm	mm	mm
nn	nn	nn
hh	hh	hh
ww	ww	ww
jj	yy	jj
ʧʧ	chch	tStS
ʤʤ	djdj	dZdZ
vv	vv	vv
pp	pp	pp

IPA	ATR System	X-SAMPA
iː	ii	i:
ɪː	II	I:
eː	ee	e:
ɛː	EE	E:
æː	aeae	{:
aː	aa	a:
ɑː	AA	A:
ɔː	OO	O:
oː	oo	o:
uː	uu	u:
ʊː	UU	U:
əː	@@	@:

Arabic WebMAUS I

Used the WebMAUS technique Schiel (1999) ⇒ HMM-based ASR systems based on MFCCs features obtained from audio signals using Gaussian Mixture Models (GMM)

- Acoustic model ⇒ estimates the posterior probability for a phone class given a segment of speech
- Pronunciation (language) model ⇒ estimates the probability of a sequence of spoken phones
- AM ⇒ 98 phoneme classes to represent nearly all Arabic varieties
- Ideally ⇒
  - Verified segmented and labelled training set of speech recordings
  - Enough samples of each phoneme class from every Arabic variety
  - Spoken by at least 50 native speakers of both sexes

Arabic WebMAUS I

Used the WebMAUS technique Schiel (1999) ⇒ HMM-based ASR systems based on MFCCs features obtained from audio signals using Gaussian Mixture Models (GMM)

- Acoustic model ⇒ estimates the posterior probability for a phone class given a segment of speech
- Pronunciation (language) model ⇒ estimates the probability of a sequence of spoken phones
- AM ⇒ 98 phoneme classes to represent nearly all Arabic varieties
- Ideally ⇒
  - Verified segmented and labelled training set of speech recordings
  - Enough samples of each phoneme class from every Arabic variety
  - Spoken by at least 50 native speakers of both sexes
  - Issue ⇒ no publicly available resources to fulfil these requirements!

Arabic WebMAUS I

Used the WebMAUS technique Schiel (1999) ⇒ HMM-based ASR systems based on MFCCs features obtained from audio signals using Gaussian Mixture Models (GMM)

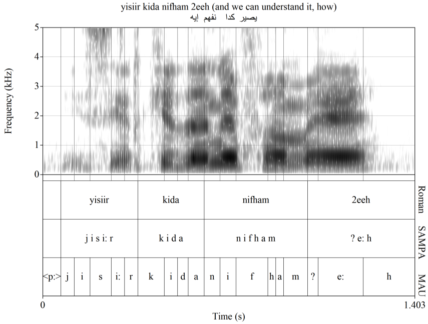
- Acoustic model ⇒ estimates the posterior probability for a phone class given a segment of speech
- Pronunciation (language) model ⇒ estimates the probability of a sequence of spoken phones
- AM ⇒ 98 phoneme classes to represent nearly all Arabic varieties
- Ideally ⇒
  - Verified segmented and labelled training set of speech recordings
  - Enough samples of each phoneme class from every Arabic variety
  - Spoken by at least 50 native speakers of both sexes
  - Issue ⇒ no publicly available resources to fulfil these requirements!
- Solution ⇒
  - Collected speech recordings from various Arabic varieties.
  - Bahraini, Saudi Arabian, Lebanese, Levantine (comprised of Lebanese, Syrian, Palestinian Arabic).
  - Recordings + transcriptions ⇒ collected, unified and merged into a common annotation format
  - Automatically segmented using the language-independent system of WebMAUS + manual verification
  - Time-aligned speech signal + orthographic/phonetic transliteration and segmentation
  - Transcription convention ⇒ broad phonetic transcription of the incoming signal; accommodated within the ATR system ⇒ Optimal bottom-up approach to the transcription which relied on what speakers said and how they said it

Arabic WebMAUS I

Used the WebMAUS technique Schiel (1999) ⇒ HMM-based ASR systems based on MFCCs features obtained from audio signals using Gaussian Mixture Models (GMM)

