WAV2GLOSS: Generating Interlinear Glossed Text from Speech

Taiqi He[1] taiqih@andrew.cmu.edu Kwanghee Choi[1] Lindia Tjuatja[1] Nathaniel R. Robinson[2] Jiatong Shi[1]
Shinji Watanabe[1] Graham Neubig[1] David R. Mortensen[1] Lori Levin[1] levin@andrew.cmu.edu

[1]Language Technologies Institute, Carnegie Mellon University [2]Center for Language and Speech Processing, Johns Hopkins University

Summary

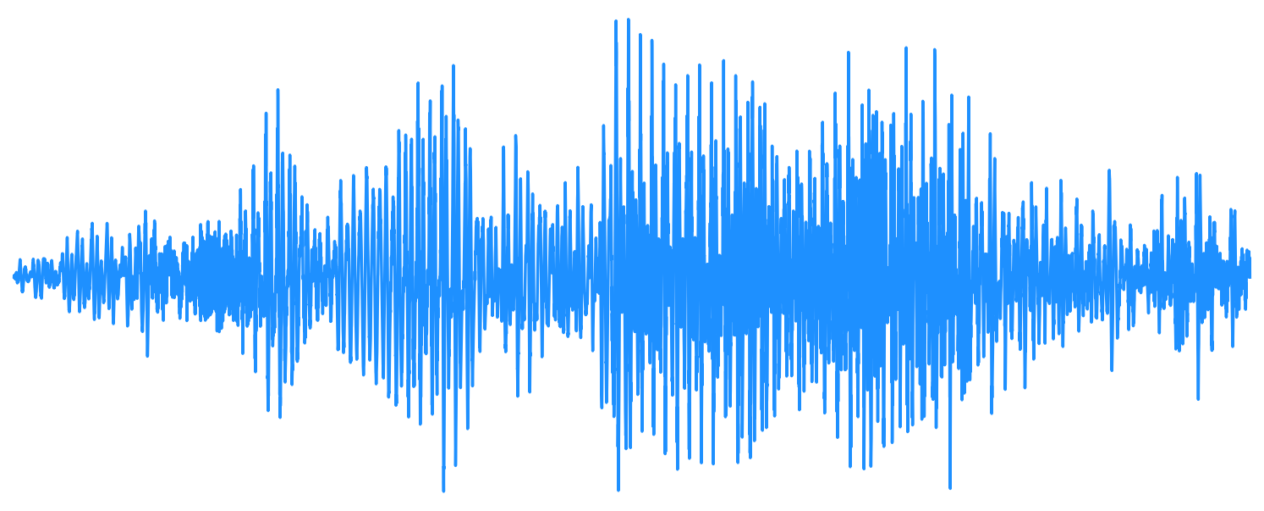
- **Interlinear Glossed Text (IGT)** is the standard notation for linguistic documentation. IGT is necessary to document **endangered languages**.
- **Wav2Gloss task**: input raw speech and output detailed linguistic annotations. The task is a tractable yet challenging problem.
- **Fieldwork dataset** (CC-BY-NC-SA-4.0): a corpus of preexisting linguistic field data (71 hours of annotated speech in 37 languages) that we formatted and **benchmarked** for the Wav2Gloss task.

Field Linguistic Recordings

Figure 1. Fieldwork in action. *Credit: Jonathan Amith, Gettysburg College*

- 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.
- To complete the documentation process, transcription and other expert annotations are needed. However, transcription alone takes **up to 10 hours per hour of speech**. Manually annotating a corpus with IGT takes years.

Interlinear Glossed Text (IGT)



`wd`:	n	sìginde		yan	de
`sr`:	n	sìgi	-nde	yan	de
`ur`:	n	sìgi	-len	yan	le
`gl`:	1.SG	sit	-PC.RES	that	FOC
`tr`:	"I live here."				

Figure 2. A representation of a single Kakabe utterance in our corpus: speech paired with annotations.

- **Transcription** (`wd`): ASR transcription, following the native orthography (writing system) of each language.
- **Surface** (`sr`): The transcription is additionally segmented into morphemes.
- **Underlying** (`ur`): Same segmentation as surface form, but illustrating a form where phonological rules have not applied.
- **Gloss** (`gl`): Same segmentation as underlying form, with brief grammatical and/or lexical explanations (IGT) for each morpheme.
- **Translation** (`tr`): Utterance translated to meta-language (English for ours).

