

Lithuanian speech recognition: effects of dialect training and transcript spelling

Saulius Jarasiunas

Study Lead, Digisensus, Vilnius, Lithuania

<https://digisensus.com> · info@digisensus.com

5 October 2026

Abstract

Enterprise conversation intelligence depends on accurate speech recognition, including when speakers use regional dialects. We test whether adding dialect speech improves Lithuanian speech recognition and whether dialect transcripts should use dialect or standard spelling.

We fine-tune Parakeet-TDT on 420.17 hours of Lithuanian telephone speech, adding up to 82.30 hours of dialect recordings. We compare this with adding other spontaneous speech and test dialect and standard transcript spelling. Training settings are fixed. We evaluate dialect recognition against both spellings and an “either-form” score that accepts either reference form. A separate 2.78-hour Customer service speech test contains previously unseen speakers.

Adding dialect speech reduces word error rate (WER) against dialect references from 41.62 % to 31.30 %, averaged over two training runs (Table 1). Adding other spontaneous speech lowers it only to 40.04 %. In the single-run comparison, the first 25.43 hours provide about two thirds of the dialect-test improvement. Standard-spelling dialect transcripts give the lowest either-form WER among models also trained on telephone speech: 25.97 %.

On Customer service speech, adding dialect recordings lowers WER in each of two training runs, from an average of 44.89 % to 39.43 % with dialect spelling and to 39.06 % with standard spelling; run-to-run variation on this test is larger than on the dialect test. WER on the LIEPA-3 telephone test changes little; that test may share speakers with training. These results support adding dialect recordings for regional speech and customer-service conversations. Transcript spelling affects the output learned and the accuracy measured.

Additional training data	Dialect test: accepted spelling			LIEPA-3	Customer
	Dialect	Standard	Either form	telephone	service speech
None	41.62	35.82	34.22	10.61	44.89
82.36 h other spontaneous speech	40.04	34.10	32.40	10.34	41.66
82.30 h dialect speech, dialect-spelling transcripts	31.30	36.73	27.82	10.76	39.43
82.30 h dialect speech, standard-spelling transcripts	37.48	26.94	25.97	10.56	39.06

Table 1: Main findings: word error rate (%), averaged over two training runs; lower is better. Every model also trains on 420.17 hours of standard telephone speech. The three dialect-test columns score the same recordings using different rules for accepted spelling.

1 Introduction

Digisensus provides conversation intelligence to enterprise customers. This work depends on accurate transcription of telephone conversations, including speech from people who use regional dialects. Recording quality and regional pronunciation can both challenge speech recognition. This study focuses on regional speech: whether dialect recordings improve recognition and how their transcripts should be written. Callers may retain regional pronunciation features even when speaking mostly standard Lithuanian or living outside their region of origin. We therefore also test whether dialect training helps on independent customer-service recordings. We share these findings to help others developing speech recognition for languages with limited training resources.

Lithuanian provides a useful setting for this question. Its regional varieties differ from standard Lithuanian in pronunciation and inflection [3, 4, 8]. A dialect recording can therefore be transcribed using forms that reflect the speaker’s pronunciation or the corresponding standard Lithuanian forms. This choice affects both the model’s training and its evaluation: a recogniser may identify the intended word but receive an error because its spelling differs from the reference transcript.

We investigate three practical questions. How much does recognition improve as more dialect speech is added to standard-speech training data? Does dialect speech help more than the same amount of other spontaneous speech? And should dialect training transcripts use dialect or standard spelling? We also measure changes on standard Lithuanian telephone speech, explore transfer to regions absent from dialect training, and compare held-out recordings from training speakers with recordings from unseen speakers.

To distinguish the effects of the two spellings, we evaluate the same model outputs against paired dialect and standard transcripts. We also use an “either-form” word error rate that accepts either reference spelling of each word. Table 2 illustrates why these scores can differ. Together, the three scores show how strongly the measured result depends on the spelling accepted during evaluation.

Model output for the same spoken word	Dialect reference: <i>gīveno</i>	Standard reference: <i>gyveno</i>	Either form accepted
<i>gīveno</i> — dialect form of “lived”	Correct	Error	Correct
<i>gyveno</i> — standard form of “lived”	Error	Correct	Correct
<i>gavo</i> — a different word, “received”	Error	Error	Error

Table 2: Illustrative scoring example. The dialect and standard reference forms represent the same intended word.

2 Related work

Previous research has examined speech recognition across accents and dialects, including English accents [7], Egyptian Arabic [1], and Swiss German [10]. These studies address both the availability of representative speech recordings and the choice of written transcripts. For dialects without a single established spelling, the same spoken words may have several acceptable written forms.

Existing datasets take different approaches. The Arabic MGB-3 challenge transcribes Egyptian Arabic as spoken and allows multiple spellings [1]. The Swiss Parliaments Corpus pairs Swiss German speech with Standard German text, treating recognition as translation between closely related language varieties [10]. Our study compares dialect and standard spelling using the same Lithuanian recordings and training settings. Paired reference transcripts let us evaluate every model under both spelling conventions and with a score that accepts either form.

3 Data

LIEPA-3 is an openly available corpus of 10,000 hours of annotated Lithuanian speech, developed with European Union funding through NextGenerationEU / Naujos kartos Lietuva for speech recognition and research [12]. It contains read speech, spontaneous speech and 100 hours of dialect recordings, and is published by Vilnius University, Vytautas Magnus University and the Institute of the Lithuanian Language under CC BY 4.0.

We use three parts of LIEPA-3: standard telephone conversations, dialect recordings from four regions, and other spontaneous speech from radio and dictaphone recordings. Telephone speech provides the main training data. Dialect recordings let us test the value of adding regional speech, while other spontaneous recordings provide a comparison for adding a similar amount of data. We also evaluate the models on two public Lithuanian read-speech datasets, FLEURS and Common Voice. We additionally test on private Customer service speech. Table 3 summarises the data and their purposes.

Data	Hours	Purpose
Telephone speech: training	420.17	Train the standard-speech baseline
Dialect speech: training	82.30	Test dialect data and spelling choices; 97 speakers
Other spontaneous speech: training	82.36	Compare with adding a similar amount of other speech
Dialect speech: main test	5.01	Test dialect recognition on 12 unseen speakers
Telephone speech: test	4.36	Measure changes on standard telephone speech
Customer service speech: test	2.78	Independent private test on previously unseen speakers
FLEURS: test	2.97	Additional evaluation on read speech
Common Voice 19.0: test	7.15	Additional evaluation on read speech

Table 3: Main training and test data. The dialect and control training sets have approximately equal durations. Telephone training and test recordings may share speakers.

3.1 Dialect recordings and speaker separation

The dialect recordings cover Aukštaitija, Dzūkija, Suvalkija and Žemaitija. The main test contains three speakers from each region, selected to balance gender and age groups. None of these speakers appears in training.

Another 8 speakers provide 2.67 hours of development recordings used to monitor training. Their recordings are also excluded from training. All remaining recordings from the test and development speakers—9.50 hours—are unused.

