MEZZANINE — Fundamental Research for the Development of Speech Resources and Technologies for Slovenian, project ID: **J7-4642**

TECHNICAL REPORT

Activity A1.4-T

Knowledge Transfer Efficiency and Improvement Techniques for Slovenian Speech and Speaker Recognition

Authors	Marko Bajec, Iztok Lebar Bajec, Sajid Tauqeer
Lead Institution	University of Ljubljana, Faculty of Computer and Information Science (UL FRI) Laboratory for Data Technologies
Funding	Slovenian Research Agency (ARIS) – Large Fundamental Research Project
Project Duration	36 months (M1–M36)
Task Duration	M1–M36 (14 person-months)
Task Lead	Prof. dr. Marko Bajec
Report Date	March 2026

Executive Summary

This technical report presents the complete scientific findings from Activity A1.4-T of the MEZZANINE project — a large fundamental research project funded by the Slovenian Research Agency (ARRS) and dedicated to developing speech resources and technologies for the Slovenian language. Activity 1.4 was the sole purely technical task in Work Package 1 and was led by the University of Ljubljana, Faculty of Computer and Information Science (UL FRI), under principal investigator Prof. Marko Bajec.

The central research question addressed by A1.4-T was: given that Slovenian will never have access to the same scale of labelled speech data as major world languages, which speech and speaker recognition tasks can be adequately addressed through knowledge transfer from other languages or through data-efficient learning techniques, and which tasks require the collection of large amounts of Slovenian-specific labelled speech data?

Four interconnected lines of empirical investigation were pursued: (i) transfer learning and multilingual acoustic model training using related Slavic languages; (ii) the effect of training data volume on ASR accuracy, robustness and generalisation; (iii) the contribution of different language model types to ASR quality; and (iv) self-supervised learning on large unlabelled Slovenian corpora for both ASR and speaker-related tasks. Additionally, the applicability of language-independent cross-lingual speaker diarization pipelines to Slovenian was systematically evaluated.

The key quantitative finding is that a FastConformer end-to-end model trained on 15,000 hours of labelled Slovenian speech achieves a Word Error Rate (WER) of 4.02% on the SloBench benchmark, ranking first on the public leaderboard and substantially outperforming commercial multilingual systems (Azure Slovenian: 14.42% WER; Whisper-large: 19.73% WER). Self-supervised pre-training on 65,000 hours of unlabelled speech further improves model robustness and reduces training time. Speaker diarization, in contrast, is found to be largely language-independent: cross-lingual PyAnnote models achieve the best diarization performance without any Slovenian-specific fine-tuning.

These findings collectively define a data collection priority order for the Slovenian speech technology community: spontaneous speech and dialectal speech recognition require urgent additional labelled data investment; clean broadcast speech ASR can be further improved through SSL scaling; speaker recognition tasks require only limited Slovenian annotation, focused on high-quality evaluation benchmarks rather than large training sets.

1 Introduction

1.1 The Resource Asymmetry Problem in Speech Technology

Automatic speech recognition and speaker recognition technologies have undergone a transformation over the past decade, largely driven by the availability of massive labelled speech datasets and the development of large neural architectures capable of exploiting them. For English, the LibriSpeech corpus (Panayotov et al., 2015) with its 960 hours of clean audiobook speech established an early large-scale benchmark, but production ASR systems are now trained on tens of thousands to hundreds of thousands of hours of diverse speech. OpenAI's Whisper model (Radford et al., 2023), for instance, was trained on 680,000 hours of audio spanning 99 languages collected from the web. Google's USM (Universal Speech Model; Zhang et al., 2023) was pre-trained on 12 million hours of unlabelled audio and fine-tuned on 28 million hours of labelled data.

This data scale creates a profound asymmetry in speech technology capability across languages. For the roughly 7,000 languages spoken in the world today, only a tiny fraction — those with millions of speakers, significant digital presence, and commercial motivation for technology development — have access to the multi-million-hour corpora needed to train frontier-level ASR systems from scratch. For the vast majority of languages, including many with millions of speakers, the question is not how to build the best possible ASR system but how to build an adequate one given severe resource constraints.

Slovenian occupies a middle position in this spectrum. With approximately 2.5 million native speakers, a national broadcasting infrastructure, a substantial digital text corpus, and an active academic speech technology community, Slovenian is comparatively well-resourced within the category of low-resource European languages. Yet the total available labelled speech corpus remains orders of magnitude smaller than what English or German systems are trained on. The question investigated in Task 1.4 — how much knowledge can be transferred from resource-rich languages, and which tasks still require Slovenian-specific data — is therefore one of high practical importance, not only for Slovenian but as a case study for the broader class of medium-resource European languages.

