"cdli/kenyan_english_nonstandard_speech_v0.9" small

Saving to: /jupyter_kernel/trained_models/en_nonstandard_tune_whisper_small_2/best_model

Training

[600/600 40:27, Epoch 2/3]

Step	Training Loss	Validation Loss	Wer	Cer	Lattescore	Input Tokens Seen
0	No log	1.416936	0.278586	0.166302	48.245614	0
50	0.999600	0.796919	0.234155	0.132961	60.526316	144000000
100	0.958800	0.756722	0.240740	0.144147	61.988304	288000000
150	0.802600	0.724311	0.227360	0.135700	64.619883	432000000
200	0.820200	0.702249	0.208163	0.121648	67.836257	576000000
250	0.882900	0.675111	0.213495	0.128683	66.959064	720000000
300	0.484300	0.662475	0.201840	0.119898	69.298246	863520000
350	0.706600	0.664432	0.201682	0.122105	69.298246	1007520000
400	0.489400	0.663069	0.202475	0.121708	67.836257	1151520000
450	0.466500	0.660234	0.200163	0.119130	68.128655	1295520000
500	0.461800	0.659746	0.197570	0.117026	69.298246	1439520000
550	0.428600	0.659702	0.197718	0.117189	69.005848	1583040000
600	0.453200	0.659704	0.197192	0.116947	69.298246	1727040000

```
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.2786
Adjusted CER: 0.1663
LATTEScore: 48.25%
Un-adjusted WER: 0.2915
Un-adjusted CER: 0.1788
===============
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.2342
Adjusted CER: 0.1330
LATTEScore: 60.53%
Un-adjusted WER: 0.2394
Un-adjusted CER: 0.1330
===============
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.2407
Adjusted CER: 0.1441
LATTEScore: 61.99%
Un-adjusted WER: 0.2798
Un-adjusted CER: 0.1715

Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.2274
Adjusted CER: 0.1357
LATTEScore: 64.62%
Un-adjusted WER: 0.2512
Un-adjusted CER: 0.1427
===============
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.2082
Adjusted CER: 0.1216
LATTEScore: 67.84%
Un-adjusted WER: 0.2476
Un-adjusted CER: 0.1524
===============
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.2135
Adjusted CER: 0.1287
LATTEScore: 66.96%
Un-adjusted WER: 0.2769
Un-adjusted CER: 0.1776
===============
```

```
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.2018
Adjusted CER: 0.1199
LATTEScore: 69.30%
Un-adjusted WER: 0.2672
Un-adjusted CER: 0.1653
===============
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.2017
Adjusted CER: 0.1221
LATTEScore: 69.30%
Un-adjusted WER: 0.2659
Un-adjusted CER: 0.1763
===============
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.2025
Adjusted CER: 0.1217
LATTEScore: 67.84%
Un-adjusted WER: 0.2551
Un-adjusted CER: 0.1688
===============
```

```
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.2002
Adjusted CER: 0.1191
LATTEScore: 68.13%
Un-adjusted WER: 0.2528
Un-adjusted CER: 0.1662
===============
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.1976
Adjusted CER: 0.1170
LATTEScore: 69.30%
Un-adjusted WER: 0.2502
Un-adjusted CER: 0.1641
===============
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.1977
Adjusted CER: 0.1172
LATTEScore: 69.01%
Un-adjusted WER: 0.2503
Un-adjusted CER: 0.1643
===============
```

```
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.1972
Adjusted CER: 0.1169
LATTEScore: 69.30%
Un-adjusted WER: 0.2498
Un-adjusted CER: 0.1640
==============
There were missing keys in the checkpoint model loaded: ['proj_out.weight'].
```

```
]: TrainOutput(global_step=600, training_loss=0.6770224614938101, metrics={'train_runtime': 2553.761, 'train_samples_per_second': 2.819, 'train_steps_per_
second': 0.235, 'total_flos': 2.07666054070272e+18, 'train_loss': 0.6770224614938101, 'epoch': 2.2988505747126435, 'num_input_tokens_seen': 172704000
0})
```