A separate 1.01-hour test contains recordings withheld from 12 speakers in the dialect training pool. This lets us compare new recordings from those speakers with recordings from unseen speakers. The withheld recordings themselves are never used for training. Models trained on smaller dialect subsets or with a region excluded may have seen only some of these speakers.

To study how improvement depends on data quantity, we build progressively larger dialect training sets (Table 4). Each larger set includes the smaller one and adds speakers. The repository records the technical split identifiers and construction rules.

Dialect training set	Hours	Speakers
Smallest subset	12.83	16
Second subset	25.43	31
Third subset	50.47	61
Full training pool	82.30	97

Table 4: Increasing amounts of dialect training speech. Each set includes all speakers and recordings from the smaller sets.

3.2 Two transcript versions for the same recordings

The original dialect transcripts represent the words as pronounced. We create a second version using the corresponding standard Lithuanian forms. The standardisation guidelines aim to preserve word order, grammatical meaning and word choice; dialect words without a standard counterpart remain unchanged.

The two versions are aligned word by word, allowing the same model output to be evaluated against either spelling. Standardisation uses language-model assistance, written guidelines and automatic alignment checks. The preparation and review procedure is documented in the repository.

We also label the differences between paired words—for example, changes to endings or pronunciation—to examine which word types remain difficult to recognise (Table 5).

Word class	Meaning	Example: dialect → standard
Unchanged	Same word form in both spellings	–
Ending change	Only the ending differs	būdava → būdavo; seseris → seserys
Letters ą, ę, į, ū	Dialect form uses a, e, i, u in place of these letters	viska → viską; mūsu → mūsų
Other pronunciation change	Another sound change in the stem	giveno → gyveno; teip → taip
Dialect-only word	No standard counterpart; kept unchanged	rozu; bajino

Table 5: Word classes used to analyse recognition errors.

3.3 Telephone speech and the comparison dataset

The telephone data follow the published training and test split of the dataset used in this study. Speaker identities are unavailable, so the two sets may contain recordings from the same callers. We remove training clips whose transcripts of six or more words repeat a test or development transcript. Telephone results therefore describe performance on this particular recording collection; they do not establish performance on entirely unseen callers.

The comparison dataset contains other spontaneous speech, mainly from radio and dictaphone recordings. Its duration, gender distribution, age groups and clip lengths are matched approximately to the dialect training data. This comparison helps assess whether dialect recordings offer more benefit than simply adding more speech.

3.4 Customer service speech

Digisensus provides 2.78 hours of private Customer service speech from speakers absent from the training, development and earlier test sets. We evaluate the existing models on these independent recordings to test whether dialect-training benefits extend beyond LIEPA-3.

3.5 Public read-speech tests

FLEURS [6] and Common Voice [2] test recognition of read speech, which differs from spontaneous telephone conversations. We report these results separately; dataset details are recorded in the protocol.

4 Experimental design

4.1 What we compare

We compare a telephone-only baseline with models fine-tuned from the same pretrained checkpoint on different combinations of training data. We vary the amount and type of added speech, or the spelling of its transcripts. Each comparison answers a specific question (Table 6).

Question	Comparison
How does adding more dialect speech affect recognition?	Add progressively larger dialect datasets: 12.83, 25.43, 50.47 and 82.30 hours.
Does dialect speech help more than simply adding data?	Compare adding 82.30 hours of dialect speech with 82.36 hours of other spontaneous speech.
Which transcript spelling works better?	Train on the same dialect recordings with dialect-spelling or standard-spelling transcripts.
Does the benefit extend to customer-service conversations?	Evaluate all models on independent Customer service speech from previously unseen speakers.
Does the benefit extend to an unrepresented region?	Compare models trained with and without recordings from the region being tested.
Are training speakers easier to recognise?	Compare withheld recordings from speakers in the training pool with recordings from unseen speakers.
Is standard telephone training necessary?	Compare combined training with training on dialect recordings alone.

Table 6: Questions and experimental comparisons.

We also evaluate the original Parakeet-TDT [9] and Whisper large-v3 [11] models without additional training, providing reference points for the fine-tuned models.

4.2 How we train the models

All fine-tuned models start from the same multilingual Parakeet-TDT checkpoint and use the same training settings. Each run trains for 10,000 updates. We evaluate the final checkpoint rather than selecting a checkpoint based on development or test results.

We repeat four main conditions twice: telephone speech alone, telephone speech plus other spontaneous speech, and telephone speech plus the full dialect dataset under each transcript spelling. The repeats use different random seeds, which change aspects such as training-data order. Other conditions have one run.

Keeping the training procedure fixed makes the comparisons consistent, although different training datasets may need different amounts of training to reach their best performance. Full training settings are recorded in the repository.

4.3 How we measure recognition

Word error rate (WER) counts substituted, missing and extra words, divided by the number of reference words, expressed as a percentage. Lower is better. We calculate it across all recordings in each test set.

Before scoring, we normalise capitalisation, punctuation and spacing.

Score	What counts as the correct word form?
Dialect-spelling WER	The form in the dialect transcript.
Standard-spelling WER	The form in the standard transcript.
Either-form WER	Either of the two aligned reference forms.

Table 7: Three scoring rules applied to the same dialect recordings and model outputs. We do not average these scores.

For dialect speech, we report three scores separately (Table 7).

These scores show how the measured result depends on the accepted spelling. The either-form score reduces penalties for choosing between the two reference forms; it does not establish that every remaining error is purely a recognition error.

We describe changes using before-and-after WERs. When reporting a difference, we use percentage points: a fall from 40 % to 30 % is a reduction of 10 percentage points.

4.4 How we assess uncertainty

We compare models on the same test recordings. We estimate 95 % confidence intervals by repeatedly resampling speakers, or individual recordings where speaker identities are unavailable [5]. For Customer service speech, we resample whole calls. These intervals describe uncertainty from the test sample; they do not include variation between training runs or errors in the reference transcripts.

We report variation between repeated training runs separately. Regional comparisons contain only three test speakers per region and should be treated as exploratory. Likewise, the comparison between training speakers and unseen speakers uses different recordings and cannot isolate speaker familiarity from differences in recording difficulty.

4.5 Protocol and changes

The original experimental protocol was frozen after the initial training runs and before their test evaluations. One unadapted model had already been evaluated to check the scoring pipeline. The Suvalkija exclusion experiment was added after the other regional results had been inspected, using the same training settings.

The Customer service speech evaluation used the already-trained models. Its analysis was exploratory: the plan was written after an earlier pilot evaluation on the same recordings.

After scoring, we moved the public read-speech benchmarks to secondary results because their scores varied substantially between training runs. We also did not perform the additional training runs required by the protocol for resolving small effects on telephone speech. We therefore leave the telephone-speech comparison between transcript spellings unresolved. The full protocol history and reference versions are recorded in the repository.

