Wav2Gloss: Generating Interlinear Glossed Text from Speech

ACL 2024 Main (Long Papers)
Taiqi He (CMU)

Carnegie Mellon University
Johns Hopkins University

Our team

Taiqi He

Kwanghee Choi

Lindia Tjuatja

Nate Robinson

Jiatong Shi

Graham Neubig

Shinji Watanabe

David Mortensen

Lori Levin

Contents

1. Background: What is Interlinear Glossed Text (IGT)?
2. *Fieldwork* Dataset
3. *Wav2Gloss* Task
4. Experiments
5. Takeaways

Background

Field Linguistic Recordings

- Field data is collected by linguists for indigenous language documentation.
- It consists of audio and sometimes video recordings—crucial for documentation, preservation, revitalization of languages at risk.
- Needs expert annotations, and annotation process is expensive.

Credit: Jonathan Amith, Gettysburg College

Interlinear Glossed Text

- *Lingua franca* of documentary linguistics
- Especially important for illustrating the morphology of documented languages

Transcription (`wd`):	n	sìginde		yan	de
Surface (`sr`):	n	sìgi	-nde	yan	de
Underlying (`ur`):	n	sìgi	-len	yan	le
Gloss (`gl`):	1.SG	sit	-PC.RES	that	FOC
Translation (`tr`):	I live here.				

Segment by morphemes

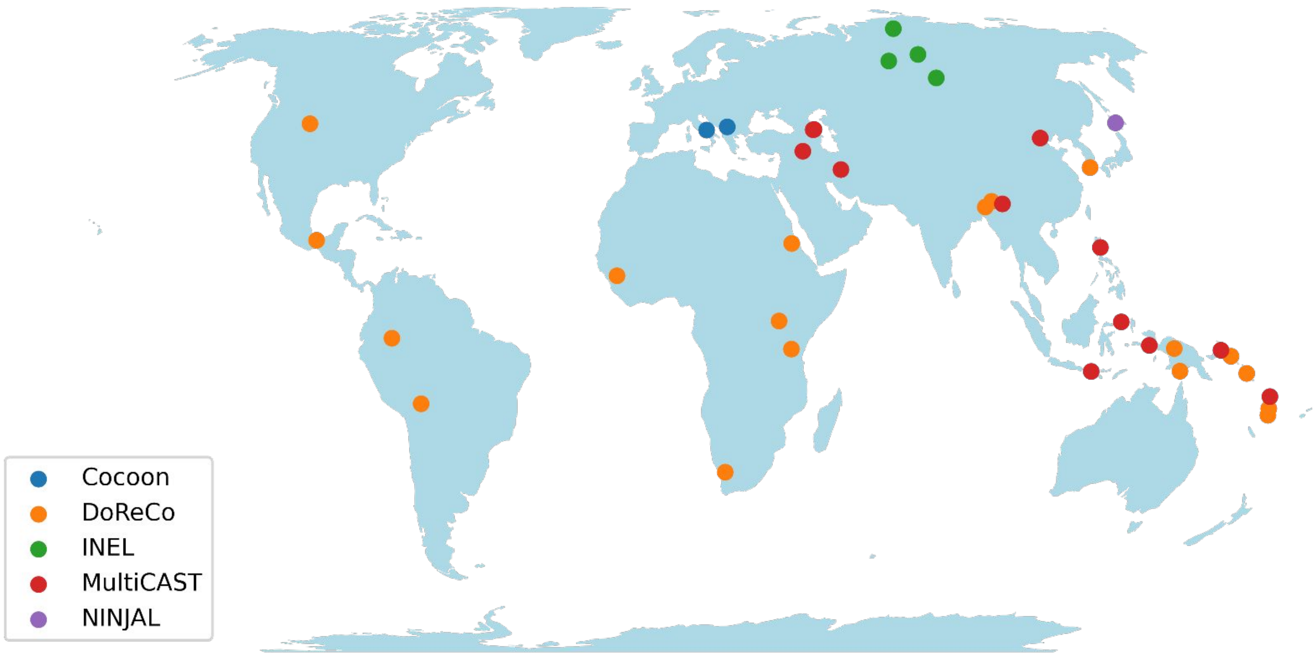
Abstract form of morpheme
(before applying phonological rules)