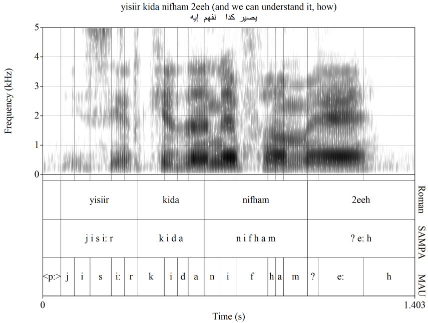
- Acoustic model ⇒ estimates the posterior probability for a phone class given a segment of speech
- Pronunciation (language) model ⇒ estimates the probability of a sequence of spoken phones
- AM ⇒ 98 phoneme classes to represent nearly all Arabic varieties
- Ideally ⇒
  - Verified segmented and labelled training set of speech recordings
  - Enough samples of each phoneme class from every Arabic variety
  - Spoken by at least 50 native speakers of both sexes
  - Issue ⇒ no publicly available resources to fulfil these requirements!
- Solution ⇒
  - Collected speech recordings from various Arabic varieties.
  - Bahraini, Saudi Arabian, Lebanese, Levantine (comprised of Lebanese, Syrian, Palestinian Arabic).
  - Recordings + transcriptions ⇒ collected, unified and merged into a common annotation format
  - Automatically segmented using the language-independent system of WebMAUS + manual verification
  - Time-aligned speech signal + orthographic/phonetic transliteration and segmentation
  - Transcription convention ⇒ broad phonetic transcription of the incoming signal; accommodated within the ATR system ⇒ Optimal bottom-up approach to the transcription which relied on what speakers said and how they said it
- Arabic WebMAUS (version 2) ⇒ 6610 recordings, from 94 speakers, with a total duration of 16h10min and 509804 labelled phone segments

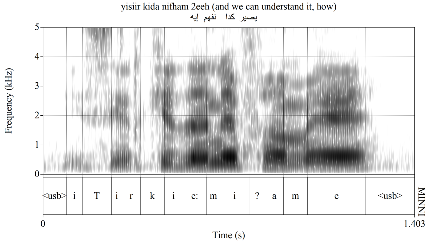
Arabic WebMAUS II



Sentence

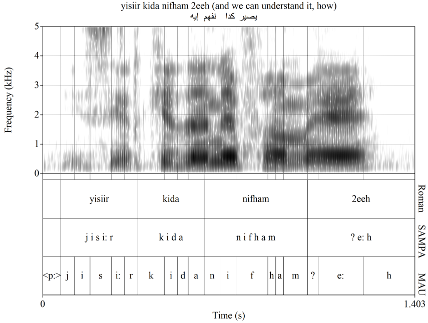
Arabic WebMAUS II

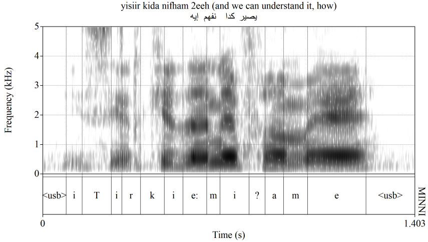




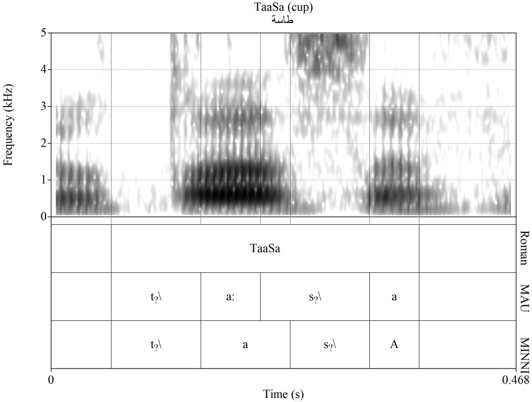
Sentence

Arabic WebMAUS II





Sentence



Word

Arabic WebMAUS III

- Performance
  - Arabic WebMAUS $\Rightarrow \approx 95\%$ accuracy at 20ms, comparable to other systems; increasing to 100% for nasals, laterals, and some back consonants; much lower for other!

Arabic WebMAUS III

- Performance
  - Arabic WebMAUS $\Rightarrow \approx$ 95% accuracy at 20ms, comparable to other systems; increasing to 100% for nasals, laterals, and some back consonants; much lower for other!
  - Arabic WebMINNI $\Rightarrow$ Variable $\approx$ Baseline 50% - 100%

Arabic WebMAUS III

- Performance
  - Arabic WebMAUS $\Rightarrow \approx$ 95% accuracy at 20ms, comparable to other systems; increasing to 100% for nasals, laterals, and some back consonants; much lower for other!
  - Arabic WebMINNI $\Rightarrow$ Variable $\approx$ Baseline 50% - 100%
- Increase performance and sample size $\Rightarrow$
  - Various new datasets over 200 participants $\rightarrow$ 100 Jordanian; 20 Saudi; 30 Egyptian, 20 Lebanese; 20 Algerian; 10 Moroccan
  - Variable types of data $\Rightarrow$ Word lists, spontaneous, read and retold stories, etc.