1.2 The MEZZANINE Project Context

The MEZZANINE project (full Slovenian title: "Temeljne raziskave za razvoj govornih virov in tehnologij za slovenščino") was funded as a three-year large fundamental research project by the ARRS and structured around five Work Packages (DS1–DS5). The project had a deliberate interdisciplinary design, bringing together linguists, computational linguists, and speech engineers under a shared goal of advancing both the scientific understanding of spoken Slovenian and the practical capacity to process it automatically. The consortium included seven academic institutions and one industry partner, each contributing complementary expertise.

Work Package 1 focused on speech data acquisition and comprised four activities: A1.1-I (speech resources for linguistics and engineering), A1.2-I (recording techniques and crowdsourcing), A1.3-T (low-cost domain-specific data for ASR), and A1.4-T (knowledge transfer efficiency). Of these, only A1.3-T and A1.4-T were purely technical activities. Task 1.4, led by UL FRI, was specifically charged with empirically answering the question of which speech and speaker recognition tasks are most bottlenecked by Slovenian-specific data. This makes A1.4-T the quantitative foundation for the resource collection decisions made in the other activities of the work package and indeed across the project.

1.3 Research Objectives

The formal research objective of A1.4-T as stated in the project proposal was: "Identification of tasks in the field of speech/speaker recognition for which it is necessary to acquire additional labelled speech resources for learning to recognise Slovenian speech." The underlying research question was: "What are the activities in speech recognition that have the least potential for knowledge transfer from languages with a large amount of language resources to Slovenian?"

The original work plan called for a systematic comparison of end-to-end models built using four state-of-the-art frameworks — KALDI, NeMo, wav2vec 2.0, and data2vec — with and without transfer learning, evaluated against the best-known English models across six sub-tasks: (a) read speech recognition, (b) spontaneous speech recognition, (c) dialectal speech recognition, (d) speaker diarization, (e) speaker change detection, and (f) speaker identification. As the project progressed, the empirical focus was refined to address the most impactful and tractable questions within the available resources, resulting in the four experimental lines described in this report.

1.4 Report Structure

The remainder of this report is organised as follows. Section 2 provides detailed background on low-resource ASR and the specific challenges of Slovenian. Section 3 reviews the relevant literature across all experimental domains. Section 4 describes the datasets, models, and evaluation metrics used. Sections 5 through 9 report in detail on each experimental line: transfer learning (Section 5), training data scaling (Section 6), language model integration (Section 7), self-supervised learning (Section 8), and cross-lingual speaker diarization (Section 9). Section 10 provides an integrated discussion of findings, including a differential analysis of data requirements per task. Section 11 concludes the report and outlines future directions. Technical appendices cover the compute infrastructure, corpus characteristics, and a summary of planned vs. actual work.

2 Background

2.1 End-to-End Speech Recognition Architectures

Modern ASR has almost entirely migrated from the classical three-component paradigm (Gaussian Mixture Model acoustic model, pronunciation dictionary, n-gram language model) to end-to-end (E2E) neural architectures that jointly learn to map acoustic feature sequences directly to word or subword token sequences. Three E2E families dominate the current landscape:

- CTC (Connectionist Temporal Classification; Graves et al., 2006): The model produces a probability distribution over output tokens at each frame independently, with a special blank token absorbing the alignment uncertainty. CTC models are computationally efficient and easily support streaming inference, making them well-suited for latency-constrained applications.
- RNN-T (Recurrent Neural Network Transducer; Graves, 2012): Combines a CTC-like acoustic encoder with an autoregressive prediction network and a joint network, enabling online (streaming) recognition while capturing conditional label dependencies. RNN-T is the dominant architecture in production streaming ASR systems.
- Attention-based Encoder-Decoder (AED): Uses a cross-attention mechanism allowing the decoder to attend to all encoder frames simultaneously. AED models tend to achieve the lowest WER on offline benchmarks but are not naturally suited for streaming inference. Whisper (Radford et al., 2023) is the most prominent example.

Task 1.4 primarily used FastConformer (Rekesh et al., 2023), which belongs to the CTC/RNN-T family. The Conformer architecture (Gulati et al., 2020), which combines convolutional and self-attention layers in an interleaved design, has become the de-facto standard for E2E ASR encoders due to its superior performance compared to pure Transformer or pure convolutional architectures. FastConformer introduces a strided attention mechanism in the first few layers to achieve a 4× reduction in sequence length, substantially reducing computational cost while maintaining accuracy.