Morpheme-wise annotation
(with translation and/or grammatical function)

Kakabe (Vydrina, 2022)

Fieldwork Dataset

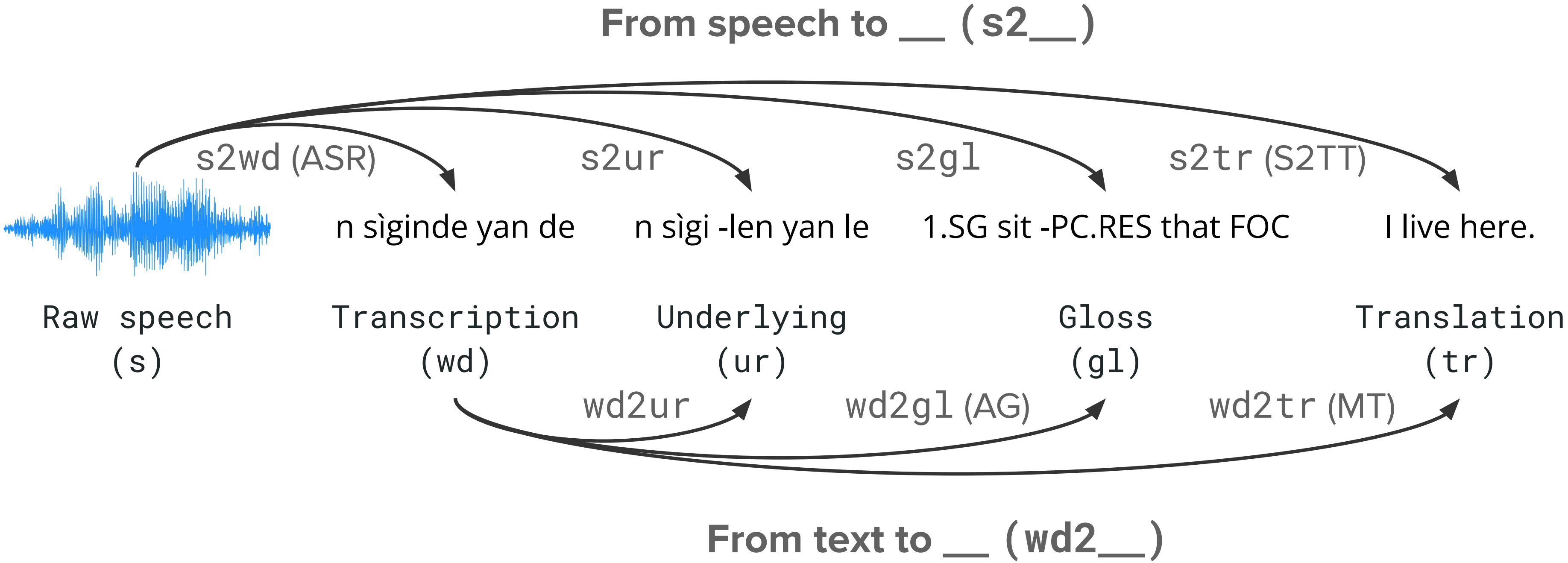
Fieldwork Dataset Statistics



- 37 Languages from 5 linguistic fieldwork repositories
- 71.35 hours of data in total
- Train/dev/test split
- 22 seen languages and 15 unseen languages
- May contain very personal information about the speakers and their families.

Wav2Gloss Task

Wav2Gloss Task Definition



From speech to __ (s2__)

s2wd (ASR)

s2ur

s2gl

s2tr (S2TT)

n sìginde yan de

n sìgi -len yan le

1.SG sit -PC.RES that FOC

I live here.

Raw speech (s)

Transcription (wd)

Underlying (ur)

Gloss (gl)

Translation (tr)

wd2ur

wd2gl (AG)

wd2tr (MT)

From text to __ (wd2__)

Experiments

Baseline design

E2E (End-to-end) vs. Cascade

- Can wav2gloss tasks be solved in an end-to-end manner?

Single task vs. Multi-task

- Do different tasks help each other?