Wav2Gloss Task

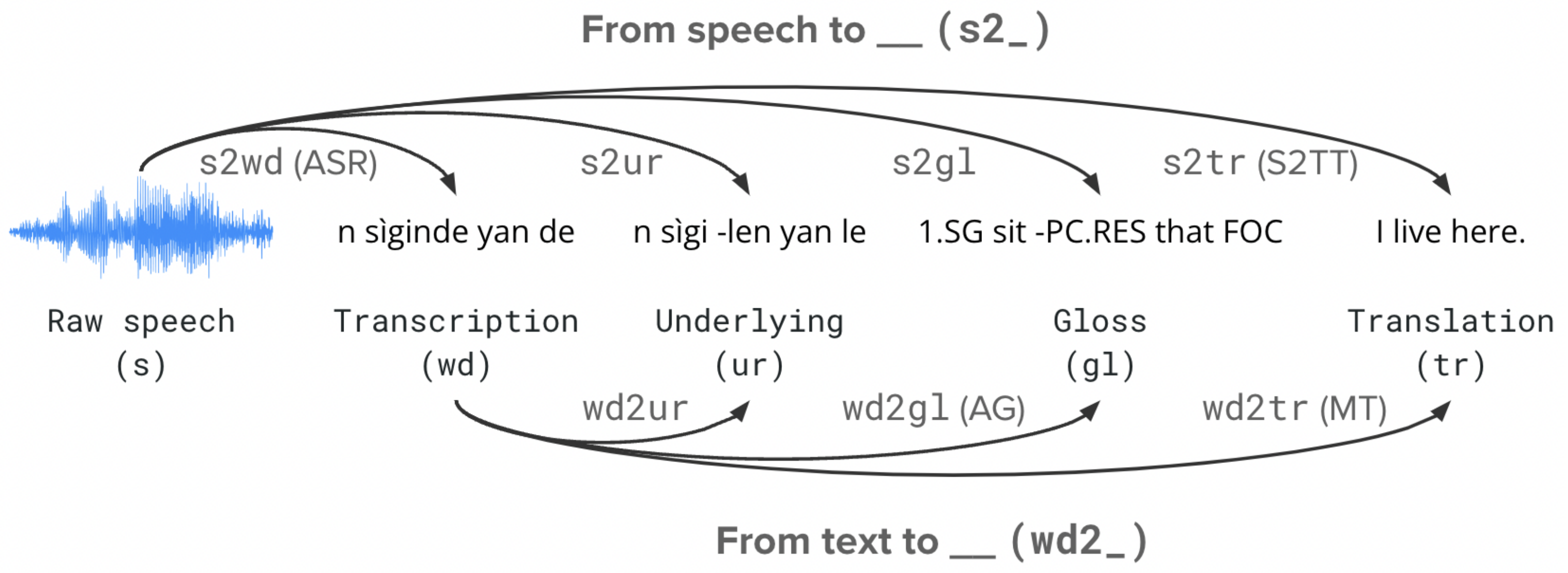


Figure 3. Various subtasks of Wav2Gloss, including automatic speech recognition (ASR), speech-to-text translation (S2TT), automatic glossing (AG), and machine translation (MT).

- We set up the task such that each line of the IGTs is treated as a target.
- Then, we treat them each as a sequence to sequence task, from raw speech signals to linearized texts.
- We also consider starting from text, *i.e.*, the cascade setup.
- One can expand to other tasks as we provide the parallel dataset.
- For evaluation, we used chrF++ for translation and CER for others.