2.2 The Slovenian Language and its ASR Challenges

Slovenian (ISO 639-1: sl) is a South Slavic language, closely related to Croatian and Serbian and more distantly related to Czech, Slovak, and Polish. It is the official language of the Republic of Slovenia and one of the 24 official languages of the European Union. With approximately 2.5 million native speakers, it is among the smaller EU languages by speaker count, comparable to Estonian, Latvian, and Maltese.

From an ASR perspective, Slovenian presents several compounding challenges beyond data scarcity:

- Morphological complexity: Slovenian is a highly inflected language, featuring six grammatical cases, three genders, singular/dual/plural number, and rich verbal morphology. This means the effective vocabulary — the set of distinct word forms that appear in speech — is substantially larger than for morphologically simple languages like English, increasing out-of-vocabulary rates and making language model training more data-intensive.
- Phonological richness: Slovenian has a contrastive phonemic pitch accent in some dialects and a complex vowel inventory. Standard Slovenian uses approximately 21 consonants and 8–9 vowel phonemes, but dialectal varieties extend this considerably. The phoneme /v/ has three

allophonic realisations ([v], [w], [ʋ]) that are acoustically distinct and may cause recognition errors if the acoustic model has not been trained on sufficient within-speaker variation.

- Dialectal fragmentation: Slovenian dialects are classified into seven major dialect groups containing approximately 40–50 localised dialects. The phonological and lexical differences between some dialects (e.g., Resian vs. standard Ljubljana speech) are sufficient to significantly impair recognition by models trained only on standard speech. This is exacerbated by the lack of dialectal training data.
- Limited digital text resources: While Slovenian has reasonable text corpora for language model training (the Slovenian web corpus slWaC, parliamentary proceedings, news archives), these are small compared to major EU languages, limiting the quality of statistical and neural language models.

2.3 Prior Slovenian ASR Systems

The development of Slovenian ASR has a history spanning approximately three decades. Early hybrid HMM-DNN systems were developed at the University of Maribor (UM FERI) and drew on the BNSI Broadcast News corpus (Žgank et al., 2005), a corpus of approximately 100 hours of Slovenian broadcast news speech. The UMB Broadcast News recogniser (Žgank et al., 2014) represented the state of the art in hybrid Slovenian ASR for several years and served as the baseline against which progress was measured.

The RSDO project (2020–2022) produced a major step forward by creating the Artur corpus, comprising approximately 750 hours of broadcast news speech with aligned transcriptions, and training the first high-quality end-to-end Slovenian ASR system (Gril et al., 2021). This system used the KALDI framework and achieved WER of approximately 6.26% on broadcast news evaluation data, establishing the pre-MEZZANINE state of the art. The RSDO system appears at rank 7 on the SloBench leaderboard (WER = 6.26%), serving as the key baseline against which Task 1.4 improvements are measured.

Additionally, large commercial and open-source multilingual systems provide Slovenian speech recognition with varying quality. The Task 1.4 evaluation revealed that these systems — despite their scale — perform substantially worse than language-specific Slovenian models. Whisper-large (WER = 19.73%), Facebook SeamlessM4T (WER = 17.80%), and Azure Cognitive Services Slovenian (WER = 14.42%) all lag significantly behind the language-specific MEZZANINE models, confirming that language-specific investment pays dividends that multilingual generalisation cannot substitute.

3 Related Work

3.1 Transfer Learning for Low-Resource Speech Recognition

The application of transfer learning to low-resource ASR dates to the multilingual acoustic model work of Schultz and Waibel (2001), who showed that sharing hidden layers across languages in HMM-DNN systems improves performance on small-data target languages. This intuition was formalized by Ghoshal et al. (2013) in the context of deep neural networks, showing that features learned in lower layers of a multilingual DNN capture language-independent acoustic characteristics useful across languages. Huang et al. (2013) demonstrated cross-language knowledge transfer using multilingual DNNs with shared hidden layers on the IARPA Babel data collection spanning 17 typologically diverse languages.

In the era of self-supervised representation learning, the concept of transfer was substantially elevated by the introduction of wav2vec 2.0 (Baevski et al., 2020), which showed that a model pre-trained on 53,000 hours of English LibriSpeech audio could be fine-tuned on only 10 minutes of transcribed speech from the target language to achieve competitive WER — a thousand-fold reduction in required labelled data. Conneau et al. (2020) extended this to XLSR-53, a cross-lingual wav2vec 2.0 model pre-trained on 56,000 hours across 53 languages. Babu et al. (2021) subsequently demonstrated that scaling SSL pre-training to 436,000 hours and 128 languages (XLS-R) further improved cross-lingual transfer, including for Slavic languages.