Arabic WebMAUS III

- Performance
  - Arabic WebMAUS $\Rightarrow \approx$ 95% accuracy at 20ms, comparable to other systems; increasing to 100% for nasals, laterals, and some back consonants; much lower for other!
  - Arabic WebMINNI $\Rightarrow$ Variable $\approx$ Baseline 50% - 100%
- Increase performance and sample size $\Rightarrow$
  - Various new datasets over 200 participants $\rightarrow$ 100 Jordanian; 20 Saudi; 30 Egyptian, 20 Lebanese; 20 Algerian; 10 Moroccan
  - Variable types of data $\Rightarrow$ Word lists, spontaneous, read and retold stories, etc.
- Issues related to transcription

Overview

Variability in Arabic

- Working on obtaining phonetically-informed automatic transcription of dialectal Arabic
- Accounting for dialectal variation (see M2 thesis work by Maatallaoui, 2025:24)

Variability in Arabic

- Working on obtaining phonetically-informed automatic transcription of dialectal Arabic
- Accounting for dialectal variation (see M2 thesis work by Maatallaoui, 2025:24)

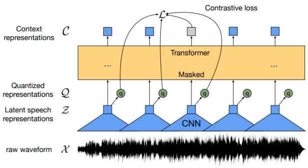
Phoneme	Realizations	Example	Dialects	Notes
/q/	[ʔ], [g], [q]	قلب /qalb/ → [ʔalb], [galb]	Egyptian, Gulf, Levantine	Urban vs. rural distinction
/d͡ʒ/	[ʒ], [g], [j]	جمل /d͡ʒamal/ → [ʒamal], [gamal], [jamal]	Maghreb, Egypt, Levant	Sociolinguistic variation
/θ/	[s], [t], [z]	ثلاثة /θala:θa/ → [tala:ta], [sala:sa]	Egypt, Sudan, Gulf	Fricative → stop
/k/	[t͡ʃ]	كبير /kabi:r/ → [t͡ʃbi:r]	Gulf, Iraqi	Gender/context-based shift
/ð/	[d], [z]	هذا /ha:ða:/ → [haza], [hada]	Egyptian, Levantine	Fricative → voiced stop
/ɣ/	[ʁ], [ʕ]	غريب /ɣari:b/ → [ʁari:b], [ʕari:b]	Gulf, Moroccan	Uvular vs. pharyngeal
/r/	[r], [ɾ], [ʁ]	رجل /raʒul/ → [ra ul], [ʁaʒul]	Moroccan, Iraqi	Trill, tap, or uvular
Emphatics	Vowel backing, spread	صديق /sˤadi:q/ → [sˤdˤi:q], [sadi:q]	MSA vs. dialects	Stronger in Maghreb
Short vowels	Elision, centralization	كتب /kataba/ → [ktib]	Moroccan, Levantine	Causes alignment errors

wav2vec2 and whisper I

- Used state-of-the-art approaches wav2vec2 and whisper $\Rightarrow$ Constitute foundation of recent advances in multilingual and low-resource speech processing

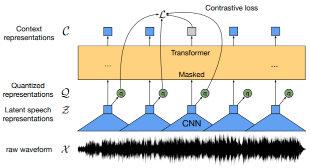
wav2vec2 and whisper I

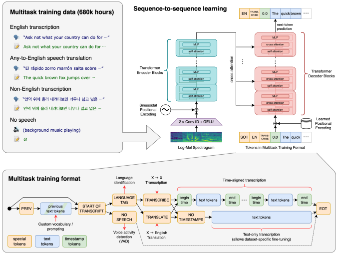
- Used state-of-the-art approaches wav2vec2 and whisper ⇒ Constitute foundation of recent advances in multilingual and low-resource speech processing
- wav2vec2
  - Self-supervised pretrained model for ASR (Baevski et al., 2020)
  - Trained on large amount of raw audios ⇒ masking parts of the representations and learning to predict the true representations using contrastive learning
  - Multiple layers initialised through the Transformer encoder ⇒ mapping features to corresponding tokens through the Connectionist Temporal Classification (CTC)