Post training

On DEV set

```
# (should give the same result shown in trainig progress on dev set)
trainer.evaluate(dev_dataset, language=LANGUAGE)
```

```
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.1972
Adjusted CER: 0.1169
LATTEScore: 69.30%
Un-adjusted WER: 0.2498
Un-adjusted CER: 0.1640
===============
{'eval_loss': 0.6597035527229309,
 'eval_wer': 0.19719239518466153,
 'eval_cer': 0.1169473428980371,
 'eval_lattescore': 69.2982456140351,
 'eval_runtime': 116.9009,
 'eval_samples_per_second': 2.926,
 'eval_steps_per_second': 0.368,
 'epoch': 2.2988505747126435,
 'num_input_tokens_seen': 1727040000}
```

On TEST set

```
# run on dev-set
# (should give the same result shown in trainig progress on dev set)
trainer.evaluate(test_dataset, language=LANGUAGE)
```

```
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.1228
Adjusted CER: 0.0681
LATTEScore: 77.73%
Un-adjusted WER: 0.1358
Un-adjusted CER: 0.0717
===============
{'eval_loss': 0.5866317749023438,
 'eval_wer': 0.12280530354136088,
 'eval_cer': 0.06807830036387313,
 'eval_lattescore': 77.73049645390071,
 'eval_runtime': 242.9586,
 'eval_samples_per_second': 2.902,
 'eval_steps_per_second': 0.366,
 'epoch': 2.2988505747126435,
 'num_input_tokens_seen': 1727040000}
```

speaker_metadata.tsv: 100%

9.27k/9.27k [00:00<00:00, 688kB/s]

Generating train split:

52/0 [00:00<00:00, 1910.89 examples/s]

```
Speaker metadata loaded: (52, 6)
Columns: ['speaker_id', 'gender', 'age', 'severity_speech_impairment',
'type_nonstandard_speech', 'etiology']
  speaker_id  gender    age  \
0     KES001  Female  30-40
1     KES002  Female  30-40
2     KES003    Male  25-30
3     KES004    Male  25-30
4     KES005    Male  18-24

                         severity_speech_impairment  \
0                      Severe (frequent breakdowns)
1                      Severe (frequent breakdowns)
2  Profound (communication very difficult or impo...
3                      Severe (frequent breakdowns)
4           Moderate (requires effort to understand)

             type_nonstandard_speech              etiology
0                         Dysarthria         Cerebral Palsy
1                         Dysarthria         Cerebral Palsy
2  Stuttering (Disfluency Disorders)         Cerebral Palsy
3  Stuttering (Disfluency Disorders)  Neurological disorder
4  Stuttering (Disfluency Disorders)  Neurological disorder
Generating predictions...

Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.1972
Adjusted CER: 0.1169
LATTEScore: 69.30%
Un-adjusted WER: 0.2498
Un-adjusted CER: 0.1640
===============

Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.1228
Adjusted CER: 0.0681
LATTEScore: 77.73%
Un-adjusted WER: 0.1358
Un-adjusted CER: 0.0717
===============
Dev WER: 0.361 | Word Accuracy: 63.9% | LATTEScore: 50.6%
```

```
Test WER: 0.274 | Word Accuracy: 72.6% | LATTEScore: 64.4%

Dev DataFrame shape: (342, 11)
Columns: ['speaker_id', 'reference', 'prediction', 'wer', 'word_accuracy',
'lattescore_meaning_preserved', 'gender', 'age', 'severity_speech_impairment',
'type_nonstandard_speech', 'etiology']

Test DataFrame shape: (705, 11)
Columns: ['speaker_id', 'reference', 'prediction', 'wer', 'word_accuracy',
'lattescore_meaning_preserved', 'gender', 'age', 'severity_speech_impairment',
'type_nonstandard_speech', 'etiology']

=== Model Deployment Analysis ===
LATTEScore: 64.4%
Deployment Threshold: 80.0%
❌ RECOMMENDATION: Model does not meet quality standards
   Consider: More training data, hyperparameter tuning, or different architecture

=== FILES SAVED ===
dev_predictions.csv: 342 samples
test_predictions.csv: 705 samples

=== DATA PREVIEW ===
  speaker_id                                          reference  \
0     KES004  it seems like some some fish or some seafood i...
1     KES004  maybe it is kinda some milk or some some some ...
2     KES004  it s a cake of course but am not a good fan of...
3     KES004  evening walks maybe around the city or in the ...
4     KES004  maasai culture of course you will find them do...
5     KES004  i can see a giraffe there i think it it it is ...
6     KES004  i can say thats a a burger there on the left h...
7     KES004  fancy and large it is a a large fish there som...
8     KES004  the maasai the kids i think they were in a gat...
9     KES004  whenever i go to the local market i make sure ...

                                        prediction       wer  \
0  It seems like some some some some some some so...  7.933333
1  Maybe it is kind of some meal or some some som...  3.774194
2  It’s a cake of course but I’m not a fan of cak...  0.375000
3  Evening walks maybe, around the city or in the...  0.242424
4  Maasai culture of course you you you will find...  0.382353
5  I can see a giraffe there. I think it it it it...  0.428571
6  I can say that that's a a burger there, there ...  3.896552
7  Fancy and lads it would be a lads fish there s...  0.291667
8  The the Maasai the kings I think the the they ...  0.461538
9  Whenever I go to to the local market, I I make...  0.317073
```

```
   lattescore_meaning_preserved
0                             0
1                             0
2                             0
3                             1
4                             0
5                             0
6                             0
7                             1
8                             0
9                             0


=== DOWNLOAD LINKS ===

dev_predictions.csv

test_predictions.csv

=== NEXT STEPS ===
1. Analyze LATTEScore by speaker metadata (etiology, severity, gender)
2. Compare LATTEScore with WER to see if meaning preservation differs from word accuracy
3. Use LATTEScore for model deployment decisions
4. Calculate LATTEScore breakdown by speaker characteristics
```