Monolingual vs. Multilingual training

- Do languages benefit from other languages?

Self-supervised vs. Weakly supervised speech models

- How do the pre-trained speech models influence downstream performance?

Experimental results

E2E (End-to-end) vs. Cascade

- E2E models show better performance except for translation.

Multi-task vs. Single task

- Multi-task models usually performs worse.

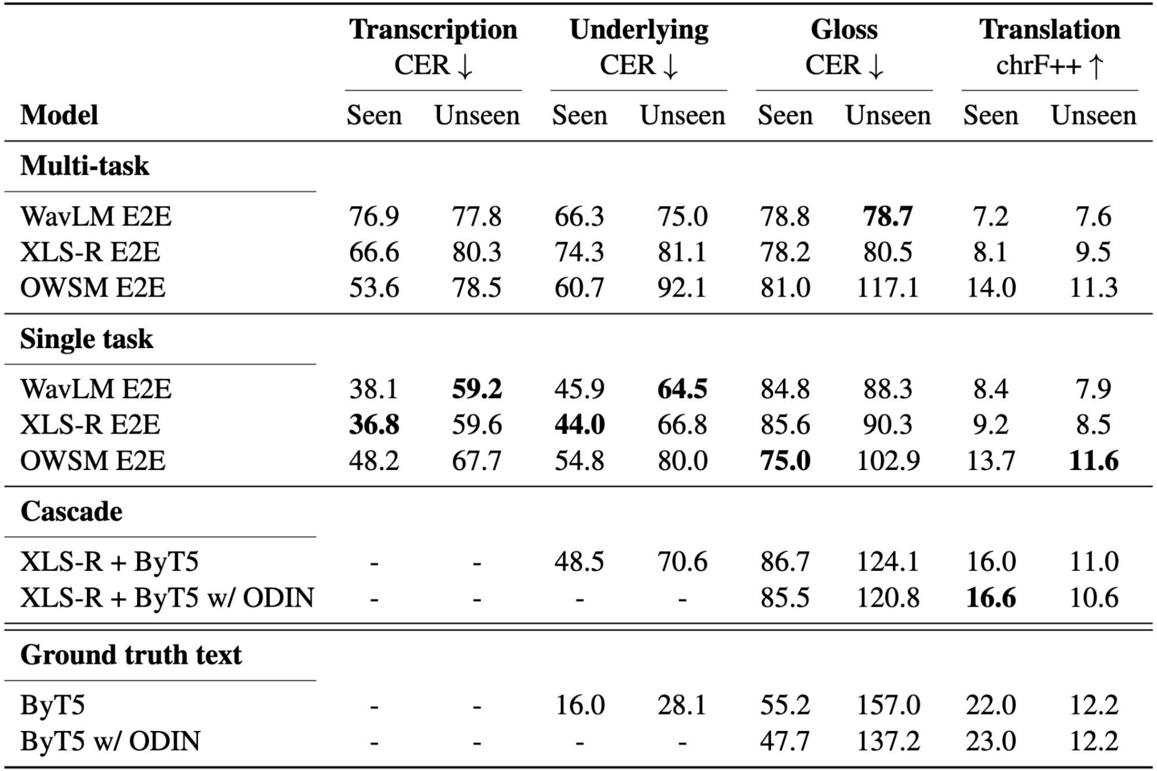
Monolingual vs. Multilingual training

- Only the lowest-resource languages benefit from multilingual training.

Self-supervised vs. Weakly supervised speech models

- Self-supervised models are better at transcription and underlying, weakly supervised models are better at gloss and translation.