wav2vec2 and whisper I

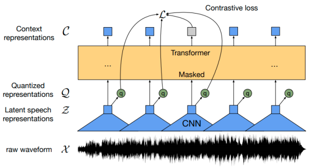
- Used state-of-the-art approaches wav2vec2 and whisper ⇒ Constitute foundation of recent advances in multilingual and low-resource speech processing
- wav2vec2
  - Self-supervised pretrained model for ASR (Baevski et al., 2020)
  - Trained on large amount of raw audios ⇒ masking parts of the representations and learning to predict the true representations using contrastive learning
  - Multiple layers initialised through the Transformer encoder ⇒ mapping features to corresponding tokens through the Connectionist Temporal Classification (CTC)
- Whisper
  - Supervised model pretrained on a large dataset of 680k hours of labeled multilingual and multitask speech (Radford et al., 2023)
  - Use of Transformer-based encoder-decoder design ⇒ Audios are first preprocessed to obtain the Log-Mel Spectogram, with feature normalization
  - Audios passed through two 1D-convolutional layers with positional embeddings and obtains contextual latent representations Output ⇒ generated transcriptions using the learned embeddings

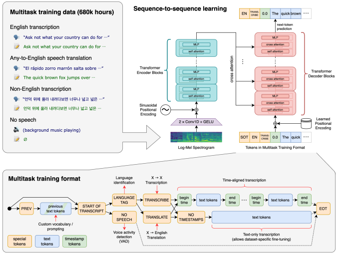




wav2vec2 and whisper I

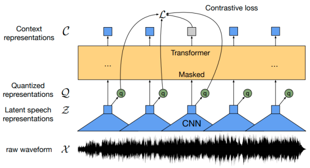
- Used state-of-the-art approaches wav2vec2 and whisper ⇒ Constitute foundation of recent advances in multilingual and low-resource speech processing
- wav2vec2
  - Self-supervised pretrained model for ASR (Baevski et al., 2020)
  - Trained on large amount of raw audios ⇒ masking parts of the representations and learning to predict the true representations using contrastive learning
  - Multiple layers initialised through the Transformer encoder ⇒ mapping features to corresponding tokens through the Connectionist Temporal Classification (CTC)
- Whisper
  - Supervised model pretrained on a large dataset of 680k hours of labeled multilingual and multitask speech (Radford et al., 2023)
  - Use of Transformer-based encoder-decoder design ⇒ Audios are first preprocessed to obtain the Log-Mel Spectogram, with feature normalization
  - Audios passed through two 1D-convolutional layers with positional embeddings and obtains contextual latent representations Output ⇒ generated transcriptions using the learned embeddings
- Both capture fine phonetic details and overall meaning
  - Wav2vec2 focuses on the acoustic details and subtle variations → Looks at a specific tree (or actually branch) and fine-grained speech characteristics ⇒ Not ideal for broader semantic context in ambiguous cases

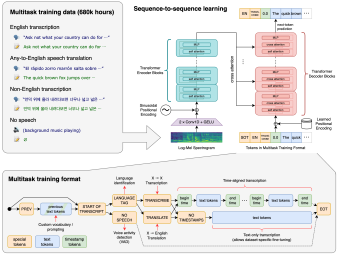




wav2vec2 and whisper I

- Used state-of-the-art approaches wav2vec2 and whisper ⇒ Constitute foundation of recent advances in multilingual and low-resource speech processing
- wav2vec2
  - Self-supervised pretrained model for ASR (Baevski et al., 2020)
  - Trained on large amount of raw audios ⇒ masking parts of the representations and learning to predict the true representations using contrastive learning
  - Multiple layers initialised through the Transformer encoder ⇒ mapping features to corresponding tokens through the Connectionist Temporal Classification (CTC)
- Whisper
  - Supervised model pretrained on a large dataset of 680k hours of labeled multilingual and multitask speech (Radford et al., 2023)
  - Use of Transformer-based encoder-decoder design ⇒ Audios are first preprocessed to obtain the Log-Mel Spectogram, with feature normalization
  - Audios passed through two 1D-convolutional layers with positional embeddings and obtains contextual latent representations Output ⇒ generated transcriptions using the learned embeddings
- Both capture fine phonetic details and overall meaning
  - Wav2vec2 focuses on the acoustic details and subtle variations
    → Looks at a specific tree (or actually branch) and fine-grained speech characteristics
    ⇒ Not ideal for broader semantic context in ambiguous cases
  - Whisper adopts a high-level, semantic-driven approach
    → Looks at the forest ⇒ producing fluent and plausible transcriptions that make sense in context, even when input audio is unclear
    ⇒ misses subtle phonetic nuances and pronunciation differences