5 Results

Adding dialect recordings improves dialect recognition more than adding a similar amount of other spontaneous speech. Much of the improvement comes from the first additions, and transcript spelling changes both the output and the measured error rate. We present each comparison below. Complete model scores are provided in Appendix A; Table 12 summarises the findings at the end of this section.

5.1 How does adding more dialect speech affect recognition?

We test how dialect recognition changes as more dialect recordings are added to the standard telephone training data. We train models with no dialect speech and with 12.83, 25.43, 50.47 or 82.30 additional hours. These models use dialect-spelling training transcripts and identical training settings.

Dialect recognition improves at every step (Table 8, Figure 1). When evaluated against dialect-spelling references, word error rate falls from 42.11 % without dialect training data to 31.28 % with 82.30 hours. The first 25.43 hours provide about two thirds of this reduction. Further additions continue to help, but the improvements become smaller.

The improvement also appears when either dialect or standard spelling is accepted: word error rate falls from 34.83 % to 27.60 %. This shows that the benefit is not limited to reproducing dialect spelling.

Customer service speech also improves as more dialect recordings are added. In the same single-run comparison, WER falls from 42.61 % without dialect data to 37.01 % with 82.30 hours. The improvement continues across the tested training amounts. This differs from the LIEPA-3 telephone test, where WER changes little.

On the LIEPA-3 telephone test, word error rate changes from 10.76 % to 10.80 %. These telephone values are low relative to the original models (Section 5.6) and likely benefit from shared callers and recording conditions between the telephone training and test sets (Section 3.3). A second training run confirms the overall pattern: dialect-reference WER falls from 41.13 % to 31.31 %, while telephone WER changes from 10.46 % to 10.72 %.

These results suggest that a relatively small dialect collection can provide a substantial improvement, with additional recordings giving diminishing returns. However, each larger dataset also includes more speakers, so this experiment does not separate the effect of recording hours from speaker diversity.

Dialect speech added	Dialect-test WER: dialect spelling	Dialect-test WER: either spelling	LIEPA-3 telephone WER	Customer service speech WER
None	42.11 %	34.83 %	10.76 %	42.61 %
12.83 hours	38.08 %	33.30 %	10.61 %	41.48 %
25.43 hours	34.77 %	30.29 %	10.61 %	40.30 %
50.47 hours	32.77 %	28.96 %	10.49 %	37.91 %
82.30 hours	31.28 %	27.60 %	10.80 %	37.01 %

Table 8: One training run per condition (seed 1), using dialect-spelling transcripts. Lower WER is better. The abstract reports averages of two runs for the baseline and full-data models, which explains the slightly different values there.

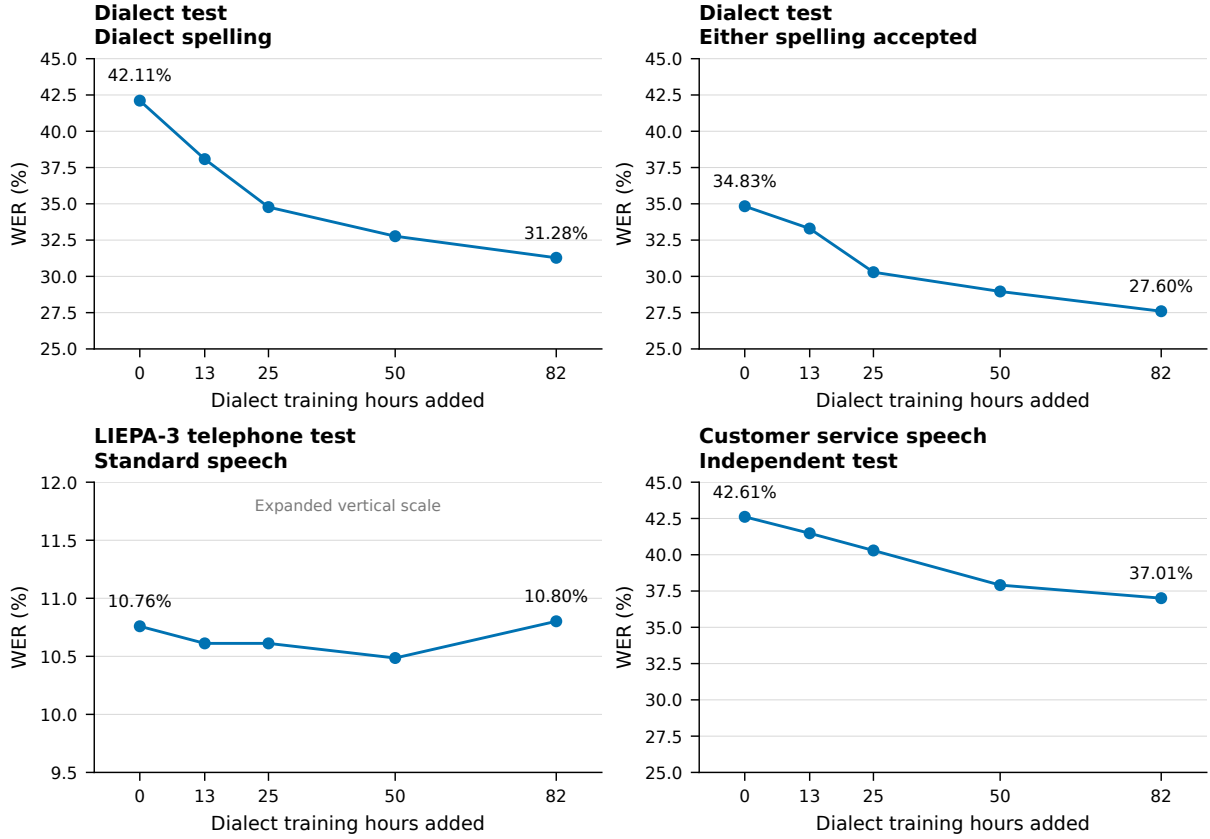


Figure 1: WER as dialect training hours increase, using dialect-spelling transcripts and seed 1 throughout. Top left: dialect references. Top right: either spelling accepted. Bottom left: LIEPA-3 telephone speech, shown on an expanded vertical scale. Bottom right: Customer service speech. All panels except the telephone panel share the same vertical scale.

5.2 Does dialect speech help more than simply adding more recordings?

We compare two additions to the same standard telephone training data: 82.30 hours of dialect speech and 82.36 hours of other spontaneous speech. The dialect recordings use dialect-spelling transcripts. This comparison tests whether the improvement comes simply from adding more training data.

Both additions improve dialect recognition, but dialect recordings provide a much larger benefit (Table 9). Averaged over two training runs, WER against dialect references falls from 41.62 % to 40.04 % with other spontaneous speech, and to 31.30 % with dialect speech.

The advantage remains when either spelling is accepted: WER reaches 32.40 % with other spontaneous speech and 27.82 % with dialect speech. The result therefore does not depend solely on matching dialect spelling.

On the LIEPA-3 telephone test, the other spontaneous recordings give the lower WER: 10.34 %, compared with 10.76 % after adding dialect recordings.