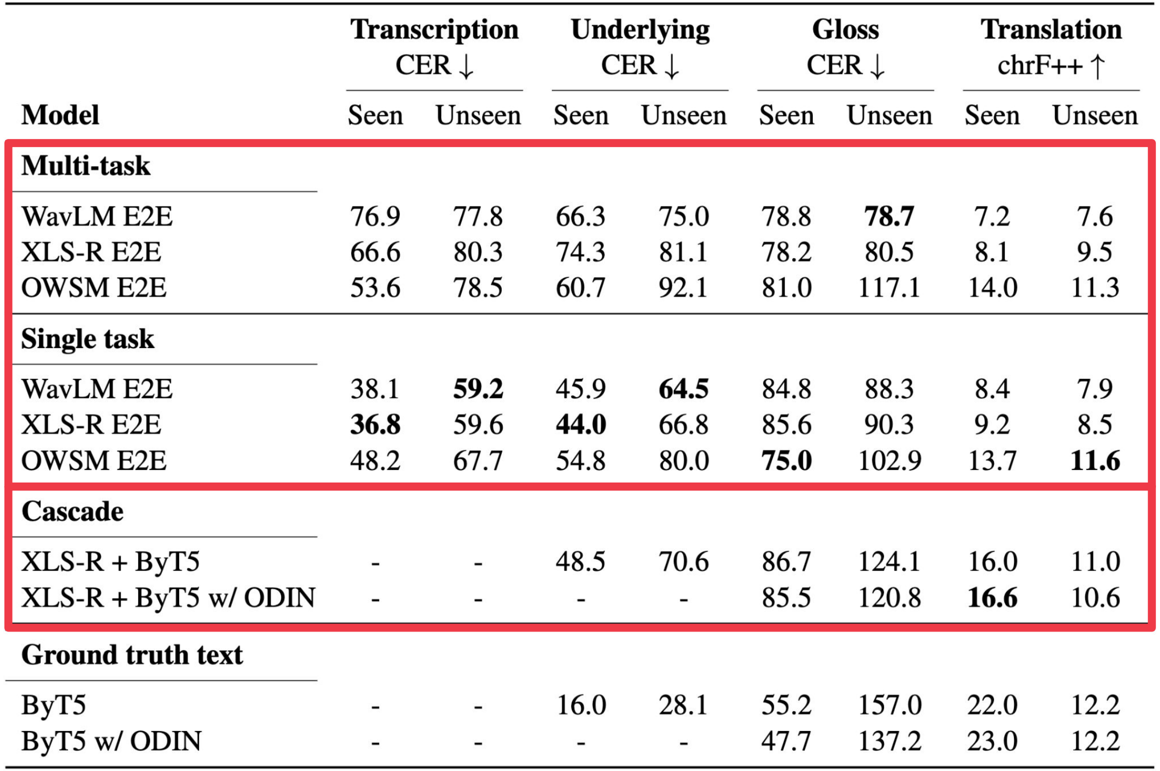
Experimental results



	Transcription CER ↓		Underlying CER ↓		Gloss CER ↓		Translation chrF++ ↑	
Model	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen
Multi-task								
WavLM E2E	76.9	77.8	66.3	75.0	78.8	**78.7**	7.2	7.6
XLS-R E2E	66.6	80.3	74.3	81.1	78.2	80.5	8.1	9.5
OWSM E2E	53.6	78.5	60.7	92.1	81.0	117.1	14.0	11.3
Single task								
WavLM E2E	38.1	**59.2**	45.9	**64.5**	84.8	88.3	8.4	7.9
XLS-R E2E	**36.8**	59.6	**44.0**	66.8	85.6	90.3	9.2	8.5
OWSM E2E	48.2	67.7	54.8	80.0	**75.0**	102.9	13.7	**11.6**
Cascade								
XLS-R + ByT5	-	-	48.5	70.6	86.7	124.1	16.0	11.0
XLS-R + ByT5 w/ ODIN	-	-	-	-	85.5	120.8	**16.6**	10.6
Ground truth text								
ByT5	-	-	16.0	28.1	55.2	157.0	22.0	12.2
ByT5 w/ ODIN	-	-	-	-	47.7	137.2	23.0	12.2