wav2vec2 and whisper II

- Maatallaoui (2025)'s M2 thesis (see Maatallaoui and Al-Tamimi, in preparation)
- Aim ⇒ Build an ASR model that preserves phonetic fidelity, unlike what is common in existing Arabic ASR systems
- Datasets
  - Data from 94 speakers (Al-Tamimi et al., 2022)
  - 10 Levantine Arabic (producing real words) (Khattab et al., 2024)
  - 10 Egyptian Arabic (Ibrahim et al., 2020)
  - 60 Jordanian Arabic (Abuoudeh et al., 2024a,b)

Variety	NbWords	Duration
Arabic Bahraini	9 532	00:52:46
Arabic Egyptian	7 108	00:41:06
Arabic Lebanese	7 323	01:17:11
Arabic Levantine 1	2 034	00:26:39
Arabic Levantine 2	3 825	01:52:30
Arabic Saudi 1	26 806	02:59:38
Arabic Saudi 2	20 968	02:16:40
Arabic Jordanian	27 159	02:48:33
Arabic Standard	120 094	22:38:29
Total	**224 849**	**35:53:30**

wav2vec2 and whisper II

- Maatallaoui (2025)'s M2 thesis (see Maatallaoui and Al-Tamimi, in preparation)
- Aim ⇒ Build an ASR model that preserves phonetic fidelity, unlike what is common in existing Arabic ASR systems
- Datasets
  - Data from 94 speakers (Al-Tamimi et al., 2022)
  - 10 Levantine Arabic (producing real words) (Khattab et al., 2024)
  - 10 Egyptian Arabic (Ibrahim et al., 2020)
  - 60 Jordanian Arabic (Abuoudeh et al., 2024a,b)

Variety	NbWords	Duration
Arabic Bahraini	9 532	00:52:46
Arabic Egyptian	7 108	00:41:06
Arabic Lebanese	7 323	01:17:11
Arabic Levantine 1	2 034	00:26:39
Arabic Levantine 2	3 825	01:52:30
Arabic Saudi 1	26 806	02:59:38
Arabic Saudi 2	20 968	02:16:40
Arabic Jordanian	27 159	02:48:33
Arabic Standard	120 094	22:38:29
Total	**224 849**	**35:53:30**

- Dialectal data ⇒ Romanized transcriptions converted to Arabic script with/out diacritics using the ATR converter tool (Maatallaoui and Al-Tamimi, in preparation); Standard Arabic with diacritics
- Full datasets ⇒ Divided into 80-10-10% (training, validation and testing)
- Wav2vec2 ⇒ trained and evaluated on the dialectal Arabic → 5 hours with 2 epochs
- Whisper (small) ⇒ finetuned on both dialectal and diacritized Standard Arabic
- Word Error Rate (WER%), Diacritization Error Rate (DER%), and Character Error Rate (CER%)

wav2vec2 and whisper III

- Maatallaoui (2025)'s M2 thesis (see Maatallaoui and Al-Tamimi, in preparation)
- Whisper on diacritised text
  - With diacritics ⇒ 0.2561 (WER); 0.1113 (DER); 0.0829 (CER)

wav2vec2 and whisper III

- Maatallaoui (2025)'s M2 thesis (see Maatallaoui and Al-Tamimi, in preparation)
- Whisper on diacritised text
  - With diacritics ⇒ 0.2561 (WER); 0.1113 (DER); 0.0829 (CER)
  - Stripped diacritics ⇒ 0.1689 WER); 0.0707 (CER)

wav2vec2 and whisper III

- Maatallaoui (2025)'s M2 thesis (see Maatallaoui and Al-Tamimi, in preparation)
- Whisper on diacritised text
  - With diacritics ⇒ 0.2561 (WER); 0.1113 (DER); 0.0829 (CER)
  - Stripped diacritics ⇒ 0.1689 WER); 0.0707 (CER)
  - Wav2vec2 ⇒ inscreased (WER (close to 50%) on evaluation
    →frequent incoherence, especially in cases involving dialectal variation and learner pronunciation errors

wav2vec2 and whisper III

- Maatallaoui (2025)'s M2 thesis (see Maatallaoui and Al-Tamimi, in preparation)
- Whisper on diacritised text
  - With diacritics ⇒ 0.2561 (WER); 0.1113 (DER); 0.0829 (CER)
  - Stripped diacritics ⇒ 0.1689 WER); 0.0707 (CER)
  - Wav2vec2 ⇒ inscreased (WER (close to 50%) on evaluation →frequent incoherence, especially in cases involving dialectal variation and learner pronunciation errors
  - Variation ⇒ Overall decrease in the three metrics in ”stripped diacritics“ condition, except from Levantine Arabic (=isolated words)