On the independent Customer service speech test, average WER falls from 44.89 % with telephone-only training to 41.66 % with other spontaneous speech and 39.43 % with dialect speech. Both dialect-training runs improve over their corresponding telephone-only run: from 42.61 % to 37.01 % and from 47.17 % to 41.86 % (Figure 2). Resampling calls gives a reduction of 5.46 percentage points with a 95 % interval of 4.85–6.13, but this interval reflects only the test sample: the two telephone-only runs differ by more than its width, so run-to-run variation, not the test sample, is the main uncertainty on this test. Average WER is lower than with other spontaneous speech, but that advantage occurs in only one of the two training runs.

These results show that the type of additional speech matters. For recognising dialect speakers in this study, dialect recordings help substantially more than a similar amount of other spontaneous speech.

Additional training speech	Dialect WER: dialect spelling	Dialect WER: either spelling	LIEPA-3 telephone WER	Customer service speech WER
None	41.62 %	34.22 %	10.61 %	44.89 %
Other spontaneous speech	40.04 %	32.40 %	10.34 %	41.66 %
Dialect speech	31.30 %	27.82 %	10.76 %	39.43 %

Table 9: Mean WER of two training runs. All models also train on 420.17 hours of standard telephone speech. Lower WER is better.

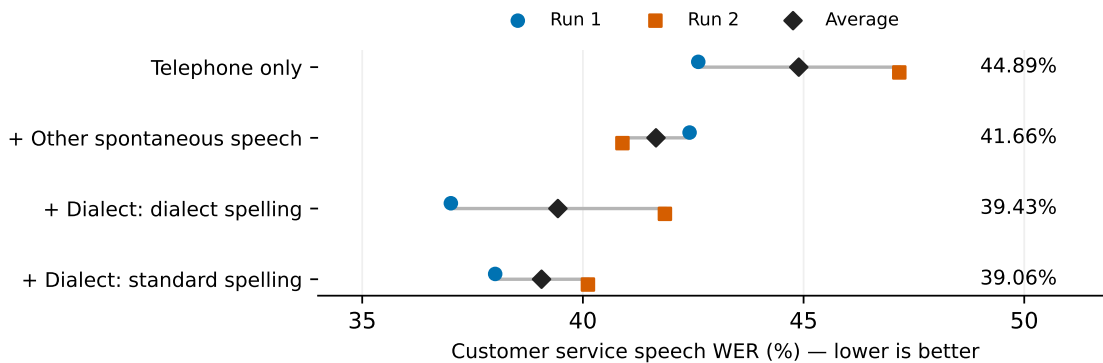


Figure 2: Customer service speech: WER for each training run and their average. All conditions include the same telephone training data; the additions use approximately equal hours. Lower is better.

5.3 Which transcript spelling works better?

We compare two training conditions, each repeated twice, using the same standard telephone data and the same 82.30 hours of dialect recordings. Only the dialect transcripts differ: one condition uses dialect spelling, the other standard spelling. We evaluate both conditions against the same test recordings using all three scoring methods (Figure 3).

The reference spelling strongly affects the measured improvement. With dialect-spelling training transcripts, WER against dialect references falls from 41.62 % to 31.30 %. Yet against standard references, WER for the same training condition changes from 35.82 % to 36.73 %. Its apparent improvement therefore depends on which spelling the test accepts.

When either spelling is accepted, both training approaches improve on the model without dialect data. WER falls from 34.22 % to 27.82 % with dialect-spelling transcripts and to 25.97 % with standard-spelling transcripts. Standard-spelling transcripts give the lower either-form WER in both repeated training runs.

Among models trained with the standard telephone data, standard-spelling dialect transcripts therefore give the best overall result under either-form scoring. The same pattern appears at the smaller training amounts of 25.43 and 50.47 hours (Figure 4). This experiment does not establish why standard spelling helps. One possibility is that the benefit is partly stylistic: the standard-spelling training transcripts and the standard-spelling test references were produced by the same standardisation procedure (Section 3.2), so a model trained on that procedure’s output is also scored against it. The dialect-spelling score and the dialect-only word class, where standard spelling does worse, are not affected by this.

On the LIEPA-3 telephone test, average WER is 10.76 % with dialect-spelling transcripts and 10.56 % with standard-spelling transcripts. Because this difference is small relative to the observed variation between training runs, we leave this comparison unresolved.

On Customer service speech, average WER is 39.43 % with dialect-spelling transcripts and 39.06 % with standard-spelling transcripts. Both improve over telephone-only training in each repeated run, but the small difference between spellings is inconclusive. The lowest observed customer-service WER is 36.14 %, using 50.47 hours of dialect speech with standard spelling. That condition has one run and does not establish an optimal training amount. Appendix A reports all evaluated models.

There is also a trade-off for dialect-only vocabulary. In the single-run word-class analysis, the error rate attributed to words without a standard counterpart is 62.12 % with dialect-spelling training transcripts and 73.84 % with standard-spelling transcripts. This measure includes extra words assigned to that word class during scoring. The better overall score therefore does not mean that every type of word improves. Detailed word-class results are retained in the repository.

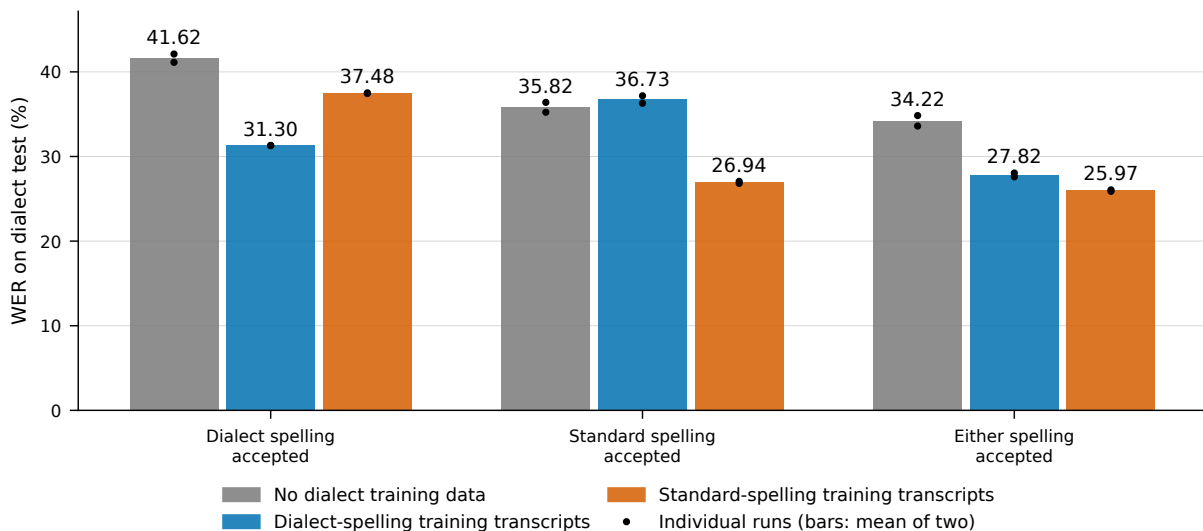


Figure 3: The same dialect test scored three ways. Bars show mean WER over two training runs; dots show the individual runs. All models also train on standard telephone speech; the two dialect-trained models add the same 82.30 hours of recordings. Lower WER is better.

