"cdli/kenyan_swahili_nonstandard_speech_v0.9"small

Saving to: /jupyter_kernel/trained_models/sw_nonstandard_tune_whisper_small_2/best_model
[600/600 39:34, Epoch 2/3]

Step	Training Loss	Validation Loss	Wer	Cer	Lattescore	Input Tokens Seen
0	No log	0.943650	0.507359	0.219101	17.279412	0
50	0.884000	0.982675	0.511222	0.205066	16.911765	144000000
100	0.885600	0.983821	0.519094	0.232393	18.382353	288000000
150	0.809000	0.925288	0.490694	0.223873	25.000000	432000000
200	0.831300	0.880293	0.455930	0.207017	27.573529	576000000
250	0.568500	0.849920	0.438330	0.196733	29.411765	719760000
300	0.544400	0.845099	0.428024	0.191377	31.985294	863760000
350	0.486500	0.838860	0.426373	0.189715	32.352941	1007760000
400	0.566400	0.837398	0.427506	0.193774	31.617647	1151760000
450	0.480800	0.835407	0.426725	0.195732	32.720588	1295760000
500	0.521400	0.834985	0.425341	0.192795	31.985294	1439520000
550	0.504400	0.835074	0.425016	0.192825	31.985294	1583520000
600	0.625800	0.835121	0.425236	0.192855	31.985294	1727520000

```
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.5074
Adjusted CER: 0.2191
LATTEScore: 17.28%
Un-adjusted WER: 0.6787
Un-adjusted CER: 0.3086
==============
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.5112
Adjusted CER: 0.2051
LATTEScore: 16.91%
Un-adjusted WER: 0.5831
Un-adjusted CER: 0.2444
==============
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.5191
Adjusted CER: 0.2324
LATTEScore: 18.38%
Un-adjusted WER: 0.6821
Un-adjusted CER: 0.3062
==============
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.4907
Adjusted CER: 0.2239
```

```
LATTEScore: 25.00%
Un-adjusted WER: 0.6670
Un-adjusted CER: 0.3226
===============
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.4559
Adjusted CER: 0.2070
LATTEScore: 27.57%
Un-adjusted WER: 0.5790
Un-adjusted CER: 0.2692
===============
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.4383
Adjusted CER: 0.1967
LATTEScore: 29.41%
Un-adjusted WER: 0.5820
Un-adjusted CER: 0.2655
===============
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.4280
Adjusted CER: 0.1914
LATTEScore: 31.99%
Un-adjusted WER: 0.5798
Un-adjusted CER: 0.2670
===============
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.4264
Adjusted CER: 0.1897
LATTEScore: 32.35%
Un-adjusted WER: 0.5726
Un-adjusted CER: 0.2595
===============
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.4275
Adjusted CER: 0.1938
LATTEScore: 31.62%
Un-adjusted WER: 0.5731
Un-adjusted CER: 0.2643
===============
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.4267
Adjusted CER: 0.1957
LATTEScore: 32.72%
Un-adjusted WER: 0.5930
```

```
Un-adjusted CER: 0.2721
===============
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.4253
Adjusted CER: 0.1928
LATTEScore: 31.99%
Un-adjusted WER: 0.5730
Un-adjusted CER: 0.2603
===============
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.4250
Adjusted CER: 0.1928
LATTEScore: 31.99%
Un-adjusted WER: 0.5727
Un-adjusted CER: 0.2604
===============
Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.4252
Adjusted CER: 0.1929
LATTEScore: 31.99%
Un-adjusted WER: 0.5729
Un-adjusted CER: 0.2604
===============
```

There were missing keys in the checkpoint model loaded: ['proj_out.weight'].

```
TrainOutput(global_step=600, training_loss=0.661701435248057, metrics={'train_runtime':
2490.2388, 'train_samples_per_second': 2.891, 'train_steps_per_second': 0.241,
'total_flos': 2.07723771150336e+18, 'train_loss': 0.661701435248057, 'epoch':
2.5210084033613445, 'num_input_tokens_seen': 1727520000})
```