With diacritics

Variety	**WER**	**DER**	**CER**
Arabic Bahraini	0.2803	0.1629	0.0868
Arabic Egyptian	0.7377	0.3301	0.3343
Arabic Lebanese	0.1571	0.0422	0.0347
Arabic Levantine	0.0443	0.0000	0.0144
Arabic Saudi	0.1477	0.0660	0.0442
Standard	0.4590	0.2085	0.1450

Stripped diacritics

Variety	**WER**	**DER**	**CER**
Arabic Bahraini	0.2443	–	0.0776
Arabic Egyptian	0.6478	–	0.3364
Arabic Lebanese	0.1148	–	0.0354
Arabic Levantine	0.0443	–	0.0152
Arabic Saudi	0.1074	–	0.0419
Standard	0.2292	–	0.1060

wav2vec2 and whisper IV

- Maatallaoui (2025)'s M2 thesis (see Maatallaoui and Al-Tamimi, in preparation)
- Whisper on undiacritised text
  - Overall $\Rightarrow$ 0.1795 (WER); 0.0778 (CER)

wav2vec2 and whisper IV

- Maatallaoui (2025)'s M2 thesis (see Maatallaoui and Al-Tamimi, in preparation)
- Whisper on undiacritised text
  - Overall ⇒ 0.1795 (WER); 0.0778 (CER)
  - Variation ⇒ Variable performance, except from Levantine/Lebanese Arabic (=isolated words)

wav2vec2 and whisper IV

- Maatallaoui (2025)'s M2 thesis (see Maatallaoui and Al-Tamimi, in preparation)
- Whisper on undiacritised text
  - Overall ⇒ 0.1795 (WER); 0.0778 (CER)
  - Variation ⇒ Variable performance, except from Levantine/Lebanese Arabic (=isolated words)

whisper

Variety	**WER**	**CER**
Arabic Bahraini	0.2827	0.0952
Arabic Egyptian	0.6141	0.2853
Arabic Lebanese	0.1107	0.0361
Arabic Levantine	0.0690	0.0189
Arabic Saudi	0.0990	0.0352
Standard	0.2172	0.1011

wav2vec2 and whisper V

- Trials on wav2vec2 on diacritised text showed many inconsistencies
- Looked at dialectal difference (with Shuhua Cao)
  - Overall $\Rightarrow$ 0.68 (WER)

wav2vec2 and whisper V

- Trials on wav2vec2 on diacritised text showed many inconsistencies
- Looked at dialectal difference (with Shuhua Cao)
  - Overall ⇒ 0.68 (WER)
  - Variation ⇒ Best performance on Levantine/Lebanese Arabic (=isolated words); difficulties with contextualised and long phrases (on other varieties) ⇒ Issues leading to clear generalisations between datasets.

wav2vec2 and whisper V

- Trials on wav2vec2 on diacritised text showed many inconsistencies
- Looked at dialectal difference (with Shuhua Cao)
  - Overall ⇒ 0.68 (WER)
  - Variation ⇒ Best performance on Levantine/Lebanese Arabic (=isolated words); difficulties with contextualised and long phrases (on other varieties) ⇒ Issues leading to clear generalisations between datasets.

wav2vec2

Variety	**WER**
Arabic Bahraini	0.0305
Arabic Egyptian	0.4119
Arabic Lebanese	0.0611
Arabic Levantine	0.0
Arabic Saudi	0.0848

- Worth exploring performance of wav2vec2 on individual dialects ⇒ Can we benefit from dialectal proximity? ⇒ Next steps