From evaluation

```
Using model: cdli/whisper-small_finetuned_kenyan_english_nonstandard_speech_v0.9 with language: en
Number of examples in dataset: 705
```

Get overall results

```
# Finalizing results...

print("Finalizing results...")
results_df = prepare_results(results)
print("Getting speaker metadata...")
results_df = add_speaker_metadata(results_df, metadata_df)
print(f"Calculating WER and CER for {len(results_df)} examples...")
results_df = calculate_error_rates(results_df, verbose=False)
results_df.head(5)
```

```
Finalizing results...
Getting speaker metadata...
Calculating WER and CER for 705 examples...
Overall WER (normalized): 0.147
Overall CER (normalized): 0.085
Avg WER (normalized): 0.123
Avg CER (normalized): 0.066
```

	wer		cer	
	mean	count	mean	count
severity				
mild	0.09	3	94.00	3
moderate	0.17	3	68.33	3
severe	0.14	3	72.67	3

	wer		cer	
	mean	count	mean	count
severity				
mild	0.04	3	94.00	3
moderate	0.10	3	68.33	3
severe	0.08	3	72.67	3

			wer		cer	
			mean	count	mean	count
speaker_id	severity	etiology				
KES013	mild	Cerebral Palsy	0.10	62	0.05	62
KES021	mild	Parkinson's Disease	0.09	125	0.04	125
KES030	mild	Neurodevelopmental disorder	0.08	95	0.04	95
KES012	moderate	Neurodevelopmental disorder	0.12	111	0.07	111
KES028	moderate	Cerebral Palsy	0.18	56	0.09	56
KES035	moderate	Multiple Sclerosis (MS)	0.20	38	0.12	38
KES001	severe	Cerebral Palsy	0.11	62	0.07	62
KES002	severe	Cerebral Palsy	0.12	69	0.07	69
KES010	severe	Neurodevelopmental disorder	0.18	87	0.10	87

	wer			cer			
	mean	count	mean	count	mean	count	mean	count
etiology								
Cerebral Palsy	0.13	4	62.25	4	0.07	4	62.25	4
Multiple Sclerosis (MS)	0.20	1	38.00	1	0.12	1	38.00	1
Neurodevelopmental disorder	0.13	3	97.67	3	0.07	3	97.67	3
Parkinson's Disease	0.09	1	125.00	1	0.04	1	125.00	1