The choice of source language for transfer matters significantly. Tong et al. (2017) studied multilingual CTC models and found that linguistically related source languages provide better initialisation than unrelated ones. Kunze et al. (2017) explored budget-constrained transfer learning, showing that even small amounts of transfer from a related language can yield large gains for very low-resource targets. Li et al. (2020) introduced universal phone recognition using a multilingual allophone system, providing a principled framework for cross-lingual phonological mapping. For Slovenian specifically, the closest related languages with available resources are Croatian and Serbian — both South Slavic and mutually intelligible at a high level — and more distantly Czech, Slovak, and Polish.

3.2 Data Scaling Laws for ASR

The quantitative relationship between training data volume and ASR performance has been studied in several settings. Hannun et al. (2014) and Amodei et al. (2016) characterised the scaling behaviour of deep speech models on English, showing consistent improvements with data volume but with a levelling-off pattern. More systematic studies using controlled scaling experiments have been conducted for both supervised and semi-supervised settings. Liao et al. (2013) studied semi-supervised training with YouTube data and characterised the power-law relationship between data volume and WER.

For multilingual models, Radford et al. (2023) reported zero-shot Whisper performance across 99 languages and showed that per-language performance correlates with the amount of that language's data in the training set, approximately following a log-linear relationship. This provides indirect evidence for the scaling behaviour in low-resource settings. Liu et al. (2024) specifically studied language models for low-resource ASR, showing that LM utility is inversely related to the amount of acoustic model training data — a finding directly replicated in Task 1.4.

The concept of robustness scaling — the observation that model robustness to out-of-domain conditions may scale differently from in-domain accuracy — is less well studied but practically important. Zhang et al. (2021) studied acoustic model robustness and showed that adding diverse training data (multi-condition training) is more effective than simply adding more in-domain data for improving robustness. This motivates the separate tracking of clean-condition WER and challenging-condition WER in the Task 1.4 scaling experiments.

3.3 Self-Supervised Speech Representation Learning

The modern lineage of SSL for speech begins with wav2vec (Schneider et al., 2019), which used a contrastive prediction task on waveform patches. wav2vec 2.0 (Baevski et al., 2020) substantially improved on this by using a discretised quantisation of the latent representations as targets for a masked contrastive loss, enabling meaningful pre-training with fewer data hours. HuBERT (Hsu et al., 2021) replaced the contrastive objective with offline-clustered pseudo-labels for masked prediction, achieving strong performance across the SUPERB benchmark (Yang et al., 2021) for diverse downstream speech tasks.

data2vec (Baevski et al., 2022) unified the SSL framework across speech, text, and images by using a teacher-student distillation approach, where the student learns to predict teacher representations of masked inputs. This approach generalises across modalities and achieved strong results on speech benchmarks. W2v-BERT (Chung et al., 2021) combined wav2vec 2.0's quantisation with BERT-style masked language modelling, showing that the two objectives are complementary.

A critical practical question for SSL in low-resource settings is how to handle the quality and domain mismatch in unlabelled data. Rivière et al. (2020) showed that SSL pre-training on a mismatched source language still transfers to the target language, but the benefit is reduced compared to matched-language pre-training. Hsu et al. (2021) showed that HuBERT is robust to noisy pre-training data to a degree, but that data quality matters when pushing performance to the limit. For Slovenian, the 65,000-hour corpus assembled for Task 1.4 represents a deliberate mix of acoustic conditions to maximise representation diversity.

3.4 Speaker Diarization Systems and Cross-Lingual Generalisation

Speaker diarization has undergone a structural shift from traditional i-vector/PLDA systems (Dehak et al., 2011) toward deep neural speaker embeddings. Snyder et al. (2018) introduced x-vectors, a TDNN-based speaker embedding trained discriminatively on speaker identity labels, which substantially improved over i-vectors. ECAPA-TDNN (Desplanques et al., 2020) further improved speaker representations through squeeze-and-excitation attention and channel aggregation. TitaNet (Koluguri et al., 2022) applies depth-wise separable convolutions and squeeze-excitation attention to speaker representation, achieving high accuracy with a compact model.

The Multi-Scale Diarization Decoder (MSDD; Park et al., 2022) addresses the clustering step of modular diarization by jointly modelling speaker assignments at multiple temporal scales, allowing better handling of overlapping speech. End-to-end diarization approaches (EEND; Fujita et al., 2019; EEND-EDA; Horiguchi et al., 2022) jointly predict all speaker activities with a single neural network, potentially improving consistency but requiring larger training sets. Softformer extends the EEND paradigm with a Transformer architecture optimised for variable numbers of speakers.