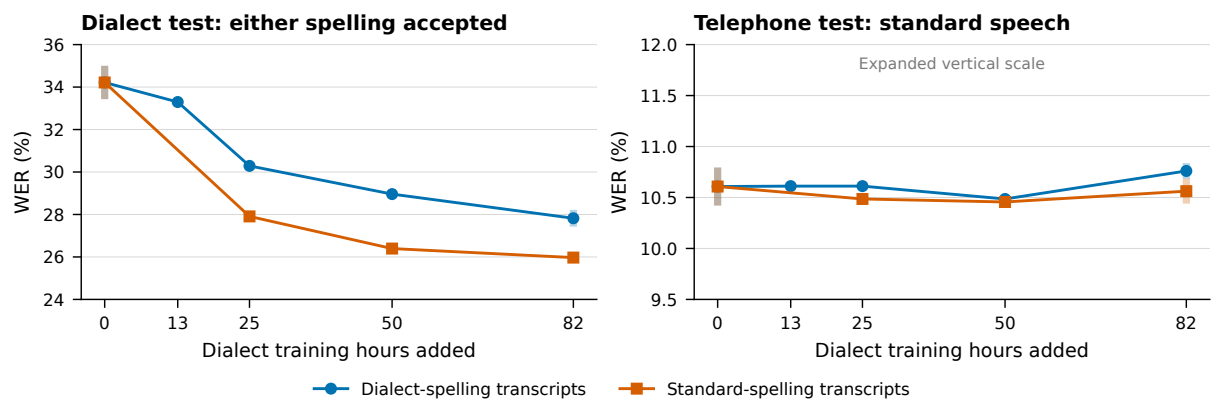


Figure 4: Transcript-spelling comparison as dialect training hours increase. Points show means where two runs are available and single runs otherwise; vertical segments show the range of the two runs. The telephone panel uses an expanded vertical scale. All models also train on standard telephone speech.

5.4 Does dialect training help a region excluded from training?

We test whether training on dialect speech from other regions helps recognise a region whose dialect recordings were excluded from training. The test covers all four regions: Aukštaitija, Dzūkija, Suvalkija and Žemaitija. It contains 12 test speakers in total, three from each region. None of these speakers appears in the training data.

For each region, Figure 5 compares three training conditions: telephone speech alone; telephone speech plus dialect recordings from the other three regions; and telephone speech plus dialect recordings from all four regions. Each condition uses one training run. Dialect training transcripts use standard spelling, and scoring accepts either dialect or standard spelling. Within each region, all three conditions are evaluated on the same recordings.

Training on the other three regions lowers WER in every region compared with telephone-only training: from 34.52 % to 26.55 % in Aukštaitija, from 40.13 % to 30.77 % in Dzūkija, from 27.46 % to 22.37 % in Suvalkija, and from 36.40 % to 29.90 % in Žemaitija. This supports a benefit from dialect training even when the tested region is absent from the dialect training data.

Training on all four regions gives the lowest measured WER in each comparison. However, the additional improvement for Žemaitija is uncertain: its 95 % confidence interval includes no improvement. These are exploratory results based on three test speakers per region and one run per condition. Including the tested region also adds training hours, so we cannot separate the value of that region’s speech from the benefit of more data. The timing of the supplementary Suvalkija experiment is documented in Section 4.5.

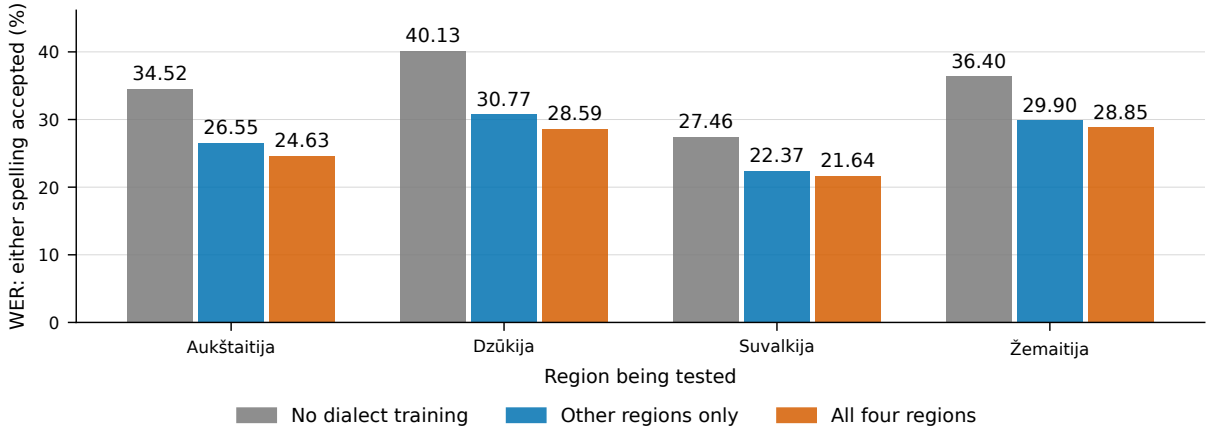


Figure 5: Dialect recognition across four regions, evaluated on 12 test speakers in total (three per region). Each group compares telephone-only training, adding dialect speech from the other three regions, and adding dialect speech from all four regions. Bars show WER with either dialect or standard spelling accepted, using one training run per condition. Lower is better.

5.5 Are speakers represented in training easier to recognise?

We compare two sets of dialect recordings: withheld recordings from 12 speakers in the dialect training pool, and recordings from 12 speakers absent from training. None of the evaluated recordings was used for training. We expected models trained on the first group’s other recordings to recognise those speakers more accurately.

The observed pattern is the opposite (Figure 6). Across all 19 fine-tuned training runs, WER is higher on the recordings from speakers in the training pool, by 0.77 to 2.41 percentage points. Both sets are scored by accepting either dialect or standard spelling.

However, the same pattern appears in models trained without dialect speech, which had not encountered either group during fine-tuning. Models trained on smaller dialect subsets or with a

region excluded had encountered only some speakers in the training pool. These results suggest that differences in the difficulty of the two recording sets affect the comparison.

We therefore cannot conclude that speaker familiarity provides no benefit. The two sets contain different speakers and recordings, so this experiment does not isolate its effect. It shows only that the withheld recordings from speakers in the training pool were harder for these models than the recordings from unseen speakers.