Experimental results - E2E (End-to-end) vs. Cascade

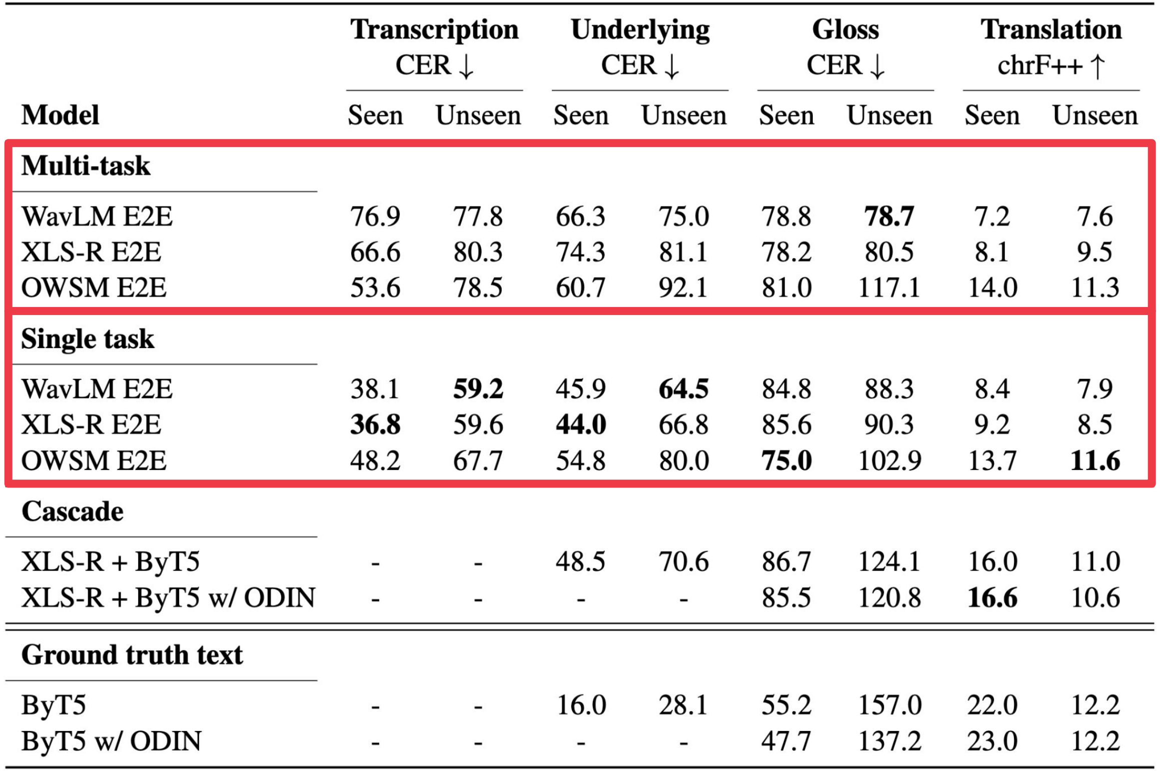
E2E models show better performance except for translation.



	Transcription CER ↓		Underlying CER ↓		Gloss CER ↓		Translation chrF++ ↑	
Model	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen
Multi-task								
WavLM E2E	76.9	77.8	66.3	75.0	78.8	**78.7**	7.2	7.6
XLS-R E2E	66.6	80.3	74.3	81.1	78.2	80.5	8.1	9.5
OWSM E2E	53.6	78.5	60.7	92.1	81.0	117.1	14.0	11.3
Single task								
WavLM E2E	38.1	**59.2**	45.9	**64.5**	84.8	88.3	8.4	7.9
XLS-R E2E	**36.8**	59.6	**44.0**	66.8	85.6	90.3	9.2	8.5
OWSM E2E	48.2	67.7	54.8	80.0	**75.0**	102.9	13.7	**11.6**
Cascade								
XLS-R + ByT5	-	-	48.5	70.6	86.7	124.1	16.0	11.0
XLS-R + ByT5 w/ ODIN	-	-	-	-	85.5	120.8	**16.6**	10.6
Ground truth text								
ByT5	-	-	16.0	28.1	55.2	157.0	22.0	12.2
ByT5 w/ ODIN	-	-	-	-	47.7	137.2	23.0	12.2