Overview

Introduction

WebMAUS

wav2vec2 and Whisper

Discussion and Conclusion

Discussion and Conclusion I

- Results promising but show several issues
  - Diacritisation $\Rightarrow$ challenging task especially for dialectal Arabic
    $\rightarrow$ Requires large-scale diacritised dataset $\Rightarrow$ Under development via Youtube channels
    $\approx$ 14h59min (see Maatallaoui and Al-Tamimi, under review, LREC)

Discussion and Conclusion I

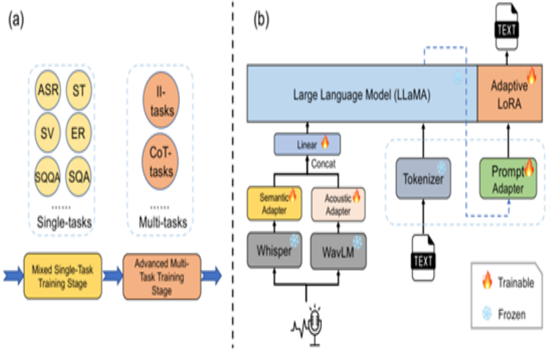
- Results promising but show several issues
  - Diacritisation ⇒ challenging task especially for dialectal Arabic
    → Requires large-scale diacritised dataset ⇒ Under development via Youtube channels ≈ 14h59min (see Maatallaoui and Al-Tamimi, under review, LREC)
  - LLM-Based Diacritisation (using DeepSeek-R1-Distill-Llama-8B) ⇒ Maatallaoui (2025)'s M2 thesis (see Maatallaoui and Al-Tamimi, in preparation) on Classical, Standard and Dialectal Arabic ⇒ Both romanised and diacritised versions (with 16, 20 and 7 hours fine-tuning) ⇒ Promising results, some errors due likely to inconsistent diacritisation/romanisation
    ⇒ Increased processing power and accessibility to large-scale datasets with both romanisation and diacritisation

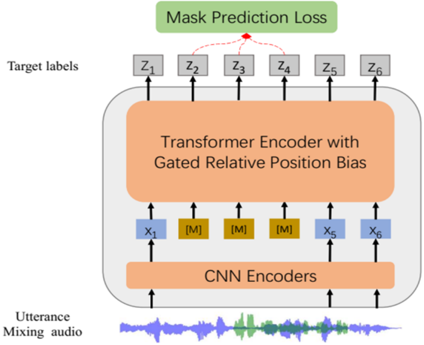
Discussion and Conclusion II

- Aim to develop a forced-alignment system leverages LLM and current developments in ASR
- Explore dialectal proximity and influence on performance of models $\Rightarrow$ Leverage on performance from wav2vec2

Discussion and Conclusion II

- Aim to develop a forced-alignment system leverages LLM and current developments in ASR
- Explore dialectal proximity and influence on performance of models ⇒ Leverage on performance from wav2vec2
- Explore the use of wavLMMs (Hu et al., 2024) based on wavLM (Chen et al., 2022)





Many thanks to:

My collaborators and participants!

Partial support from the Labex EFL - Strand 1 and inIdEx-EFL (WP4)

High Power Computing (HPC) facilities: CNRS/TGIR HUMA-NUM, IN2P3 and GENCI-IDRIS (Grant 2022-AD010613733)

Thank you... Questions, comments?