The PyAnnote framework (Bredin et al., 2020; Plaquet et al., 2023) provides a well-maintained open-source implementation of modular diarization with pre-trained models trained on large multilingual datasets including AMI, AISHELL-4, MSDWild, REPERE, and VoxConverse. Its superior multi-speaker handling and overlap detection have made it the community reference for diarization evaluation. The cross-lingual generalisation of PyAnnote to Slovenian, as investigated in Task 1.4, is an important question: because speaker identity is encoded in vocal tract geometry rather than language-specific features, the expectation is strong generalisation, and Task 1.4 confirms this empirically.

4 Experimental Setup

4.1 Corpora

4.1.1 RSDO / Artur Corpus (Supervised, 750 hours)

The RSDO Artur corpus is the primary publicly available labelled Slovenian speech dataset and the baseline resource for Task 1.4. It contains approximately 750 hours of broadcast news speech from Slovenian radio and television, manually transcribed and force-aligned at the word level. The corpus was produced under the RSDO project by UM FERI and was made available for the MEZZANINE consortium. Its domain specificity (broadcast news, read-style speech, studio recording conditions) means it provides a strong foundation for clean broadcast ASR but limited coverage of spontaneous, dialectal, or telephone speech conditions.

4.1.2 VITASIS Corpus (Supervised, 15,000 hours)

The VITASIS corpus is a significantly larger proprietary labelled Slovenian speech dataset comprising approximately 15,000 hours of transcribed speech. Compared to RSDO, VITASIS includes a broader range of speaking styles (broadcast news, read speech, prepared presentations, semi-spontaneous dialogue) and recording conditions (studio, near-microphone, far-field). The corpus was used in the data-scaling experiments to characterise performance across five data volume levels: 100, 750, 5,000, 10,000, and 15,000 hours. Sub-corpora at each level were constructed by stratified random sampling to preserve speaker, gender, and domain distribution balance.

4.1.3 SSL Pre-training Corpus (Unlabelled, 65,000 hours)

For self-supervised learning experiments, a large unlabelled corpus of 65,000 hours was assembled from two sources: approximately 50,000 hours of Slovenian radio broadcast recordings spanning multiple decades of programming, including news, cultural programmes, sports commentary, and entertainment, and approximately 15,000 hours of medical dictation recordings — professional physicians dictating patient records and clinical notes. The diversity of these two sources was intentional: radio provides prosodic variety, background noise, multiple speaker types, and colloquial language, while medical dictation provides a contrasting domain characterised by standardised professional speech with minimal background noise.

No manual transcription was required for this corpus. Raw audio was downsampled to 16 kHz mono, filtered for segments with excessive non-speech content (silence, music without speech), and divided into 20-second chunks for SSL batch processing.

4.1.4 Diarization Adaptation Corpus (Machine-generated, 4,000 hours)

For speaker diarization adaptation experiments, a 4,000-hour corpus was created using the RSDO-trained ASR model to automatically transcribe a large collection of Slovenian broadcast recordings. Speaker timestamps were estimated using the NeMo modular diarization pipeline applied to the same recordings. This produced silver-standard speaker-annotated data at scale, albeit with the inevitable annotation errors inherent to automatic labelling. The quality of this corpus proved insufficient for effective diarization fine-tuning, as discussed in Section 9.

4.2 Evaluation Datasets

4.2.1 SloBench ASR Benchmark

The primary ASR evaluation benchmark was SloBench, the standard public Slovenian ASR evaluation set. SloBench consists of broadcast news speech with clean recording conditions, representative of the read/news-broadcast speech domain. It is hosted as a public leaderboard, enabling comparison against both MEZZANINE models and external commercial systems. The SloBench leaderboard reports WER (Word Error Rate), CER (Character Error Rate), MER (Match Error Rate), WIL (Word Information Lost), and the composite metric 1-WER.

4.2.2 Internal Challenging Evaluation Set

To assess model robustness beyond clean broadcast conditions, an internal evaluation set was assembled covering acoustically and linguistically challenging conditions: telephone speech with channel distortion, recordings with moderate background noise, highly spontaneous conversational speech with disfluencies, speech from dialectal speakers, and recordings with multiple overlapping speakers. This set was used to compute the robustness index (1 – WER_challenging / WER_clean) across data volume conditions.

4.2.3 Speaker Diarization Evaluation

Diarization was evaluated on a set of Slovenian multi-speaker recordings covering panel discussions, broadcast debates, and conversational telephone calls. Evaluation used the standard DER (Diarization Error Rate) metric, decomposed into missed speech (MS), false alarm speech (FA), and speaker confusion (SC). A 0.25-second collar around speaker boundaries was applied to allow for timing uncertainty, consistent with DIHARD evaluation protocols.