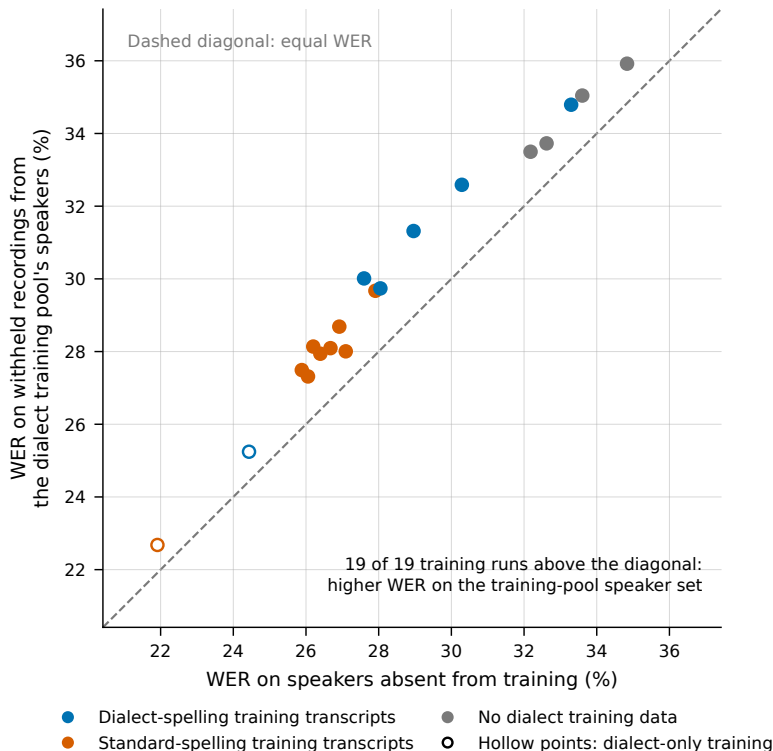


Figure 6: Each point represents one fine-tuned training run. Both axes show either-form WER (%). The dashed diagonal marks equal WER on both sets. Points above it indicate higher error on the withheld recordings from speakers in the dialect training pool. Speaker overlap depends on the model’s training data; the two recording sets are the same across all runs.

5.6 What happens when we train on dialect speech alone?

We compare fine-tuning on dialect recordings alone with fine-tuning on standard telephone recordings alone or on both datasets together. “Dialect-only” refers to the data used for fine-tuning; all these models start from the same pretrained multilingual Parakeet-TDT model.

Using only 82.30 hours of dialect speech with standard-spelling transcripts gives the lowest either-form dialect WER observed in this study: 21.91 %. However, its telephone-test WER is 19.63 %. Combining the dialect recordings with standard telephone training data gives a higher dialect WER of 25.97 %, but a much lower telephone WER of 10.56 %. The size of this telephone gap partly reflects the possible speaker and channel overlap of the telephone test (Section 3.3); on the independent Customer service test the gap is smaller (Table 10).

Dialect-only training with dialect-spelling transcripts shows a similar trade-off: either-form dialect WER is 24.43 %, while telephone WER is 28.20 %.

The original Parakeet-TDT model, without fine-tuning on our study data, reaches 68.15 % either-form dialect WER and 42.53 % telephone WER; Whisper large-v3 reaches 53.15 % and 27.43 %, respectively. Fine-tuning Parakeet-TDT on telephone speech alone improves both scores relative to these original models, reaching 34.22 % and 10.61 % (Table 10).

These results show why the best training choice depends on the intended use. Dialect-only fine-tuning performs best on this dialect test, while combined training supports both dialect and telephone recognition. The dialect-only conditions each have one training run, so their apparent advantage needs confirmation with repeated runs.

Model and fine-tuning data	Dialect WER: either spelling	LIEPA-3 telephone WER	Customer service speech WER
Original Parakeet-TDT, no study fine-tuning	68.15 %	42.53 %	59.51 %
Original Whisper large-v3, no study fine-tuning	53.15 %	27.43 %	54.67 %
Parakeet-TDT: telephone speech only	34.22 %	10.61 %	44.89 %
Parakeet-TDT: dialect speech only, dialect spelling	24.43 %	28.20 %	51.36 %
Parakeet-TDT: dialect speech only, standard spelling	21.91 %	19.63 %	42.59 %
Parakeet-TDT: telephone + dialect speech, standard spelling	25.97 %	10.56 %	39.06 %

Table 10: Lower WER is better. Telephone-only and combined-training values are means of two runs; dialect-only values come from one run. Original models are evaluated without additional training. Telephone and dialect training sets contain 420.17 and 82.30 hours, respectively.

5.7 Do the improvements carry over to public read-speech tests?

We also evaluate every model on FLEURS and Common Voice, two public Lithuanian datasets containing people reading aloud. These tests examine performance on a different type of speech from the spontaneous telephone conversations used for most training.

Table 11 distinguishes two comparisons: the effect of telephone fine-tuning relative to the stock model, and the effect of adding dialect speech relative to telephone-only training. Original Parakeet-TDT reaches 22.10 % WER on FLEURS and 16.52 % on Common Voice. After fine-tuning on telephone speech alone, the two training runs reach 28.62 % and 28.65 % on FLEURS, and 35.20 % and 44.74 % on Common Voice.

Adding the full dialect dataset improves both benchmarks relative to telephone-only training in both runs, under each transcript spelling. The size of the improvement varies between runs. Every fine-tuned model has higher WER than stock Parakeet on Common Voice, while some achieve lower WER on FLEURS. For example, with standard-spelling dialect transcripts, FLEURS WER is 26.87 % in one run and 21.51 % in the other. This variation makes small differences between training conditions difficult to interpret.

The recording conditions also differ. Most training speech has telephone bandwidth and is spontaneous, whereas the benchmarks contain wideband read speech. Dialect recordings are also wideband, so adding them changes both regional coverage and recording conditions. This experiment does not establish which differences cause the benchmark behaviour.

We therefore treat these tests as secondary results. They show that adding the full dialect dataset improves read-speech recognition relative to telephone-only training in both repeated runs. Compared with stock Parakeet, improvements occur on FLEURS for some conditions, but not on Common Voice. Complete scores are provided in Appendix A; the additional benchmark chart and diagnostics remain in the repository.

Fine-tuning data	FLEURS WER, run 1 / run 2	Common Voice WER, run 1 / run 2
None: original Parakeet-TDT	22.10 %	16.52 %
Telephone speech only	28.62 % / 28.65 %	35.20 % / 44.74 %
Telephone + dialect speech, dialect spelling	23.24 % / 22.58 %	25.77 % / 27.99 %
Telephone + dialect speech, standard spelling	26.87 % / 21.51 %	27.75 % / 23.85 %

Table 11: Lower WER is better. Each fine-tuned condition has two training runs; the original model has no additional training. Combined models add 82.30 hours of dialect speech to 420.17 hours of telephone speech.

5.8 Summary of findings

The main findings are summarised in Table 12. Conclusions apply to the datasets, model and training settings tested in this study. The original protocol hypotheses and verdicts are preserved in the repository.