Experimental results - Multi-task vs. Single Task

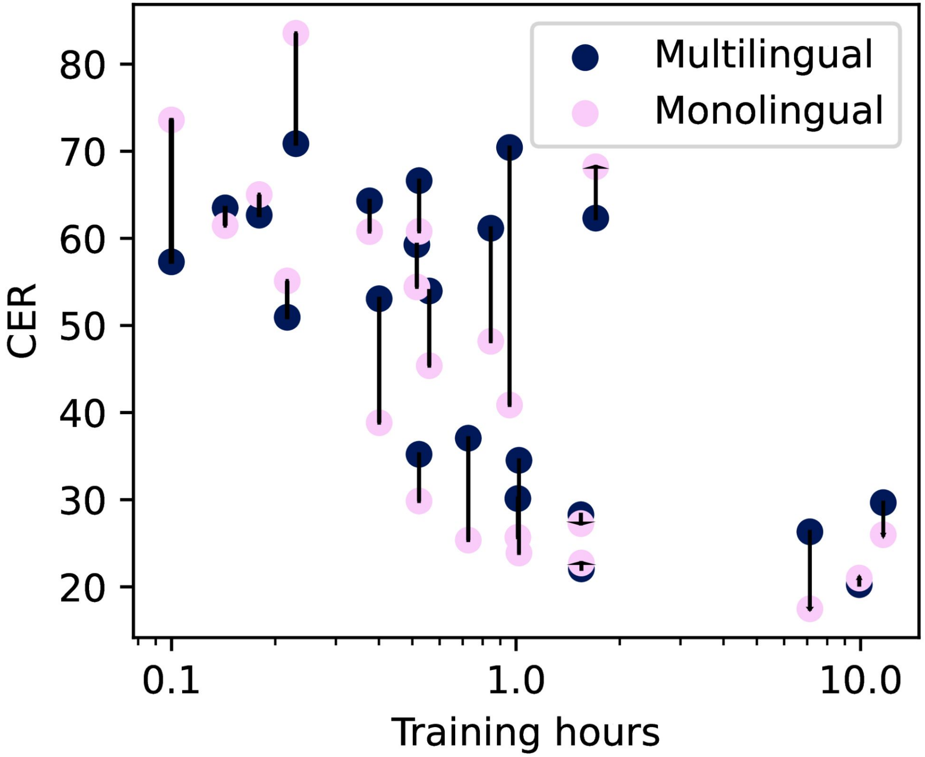
Multi-task models usually performs worse.



	Transcription CER ↓		**Underlying** CER ↓		**Gloss** CER ↓		**Translation** chrF++ ↑	
Model	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen
Multi-task								
WavLM E2E	76.9	77.8	66.3	75.0	78.8	**78.7**	7.2	7.6
XLS-R E2E	66.6	80.3	74.3	81.1	78.2	80.5	8.1	9.5
OWSM E2E	53.6	78.5	60.7	92.1	81.0	117.1	14.0	11.3
Single task								
WavLM E2E	38.1	**59.2**	45.9	**64.5**	84.8	88.3	8.4	7.9
XLS-R E2E	**36.8**	59.6	**44.0**	66.8	85.6	90.3	9.2	8.5
OWSM E2E	48.2	67.7	54.8	80.0	**75.0**	102.9	13.7	**11.6**
Cascade								
XLS-R + ByT5	-	-	48.5	70.6	86.7	124.1	16.0	11.0
XLS-R + ByT5 w/ ODIN	-	-	-	-	85.5	120.8	**16.6**	10.6
Ground truth text								
ByT5	-	-	16.0	28.1	55.2	157.0	22.0	12.2
ByT5 w/ ODIN	-	-	-	-	47.7	137.2	23.0	12.2