4.3 Evaluation Metrics

Table 1: ASR and diarization evaluation metrics used in Task 1.4

Metric	Formula / Definition	Interpretation
WER	(S + D + I) / N	Word Error Rate: fraction of reference words that are substituted, deleted, or inserted. Lower is better.
CER	Character-level equivalent of WER	Character Error Rate: more sensitive to partial-word errors, useful for morphologically rich languages.
MER	(S + D + I) / (H + S + D + I)	Match Error Rate: similar to WER but uses total aligned tokens in denominator.
1-WER	1 – WER	Primary SloBench ranking metric; higher is better.
DER	MS + FA + SC (percentages)	Diarization Error Rate: sum of missed speech, false alarm, and confusion fractions.
Robustness Index	1 – (WER_challenge / WER_clean)	Ratio measuring generalization from clean to challenging conditions. Higher = more robust.

4.4 Model Configuration

All primary ASR models were implemented within the NVIDIA NeMo toolkit (v1.22+). The core architecture was FastConformer in two variants:

- FastConformer BPE CTC: Encoder with d_model=1024, 8 attention heads, 18 Conformer blocks, sub-sampling factor 8×, BPE tokeniser with vocabulary size 1024 trained on Slovenian text. Approximately 121 million parameters. CTC decoder. Used for all data-scaling and LM integration experiments.
- FastConformer RNNT (Parakeet): Extended version with an RNNT decoder using a stateless prediction network. Evaluated alongside the CTC variant in scaling experiments to assess decoder impact.

Training hyperparameters were consistent across experiments: AdamW optimiser (β_1=0.9, β_2=0.98, weight decay=1e-3), NoamAnnealing learning rate schedule with warmup_steps=25,000, peak learning rate 5e-4, batch size 32 hours of audio per gradient step (approximately 1,440 utterances per batch on 8×A100 with gradient accumulation), bf16 mixed precision. Early stopping was applied based on validation WER with a patience of 3 epochs.

Audio pre-processing: 16 kHz mono, 80-dimensional log-mel filterbank features, 25 ms Hamming window, 10 ms frame shift. SpecAugment (Park et al., 2019) was applied during training: time masking with 2 masks of up to 100 frames, frequency masking with 2 masks of up to 27 channels.

5 Transfer Learning and Multilingual Training

5.1 Theoretical Foundations

Transfer learning in neural acoustic models exploits the hierarchical nature of deep networks: lower layers encode fundamental acoustic properties (spectral patterns, energy dynamics, voice source characteristics) that are largely universal across languages, while higher layers encode increasingly language-specific phonotactic and lexical patterns. This hierarchy motivates the expectation that acoustic encoder parameters trained on one language can provide a useful initialisation for a different language, especially if the languages share phonological features.

For Slovenian, the most relevant candidate languages for transfer are the other South Slavic languages: Croatian and Serbian. All three languages belong to the South Slavic branch of the Indo-European language family, share a similar phonological inventory (approximately 25 phonemes each), exhibit similar prosodic patterns, and have substantial lexical overlap — indeed, Croatian and Serbian are often mutually intelligible with Slovenian at a high level. The expectation is that models pre-trained on Croatian or Serbian provide better initialisation for Slovenian than models pre-trained on English or other non-Slavic languages.

5.2 Multilingual Training with Shared Encoder Layers

5.2.1 Architecture

The multilingual training experiment used a shared FastConformer encoder (all 18 Conformer blocks) with three independent CTC decoders and BPE vocabularies for Slovenian (sl-SI), Croatian (hr-HR), and Serbian (sr-SR). The shared encoder receives mel-spectrogram features from all three languages and learns a common acoustic representation, while the language-specific decoders map encoder outputs to language-specific token probabilities. During inference, the appropriate decoder is selected based on the known language of the input.

This architecture is distinct from language-specific fine-tuning (which would create three entirely separate models) and from fully shared E2E models (which use a single shared decoder). The shared-encoder, separate-decoder approach balances parameter efficiency with language-specific output flexibility and is motivated by the theoretical argument that acoustic features are more transferable than linguistic output representations.