Fieldwork Dataset

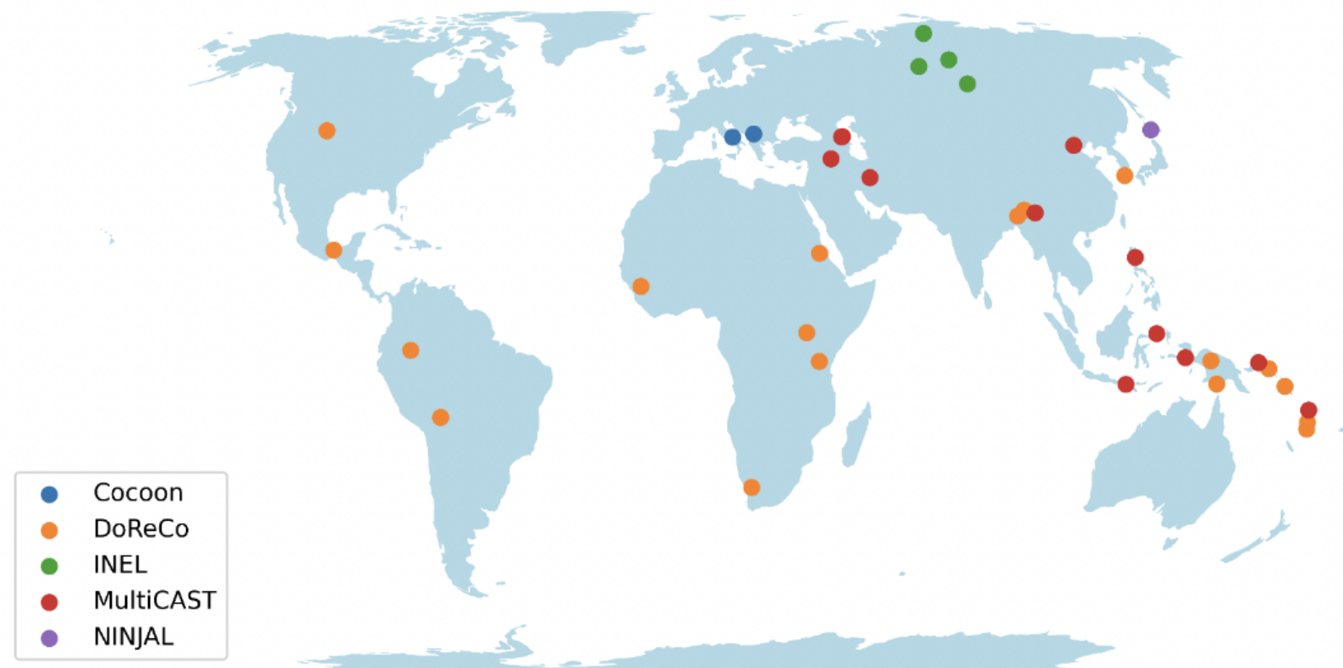


Figure 4. Supported languages

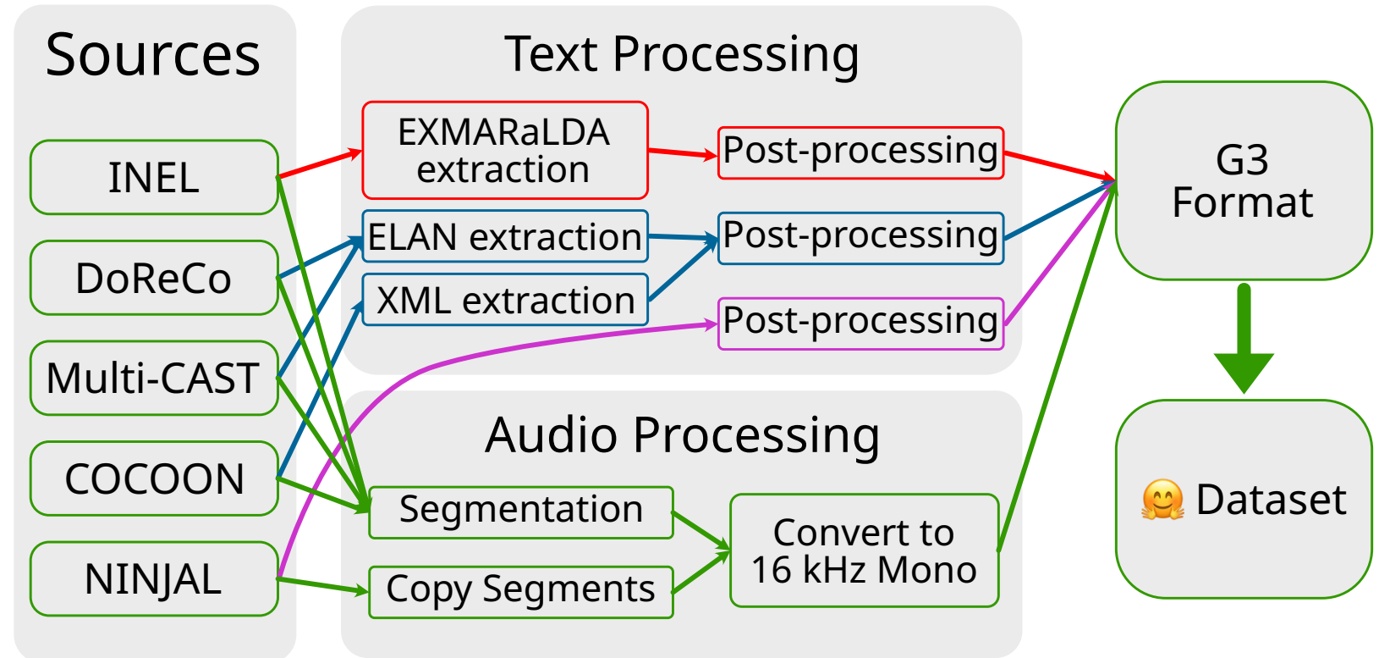


Figure 5. Dataset flowchart

- To facilitate glossing as a machine learning task, we have gathered a dataset from field linguistic work.
- Our dataset contains all of the representations (transcription, underlying, gloss, and translation) in parallel.
- The dataset contains 37 languages around the world from 5 different sources, totalling around 71 hours of data.
- We did a careful split on the dataset to avoid data spilling.
- The dataset may contain very personal information about the speakers and their communities. We therefore urge all users of the dataset to keep that in mind, and also respect the licenses set by the researchers.