Jalal.Al-Tamimi@u-paris.fr

Tutorial material `https://tinyurl.com/2ymfy7ys`

Université Paris Cité

EFL

INIDEX UNIVERSITÉ PARIS CITÉ

Atrium Humanités et Sciences Sociales

FRANCE 2030

Laboratoire de linguistique formelle

Model name	Dataset	Newly added Variety	WER Arabic_Saudi	WER Arabic_Lebanese	WER Arabic_Levantine	WER Arabic_Bahraini	WER Arabic_Egyptian
model_on_Saudi_1and2	Saudi-1_and 2	Arabic_Saudi	0.11737915349417588(epoch 7)	-	-	-	-
model_saudi_and_leba	Leb :Saudi =1:1	Arabic_Lebanese	0.12822335758848177	0.15163043478260868 (epoch 4)	-	-	-
model_on_Saudi_leba_levan_2	Leba:saud:levan = 1:1	Arabic_Levantine	0.1334833608177385	0.1448369565217391	0.019230769230769232 (epoch6)	-	-
model_on_Saudi_leba_levan_Bahr	Bahr_leba_saud_levan_balanced	Arabic_Bahraini	0.07424659806655369	0.09130434782608696	0.0	0.29442971811018015	-
model_on_Saudi_leba_levan_Bah_2	Bahr_leba_saud_levan_major Bahr_train	Arabic_Bahraini	0.07765388790293004	0.09103260869565218	0.0	0.2789360094366812	-
model_on_Saudi_leba_levan_Bah_Egpt	Egyp_leba_saud_levan_Bah_balanced_dataset_train	Arabic_Egyptian	0.07180537139969188	0.05570652173913043	0.0	0.04193228165904457	0.7690456056161399
model_on_Saudi_leba_levan_Bah_Egpt_2	Egyp_dataset	Arabic_Egyptian	0.0801510235897209	0.04755434782608696	0.0	0.04358400537921554	0.5741017142345557
model_on_Saudi_leba_levan_Bah_Egpt_3	Egyp_Bahr_balanced_dataset_train	Arabic_Egyptian	0.07482404958458598	0.059782608695652176	0.0	0.03238278338308023	0.4583515626344327(epoch1)
model_on_Saudi_leba_levan_Bah_Egpt_4	Egyp_leba_balanced_dataset_train	Arabic_Egyptian	0.072028030505042	0.0625	0.0	0.04713657156698444	0.43222794512770685
model_on_Saudi_leba_levan_Bah_Egpt_5	Egyp_bahr_leba_balanced_dataset_train	Arabic_Egyptian	0.08480156497397877	0.06114130434782609	0.0	0.030544861824919733	0.4119391451763523

Overview

References I

Abuoudeh, M., Al-Tamimi, J., and Crouzet, O. (2024a). L'impact du style de parole sur l'opposition de longueur des voyelles en arabe jordanien. In *Actes Des 35èmes Journées d'Études Sur La Parole (JEP 2024) 31ème Conférence Sur Le Traitement Automatique Des Langues Naturelles (TALN 2024) 26ème Rencontre Des Étudiants Chercheurs En Informatique Pour Le Traitement Automatique Des Langues (RECITAL 2024), Jul 2024, Toulouse, France.*, pages 421–430, Toulouse 8-12 juillet 2024.

Abuoudeh, M., Al-Tamimi, J., and Crouzet, O. (2024b). Speaking style influence on vowel length opposition in Jordanian Arabic. In *Proceedings of the 13th International Seminar on Speech Production*, Autrans, France (13-17 May 2024).

Al-Tamimi, J., Schiel, F., Khattab, G., Sokhey, N., Amazouz, D., Dallak, A., and Moussa, H. (2022). A Romanization System and WebMAUS Aligner for Arabic Varieties. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), © European Language Resources Association (ELRA), Licensed under CC-BY-NC-4.0*, pages 7269–7276, Marseille, 20-25 June 2022.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., and Wei, F. (2022). WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Hu, S., Zhou, L., Liu, S., Chen, S., Meng, L., Hao, H., Pan, J., Liu, X., Li, J., Sivasankaran, S., Liu, L., and Wei, F. (2024). WavLLM: Towards Robust and Adaptive Speech Large Language Model. In *EMNLP2024 Findings*.

References II

Ibrahim, O., Asadi, H., Kassem, E., and Dellwo, V. (2020). Arabic Speech Rhythm Corpus: Read and Spontaneous Speaking Styles. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5337–5342, Marseille, France. European Language Resources Association.

Khattab, G., Xing, K., Turton, D., Al-Tamimi, J., and Alsharif, B. (2024). Syllabic and emphatic conditioning of /l/ in Levantine Arabic: An auditory, acoustic and articulatory analysis. In *Proceedings of Ultrafest XI*, University of Aizu - 24-25 June 2024.

Kisler, T., Reichel, U., and Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347.

Maatallaoui, Y. (2025). Advancing Arabic Speech and Text Processing: LLM-Based Diacritization and Fine-Tuning a Dialect- and Variation-Aware Whisper Model. M2 Thesis, Language Sciences, strand: Computational Linguistics, Université Paris Cité, Paris, 26 Juin 2025.

Maatallaoui, Y. and Al-Tamimi, J. (in preparation). Current developments of Arabic Forced-alignment systems.

Maatallaoui, Y. and Al-Tamimi, J. (under review). VoxDamas: A Diacritized Multidialectal Arabic Speech Corpus. In *LREC2026*.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Schiel, F. (1999). Automatic Phonetic Transcription of Non-Prompted Speech. In *Proc. of the ICPhS*, pages 607–610, San Francisco.