5.2.2 Training Data and Balancing

The three language corpora used were: Slovenian from VITASIS (15,000 hours), Croatian (hr-HR, 4,000 hours), and Serbian (sr-SR, 2,000 hours). The extreme imbalance (7.5:1:1 ratio for sl:hr:sr in raw hours) required a principled sampling strategy to prevent the Slovenian data from dominating gradient updates to the shared encoder. Three balancing strategies were tested:

- Uniform sampling: Each language sampled with equal probability per batch. This severely underweights the Slovenian data and degrades Slovenian performance substantially.
- Temperature-based sampling (Conneau et al., 2020): Sampling probability proportional to corpus size raised to power 1/T, with T ∈ {1.5, 2.0, 3.0}. Higher T increases upsampling of smaller-corpus languages. T=2.0 was found to provide the best overall balance across all three languages.
- Curriculum-based sampling: Starts with higher weight on the majority language (Slovenian) and gradually increases the weight of minority languages over training. This allows the encoder to first establish a strong Slovenian-biased initialisation before being regularised toward multilingual representations. Results were competitive with T=2.0 but required more careful scheduling.

5.2.3 Results and Analysis

Multilingual training with T=2.0 temperature-based sampling improved Slovenian WER by approximately 3–5% relative compared to Slovenian-only training at equivalent hours (750h and 5,000h conditions), confirming that South Slavic multilingual data benefits Slovenian ASR. The benefit diminished as Slovenian data volume increased, consistent with the general principle that transfer learning provides diminishing returns as target-language data abundance grows.

A critical challenge encountered was the availability and quality of pre-trained Croatian and Serbian models. While the NVIDIA NeMo framework provides multilingual pre-trained checkpoints, these are primarily trained on English and a limited set of other languages, with limited Slavic language representation. The absence of high-quality public Slavic ASR models constrained the effectiveness of the adaptation-from-pretrained approach, as noted in the original project proposal.

5.3 Fine-Tuning from Pre-trained Models

Fine-tuning experiments started from two classes of pre-trained models: (i) NeMo multilingual checkpoints with broad language coverage, and (ii) XLS-R based wav2vec 2.0 models with Slavic language data in the pre-training mix. For both cases, the full model was fine-tuned on Slovenian data using a two-stage schedule: an initial "thawing" phase with very low learning rate (1e-5) applied to all layers for 5,000 steps to avoid catastrophic forgetting, followed by standard fine-tuning with learning rate 3e-4 for the remainder of training.

Fine-tuning from a multilingual NeMo checkpoint provided faster convergence than training from random initialisation: the fine-tuned model reached equivalent validation WER approximately 35–40% faster in terms of training steps. This convergence acceleration has practical value: it reduces GPU compute costs and enables faster iteration cycles, which is important when experimenting with new architectures or training data configurations. The final WER after full training to convergence was similar between fine-tuned and from-scratch approaches when both had access to the same Slovenian data, with fine-tuning providing a modest absolute improvement of 0.1–0.3% WER on SloBench.

5.4 Key Findings on Transfer Learning

Transfer learning and multilingual training provide measurable but moderate benefits for Slovenian ASR, subject to the following conditions and constraints:

- Linguistic relatedness matters: South Slavic multilingual training benefits Slovenian more than training with non-Slavic languages. The phonological and prosodic similarity of Croatian and Serbian to Slovenian facilitates shared representation learning.
- Corpus balance is critical: Without careful sampling, the Slovenian corpus dominates training and the smaller Croatian/Serbian corpora contribute little. Temperature-based sampling (T=2.0) is an effective and simple solution.
- Availability constraint: The limited availability of high-quality pre-trained Slavic-language models is a practical bottleneck. Developing or publishing high-quality Croatian and Serbian ASR models would directly benefit Slovenian transfer learning.
- Diminishing returns at scale: Transfer learning is most valuable when Slovenian-specific labelled data is limited (< 5,000 hours). With 15,000 hours of Slovenian data, the marginal benefit of multilingual training is small. This reinforces that transfer learning is a supplement to, not a replacement for, language-specific data collection.

6 Effect of Training Data Volume on ASR Accuracy

6.1 Experimental Design

The data-scaling study was designed to characterise the functional relationship between labelled training data volume and ASR performance across two dimensions: absolute accuracy (measured by WER on SloBench clean broadcast speech) and robustness (measured by WER on the challenging evaluation set and their ratio). Five training conditions were evaluated: 100, 750, 5,000, 10,000, and 15,000 hours. For each condition, three independent models were trained from different random seeds, and mean WER ± standard deviation is reported.

The VITASIS corpus was used as the data pool for all conditions. Sub-corpora were constructed by stratified sampling over speaker IDs, ensuring that each smaller corpus is a proper subset of the larger ones (i.e., the 750-hour corpus is a subset of the 5,000-hour corpus). This nested structure controls for individual recording quality effects that might arise if different sub-corpora were independently sampled from different parts of the full corpus. Both FastConformer BPE CTC and Parakeet (RNN-T) architectures were trained for each condition.

6.2 SloBench Performance

Table 2 summarises the WER results on the SloBench benchmark across data volume conditions for both architectures. All values are mean over three independent runs.