Question	Finding
Does adding dialect speech help?	Yes. In the single-run series using dialect-spelling transcripts, dialect and Customer service speech WER fall at every tested addition. Later additions give smaller dialect-test gains. LIEPA-3 telephone WER changes little (§5.1).
Is the improvement simply from adding more data?	Dialect recordings improve dialect recognition substantially more than a similar amount of other spontaneous speech (§5.2).
Which transcript spelling works better?	Among models also trained on telephone speech, standard-spelling dialect transcripts give the lowest either-form WER. Dialect-only vocabulary remains a weakness (§5.3).
Does the reference spelling affect the result?	Yes. The same model can show a large improvement against dialect references and no improvement against standard references (§5.3).
Does standard spelling improve LIEPA-3 telephone recognition?	Unresolved. The observed difference is small relative to variation between training runs (§5.3).
Does dialect training help an excluded region?	Yes in all four regional comparisons, covering 12 test speakers in total. Each uses one run per condition. Regional coverage and training hours are mixed (§5.4).
Are speakers represented in training easier to recognise?	The withheld recordings from those speakers have higher WER. Different speakers and recordings prevent a conclusion about the effect of familiarity (§5.5).
Can we train on dialect recordings alone?	This gives the lowest dialect-test WER, but higher telephone-test WER than combined training. The dialect-only results each come from one run (§5.6).
Does dialect training help customer-service speech?	Yes on this independent private test: both repeated dialect-training conditions improve over telephone-only training (§5.2–5.3).
Do improvements extend to public read-speech tests?	Yes relative to telephone-only training in both repeated runs, but not consistently relative to stock Parakeet (§5.7).

Table 12: Summary of findings, with links to the supporting results.

6 Discussion

For conversation-intelligence systems such as those developed by Digisensus, these results support collecting dialect recordings alongside standard telephone speech. Dialect data improve recognition of regional speakers more than a similar amount of other spontaneous speech. Much of the measured improvement comes from the first additions, suggesting that a modest, carefully selected dialect collection can be useful. However, this study increases recording hours and speaker diversity together, so it does not establish a minimum number of hours or speakers (§5.1–5.2).

The independent Customer service speech results strengthen the practical case: adding dialect recordings also improves recognition of previously unseen customer-service speakers. Retained regional pronunciation is one possible explanation, but the test does not label dialect features or callers’ regional origins. The improvement alone does not establish its cause.

Transcript spelling should reflect the application’s intended output. In this study, standard-spelling dialect transcripts give the lowest either-form error among models also trained on telephone speech. They are therefore a promising choice for applications that produce standard Lithuanian text. The gain is not uniform, however: on words that have no standard counterpart, standard-spelling training raised the error rate from 62.12% to 73.84% in the single-run analysis. Applications that need dialect forms, or that depend on such words, have a different requirement and should evaluate this word class separately (§5.3).

Training-data choices also involve a trade-off between speech types. Dialect-only fine-tuning gives the lowest error on the dialect test, while combining dialect and telephone recordings gives much lower error on this telephone collection, whose test set may share callers with training, and lower error on Customer service speech. On public read-speech tests, adding dialect speech improves both repeated runs relative to telephone-only training, but does not consistently beat stock Parakeet. Practical model selection should therefore use test recordings representative of the intended service (§5.6–5.7).

Evaluation should make the accepted spelling explicit. Scoring the same output against dialect and standard references can produce different conclusions about improvement. Paired reference transcripts and either-form scoring help reveal this dependence. Reporting all three scores allows readers to distinguish accuracy under a required output convention from accuracy when either reference form is acceptable (§4.3, §5.3).

The findings motivate similar experiments in other languages with regional variation, but they do not establish that the same improvements will occur elsewhere. The training data come from one Lithuanian corpus; we study one pretrained model family and one training procedure, with supplementary evaluation on a private customer-service collection. They also do not directly measure improvements under noise, compression or poor telephone connections.

7 Limitations

Transcript accuracy. The standard-spelling transcripts were prepared with language-model assistance and automatic checks. Errors in these transcripts can affect both training and evaluation. The team does not include a professional dialectologist, and the standardisation guidelines reflect practical decisions that others may judge differently. Because training and test standard-spelling transcripts share this procedure, part of the advantage of standard-spelling training may reflect consistency with the references rather than better recognition (§5.3). The preparation procedure is recorded in the repository.

Test coverage. The main dialect test contains 12 speakers, with only three per region. Regional findings therefore remain exploratory. The LIEPA-3 telephone training and test sets may share speakers, so their results do not establish performance on entirely unseen callers. The comparison of familiar and unfamiliar speakers also uses different recordings and cannot isolate the effect of speaker familiarity.

Experimental coverage. We fine-tuned one pretrained model family using fixed training settings and 10,000 updates per run. Different datasets may need different training schedules to reach their best performance. Four main conditions have two training runs; the remaining conditions have one. More dialect hours also introduce more speakers, and excluding a region reduces the training data. These comparisons therefore combine several possible influences.

Scope of the conclusions. The small LIEPA-3 telephone-test difference between transcript spellings remains unresolved. Results on public read-speech datasets vary between training runs and differ from the telephone and dialect results. We did not isolate the effects of recording bandwidth, speaking style, noise or compression. The independent Customer service speech test demonstrates improvement on one private collection, but does not establish the same benefit across other customer populations, languages or models. Its speakers have no dialect labels, so we cannot attribute the gain specifically to regional pronunciation.

8 Conclusion

Adding dialect recordings improves Lithuanian dialect recognition more than adding a similar amount of other spontaneous speech. The benefit also extends to independent Customer service speech: both repeated runs improve over telephone-only training under each transcript spelling. This supports combining telephone and dialect recordings for conversation-intelligence applications.

Standard-spelling dialect transcripts give the lowest either-form WER among models also trained on telephone speech. However, no training choice is best on every test or word class, and this study does not establish an optimal amount of dialect data. Model selection should use recordings representative of the intended service, with the accepted transcript spelling stated explicitly.

Reproducibility and resources

Study code, training configurations, frozen splits, the run registry and analysis scripts are released at <https://github.com/Digisensus/lithuanian-dialect-asr> (tag v1.0, MIT licence). The public analysis regenerates the reported results and tables from the released evaluation records.

Data and models are on the Hugging Face Hub (CC BY 4.0):

- Dialect speech with speaker-disjoint splits, training subsets and the standard-spelling transcript layer: <https://huggingface.co/datasets/Digisensus/lithuanian-dialect-speech-liepa-3-100h-punctuated> (tag v2.0).
- Telephone speech with the study split lists: <https://huggingface.co/datasets/Digisensus/lithuanian-phone-speech-liepa-3-429h-punctuated> (tag v2.0).
- Evaluation records for the public test sets (references, hypotheses of all 21 evaluated systems, scores and the control pool): <https://huggingface.co/datasets/Digisensus/lithuanian-dialect-asr-eval> (tag v1.0).
- Fine-tuned models, run 1 of each headline condition: <https://huggingface.co/Digisensus/parakeet-tdt-0.6b-1t-study-s>, [-s-d80-dial](https://huggingface.co/Digisensus/parakeet-tdt-0.6b-1t-study-s-d80-dial), [-s-d80-std](https://huggingface.co/Digisensus/parakeet-tdt-0.6b-1t-study-s-d80-std), [-s-x80](https://huggingface.co/Digisensus/parakeet-tdt-0.6b-1t-study-s-x80), [-d80-dial](https://huggingface.co/Digisensus/parakeet-tdt-0.6b-1t-study-s-d80-dial) and [-d80-std](https://huggingface.co/Digisensus/parakeet-tdt-0.6b-1t-study-s-d80-std) (tag v1.0). Other checkpoints are available on request.