Post training evaluation

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: 0.4253
Adjusted CER: 0.1928
LATTEScore: 31.99%
Un-adjusted WER: 0.5730
Un-adjusted CER: 0.2603
===============
```

```
{'eval_loss': 0.8349847197532654,
 'eval_wer': 0.42534073607522993,
 'eval_cer': 0.19279478039122483,
 'eval_lattescore': 31.985294117647058,
 'eval_runtime': 117.5694,
 'eval_samples_per_second': 2.314,
 'eval_steps_per_second': 0.289,
 'epoch': 2.5210084033613445,
 'num_input_tokens_seen': 1727520000}
```

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: 0.3410
Adjusted CER: 0.1362
LATTEScore: 39.35%
Un-adjusted WER: 0.3800
Un-adjusted CER: 0.1564
===============
```

```
{'eval_loss': 0.806226909160614,
 'eval_wer': 0.34104764837601076,
 'eval_cer': 0.13624787270573063,
 'eval_lattescore': 39.35018050541516,
 'eval_runtime': 257.316,
 'eval_samples_per_second': 2.153,
 'eval_steps_per_second': 0.272,
 'epoch': 2.5210084033613445,
 'num_input_tokens_seen': 1727520000}
```

```
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: 0.4253
Adjusted CER: 0.1928
LATTEScore: 31.99%
Un-adjusted WER: 0.5730
Un-adjusted CER: 0.2603
==============

Sentence transformers not available, using WER-based LATTEScore approximation
=== Metrics ===
Adjusted WER: 0.3410
Adjusted CER: 0.1362
LATTEScore: 39.35%
Un-adjusted WER: 0.3800
Un-adjusted CER: 0.1564
==============
Dev WER: 0.698 | Word Accuracy: 30.2% | LATTEScore: 14.0%
Test WER: 0.460 | Word Accuracy: 54.0% | LATTEScore: 31.4%

Dev DataFrame shape: (272, 11)
Columns: ['speaker_id', 'reference', 'prediction', 'wer', 'word_accuracy',
'lattescore_meaning_preserved', 'gender', 'age', 'severity_speech_impairment',
'type_nonstandard_speech', 'etiology']


Test DataFrame shape: (554, 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: 31.4%
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: 272 samples
test_predictions.csv: 554 samples


=== DATA PREVIEW ===
  speaker_id                                          reference  \
0     KES004  kila disemba familia yetu hukutana kijijini kw...
1     KES004  katika shule ya msingi ya langas ni ni ni nili...
2     KES004  nilipofanya safari ya kitalii katika mbuga ya ...
3     KES004  kila jumapili mama huamka mapena na kutengenez...
4     KES004  nilipokua chuo kikuu nairobi tulikua na na uta...
5     KES004  katika kampaini za usafi mtaani kibera sisi ka...
6     KES004  nilipokuwa soko la gikomba nilikutana na mama ...
7     KES004  siku moja nilisaidia shangazi yangu kuvuna mah...
8     KES004  nilihudhuria harusi ya ndugu yangu katika kaun...
9     KES004  katika safari yangu ya ya kwenda kakamega tuli...


                                           prediction       wer  \
0  Kila Disemba familia yetu hukutana kijijini kw...  0.392857
1  Katika shule ya msingi ya lililanga nilijifunz...  0.243243
2  Nilipofanya safari yaitalika katika mbuga ya M...  0.409091
3  Kila jumapili mama wamka mapema na kutengeneza...  0.260870
4  Nilipokua chukuu Nairobi, tulikwana na utamadu...  0.555556
5  Katika kampeni za usafi mtaani kibera sisi kam...  0.208333
6  Nilipokuwa soko la Gikomba nilikukutana na mam...  0.466667
7  Siku mmoja nilisaidia shangazi yangu kuvuna ma...  0.187500
8  Nilihudhuria harusi ya ndugu yangu katika coun...  0.218750
9  Katika safari yangu ya ya kwenda kamega tulisi...  0.187500


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

=== 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