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

Table 1. Results from multilingual experiments. Each number represents an average of that metric across the languages in that group (macro-averaging).

End-to-end (speech-to-X) vs. Cascade (speech-to-text-to-X)

- Q. Can wav2gloss tasks be solved in an end-to-end manner?
- A. Yes. E2E models show better performance except for translation.
- If we can take advantage of stronger text-based models, it might have the potential to perform better.

Self-supervised vs. Weakly supervised speech models

- Q. How do the pre-trained speech models influence downstream performance?
- A. Self-supervised models are better at transcription and underlying forms, weakly supervised models are better at gloss and translation.
- We suspect it is due to the former's better acoustic modeling capabilities and the latter's exposure to the translation task during pretraining.

Single task vs. Multi-task

- Q. Do different tasks help each other?
- A. In our setup, no.
- We may need a more sophisticated modeling.

Monolingual vs. Multilingual training

- Q. Is multilingualism beneficial?
- A. Only the lowest-resource languages benefit from multilingual training.
- It seems our setup suffered from the curse of multilinguality.

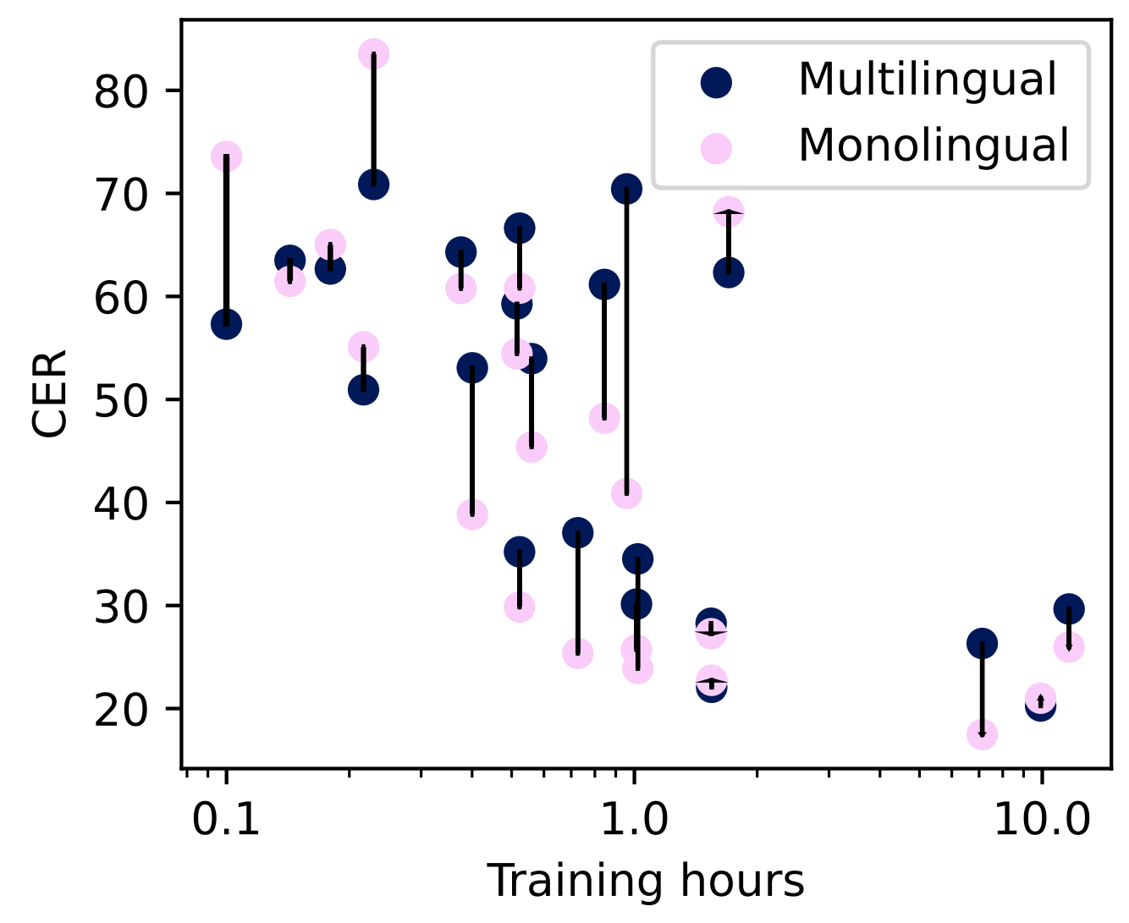


Figure 6. Comparing multilingual transcription OWSM E2E model with monolingual models.