Customer service speech audio, transcripts and model outputs remain private; only aggregate results are reported here.

References

- [1] Ahmed Ali, Stephan Vogel, and Steve Renals. Speech recognition challenge in the wild: Arabic MGB-3. In *Proc. IEEE ASRU*, pages 316–322, 2017. doi: 10.1109/ASRU.2017.8268952.
- [2] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proc. LREC*, 2020.
- [3] Rima Bakšienė, Agnė Čepaitienė, Jurgita Jaroslavienė, and Jolita Urbanavičienė. Standard Lithuanian. *Journal of the International Phonetic Association*, 54(1):414–444, 2024. doi: 10.1017/S0025100323000105.
- [4] Laimute Balode and Axel Holvoet. The Lithuanian language and its dialects. In Östen Dahl and Maria Koptjevskaja-Tamm, editors, *Circum-Baltic Languages, Volume 1: Past and Present*, Studies in Language Companion Series 54, pages 41–79. John Benjamins, 2001. doi: 10.1075/slcs.54.05bal.
- [5] Maximilian Bisani and Hermann Ney. Bootstrap estimates for confidence intervals in ASR performance evaluation. In *Proc. ICASSP*, 2004.
- [6] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. FLEURS: Few-shot learning evaluation of universal representations of speech. In *Proc. IEEE SLT*, 2022.
- [7] Arthur Hinsvark, Natalie Delworth, Miguel Del Rio, Quinten McNamara, Joshua Dong, Ryan Westerman, Michelle Huang, Joseph Palakapilly, Jennifer Drexler, Ilya Pirkin, Nishchal Bhandari, and Miguel Jette. Accented speech recognition: A survey. arXiv:2104.10747, 2021.
- [8] Gintarė Judžentytė-Šinkūnienė and Saulė Nikartaitė. The concept of Samogitianness in the Northern Samogitian dialect. *Valoda: nozīme un forma*, 11:39–62, 2020. doi: 10.22364/vnf.11.03.
- [9] NVIDIA NeMo team. parakeet-tdt-0.6b-v3: a multilingual FastConformer-TDT speech recognition model. <https://huggingface.co/nvidia/parakeet-tdt-0.6b-v3>, 2025.

- [10] Michel Plüss, Lukas Neukom, Christian Scheller, and Manfred Vogel. Swiss Parliaments Corpus, an automatically aligned Swiss German speech to Standard German text corpus. In *Proc. Swiss Text Analytics Conference (SwissText) 2021, CEUR Workshop Proceedings vol. 2957*, 2021. URL <https://ceur-ws.org/Vol-2957/paper3.pdf>.
- [11] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*, 2023.
- [12] Vilnius University, Vytautas Magnus University, and Institute of the Lithuanian Language. Large Lithuanian language speech corpus (LIEPA-3). CLARIN-LT repository, <https://hdl.handle.net/20.500.11821/101>, 2026. CC BY 4.0.

A Complete results

Table 13 reports WER for every evaluated model on every test set. Each row is one training run; the main text averages repeated runs where stated. Lower WER is better.

Model and training data	Dialect test: accepted spelling				Training	LIEPA-3	Customer	Common	
	Run	Dialect	Standard	Either	speakers	telephone	service	FLEURS	Voice
Stock Parakeet-TDT	–	71.67	69.09	68.15	69.10	42.53	59.51	22.10	16.52
Stock Whisper large-v3	–	59.36	54.57	53.15	51.77	27.43	54.67	23.17	28.09
Telephone only	1	42.11	36.40	34.83	35.92	10.76	42.61	28.62	35.20
Telephone only	2	41.13	35.23	33.60	35.04	10.46	47.17	28.65	44.74
Telephone + dialect, 12.83 h; dialect spelling	1	38.08	38.88	33.30	34.79	10.61	41.48	24.32	30.35
Telephone + dialect, 25.43 h; dialect spelling	1	34.77	37.29	30.29	32.59	10.61	40.30	21.89	24.24
Telephone + dialect, 50.47 h; dialect spelling	1	32.77	37.09	28.96	31.32	10.49	37.91	23.22	29.32
Telephone + dialect, 82.30 h; dialect spelling	1	31.28	36.29	27.60	30.01	10.80	37.01	23.24	25.77
Telephone + dialect, 82.30 h; dialect spelling	2	31.31	37.17	28.05	29.74	10.72	41.86	22.58	27.99
Telephone + other spontaneous speech, 82.36 h	1	39.90	33.86	32.18	33.50	10.31	42.42	24.95	34.07
Telephone + other spontaneous speech, 82.36 h	2	40.17	34.33	32.62	33.73	10.38	40.90	32.00	40.61
Telephone + dialect, 25.43 h; standard spelling	1	38.31	29.14	27.91	29.67	10.49	40.04	24.22	30.50
Telephone + dialect, 50.47 h; standard spelling	1	37.67	27.47	26.39	27.94	10.46	36.14	20.35	18.99
Telephone + dialect, 82.30 h; standard spelling	1	37.52	27.06	26.06	27.32	10.65	38.01	26.87	27.75
Telephone + dialect, 82.30 h; standard spelling	2	37.44	26.82	25.88	27.49	10.47	40.11	21.51	23.85
Dialect only, 82.30 h; dialect spelling	1	27.18	35.32	24.43	25.24	28.20	51.36	41.54	39.74
Dialect only, 82.30 h; standard spelling	1	35.15	22.72	21.91	22.68	19.63	42.59	30.34	26.71
Telephone + dialect excluding Aukštaitija; standard spelling	1	38.00	27.95	26.92	28.69	10.53	37.98	21.38	20.33
Telephone + dialect excluding Dzūkija; standard spelling	1	38.19	28.10	27.09	28.01	10.58	38.86	25.64	35.81
Telephone + dialect excluding Suvalkija; standard spelling	1	38.13	27.58	26.67	28.09	10.61	40.20	27.37	34.35
Telephone + dialect excluding Žemaitija; standard spelling	1	37.44	27.37	26.20	28.14	10.64	41.72	24.43	32.43

Table 13: Word error rate (%). The three dialect-test columns accept dialect spelling, standard spelling or either form. “Training speakers” means withheld recordings from speakers in the dialect training pool, scored with either form accepted; speaker exposure depends on the training subset. Stock models receive no study fine-tuning.