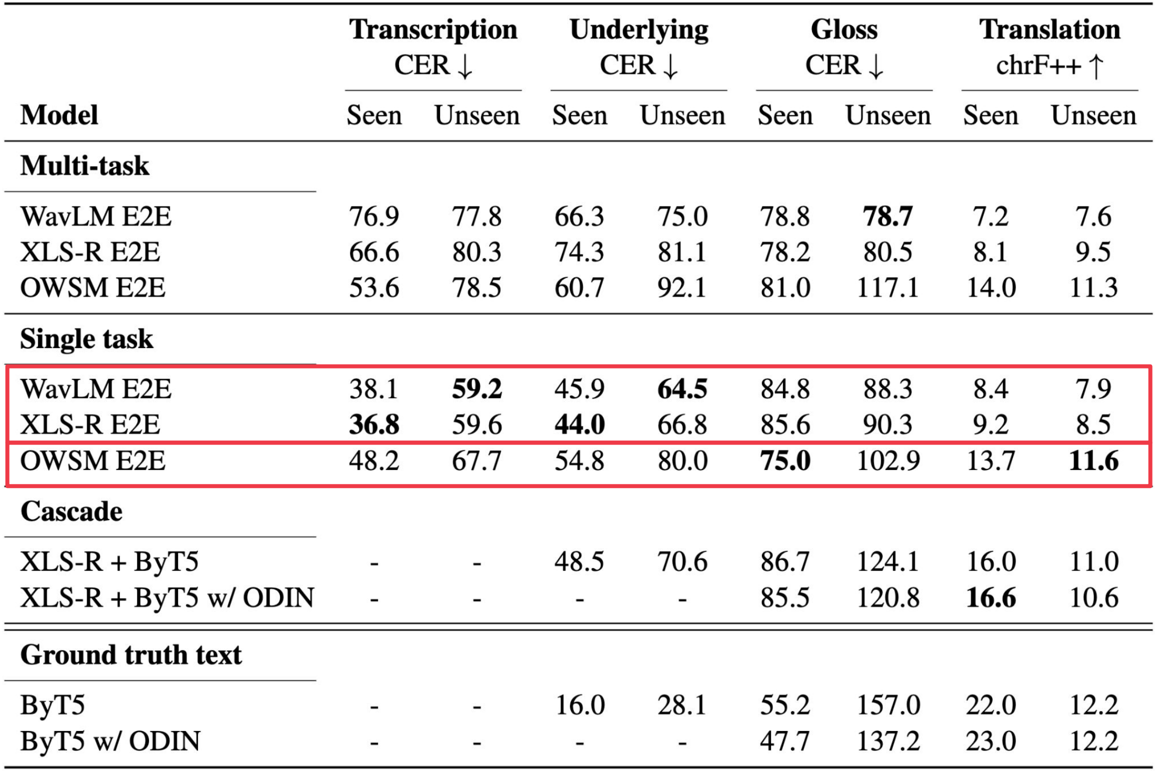
Experimental results - Monolingual vs. Multilingual

Only the lowest-resource languages benefit from multilingual training.



Experimental results - Self-supervised vs. Weakly supervised

Self-supervised models are better at transcription and underlying, weakly supervised models are better at gloss and translation.



	Transcription CER ↓		Underlying CER ↓		Gloss CER ↓		Translation chrF++ ↑	
Model	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen
Multi-task								
WavLM E2E	76.9	77.8	66.3	75.0	78.8	**78.7**	7.2	7.6
XLS-R E2E	66.6	80.3	74.3	81.1	78.2	80.5	8.1	9.5
OWSM E2E	53.6	78.5	60.7	92.1	81.0	117.1	14.0	11.3
Single task								
WavLM E2E	38.1	**59.2**	45.9	**64.5**	84.8	88.3	8.4	7.9
XLS-R E2E	**36.8**	59.6	**44.0**	66.8	85.6	90.3	9.2	8.5
OWSM E2E	48.2	67.7	54.8	80.0	**75.0**	102.9	13.7	**11.6**
Cascade								
XLS-R + ByT5	-	-	48.5	70.6	86.7	124.1	16.0	11.0
XLS-R + ByT5 w/ ODIN	-	-	-	-	85.5	120.8	**16.6**	10.6
Ground truth text								
ByT5	-	-	16.0	28.1	55.2	157.0	22.0	12.2
ByT5 w/ ODIN	-	-	-	-	47.7	137.2	23.0	12.2

Takeaways

Takeaways

- Interlinear Glossed Text (IGT) is used to document endangered languages.
- Producing IGT from raw speech (Wav2Gloss) is a necessary, tractable, yet challenging problem.
- We provide the Fieldwork dataset and various baselines to lay the groundwork for future research on IGT.

Thank you!

Paper: https://arxiv.org/abs/2403.13169
Dataset: https://huggingface.co/datasets/wav2gloss/fieldwork
Code (SSL): https://github.com/juice500ml/espnet/tree/wav2gloss
Code (WSL): https://github.com/juice500ml/finetune_owsm