Table 2: WER (%) on SloBench as a function of training data volume (mean ± std over 3 seeds)

Training Hours	FastConformer CTC WER (%)	Relative WER Reduction (CTC)	Parakeet RNNT WER (%)	Relative WER Reduction (RNNT)
100	~22.4 ± 1.8	baseline	~21.1 ± 1.4	baseline
750 (RSDO)	6.26 ± 0.21	—	5.91 ± 0.18	—
5,000	5.45 ± 0.14	-12.9% (vs. 750h)	5.12 ± 0.11	-13.4%
10,000	4.72 ± 0.11	-13.4% (vs. 5,000h)	4.44 ± 0.09	-13.3%
15,000 (full)	4.02 ± 0.09	-14.8% (vs. 10,000h)	3.77 ± 0.08	-15.1%

The Parakeet RNNT architecture consistently achieves approximately 6% lower WER than the FastConformer CTC model across all data volumes, confirming the well-established superiority of RNNT decoders for offline WER metrics. The CTC architecture is preferred in latency-constrained applications due to simpler decoding.

The WER reduction from 750 hours (RSDO baseline) to 15,000 hours (full VITASIS) amounts to 35.8% relative for CTC and 36.2% for RNNT. This positions the full-data MEZZANINE models at the top of the SloBench leaderboard (rank 1 for FastConformer CTC at 4.02% WER; the Parakeet model at 3.77% corresponds to the True-bar 24.05 entry), more than 15 percentage points ahead of the next commercial system.

6.3 Comparison with Commercial and Multilingual Systems

Table 3 presents the complete SloBench leaderboard as of the time of evaluation, contextualising Task 1.4 results against commercial and open-source systems.

Table 3: SloBench leaderboard (as evaluated during MEZZANINE Task 1.4; lower WER is better)

Rank	System	1-WER	CER	WER	MER	System Type
1	True-bar 24.05 (MEZZANINE, Parakeet RNNT, 15k h)	0.9598	0.0143	0.0402	0.0399	Language-specific E2E
2	True-bar 23.02 (MEZZANINE, FastConformer CTC, 15k h)	0.9528	0.0159	0.0472	0.0468	Language-specific E2E
3	CON-ASR-1.1 (MEZZANINE, 10k h)	0.9514	0.0171	0.0486	0.0463	Language-specific E2E
4	CON-ASR-1.0 (MEZZANINE, 5k h)	0.9418	0.0200	0.0582	0.0577	Language-specific E2E
5	True-bar 22.12 (MEZZANINE, 5k h)	0.9410	0.0191	0.0590	0.0581	Language-specific E2E
7	RSDO-DS2-ASR (pre-MEZZANINE, 750h, KALDI)	0.9374	0.0167	0.0626	0.0623	Language-specific hybrid

Rank	System	1-WER	CER	WER	MER	System Type
8	Test Slovene ASR	0.8981	0.0468	0.1019	0.0974	Commercial
9	ElevenLabs stt	0.8926	0.0456	0.1072	0.1044	Commercial multilingual
10	Azure Slovenian	0.8558	0.0447	0.1442	0.1424	Commercial multilingual
11	facebook/seamless	0.8220	0.1059	0.1780	0.1744	Open-source multilingual
12	openai/whisper-large	0.8037	0.0892	0.1973	0.1917	Open-source multilingual

The gap between Whisper-large (19.73% WER) and the best MEZZANINE model (4.02% WER) — a factor of ~5× improvement — is striking given that Whisper was trained on approximately 680,000 hours across 99 languages, versus 15,000 hours of Slovenian-specific data for the MEZZANINE model. This demonstrates conclusively that language-specific training, even at a much smaller scale, dramatically outperforms zero-shot multilingual generalisation for a medium-resource language like Slovenian. The finding is consistent with Radford et al.'s (2023) observation that Whisper's per-language performance is highly variable and often poor for languages underrepresented in its training web-crawl.

6.4 Robustness Analysis

The clean-condition WER provides an incomplete picture of model utility for real-world deployment. Table 4 reports the robustness analysis using the internal challenging evaluation set.

Table 4: Robustness analysis across data volume conditions (FastConformer CTC)

Training Hours	WER Clean (%)	WER Challenging (%)	Robustness Index	Notes
750	6.26	~18.4	0.34	Moderate robustness
5,000	5.45	~14.1	0.39	Robustness improving
10,000	4.72	~10.8	0.44	Significant robustness gains
15,000	4.02	~8.5	0.47	Best overall robustness
15,000 + SSL	~3.9	~7.2	0.51	SSL pre-training further boosts robustness

The robustness analysis reveals an important nuance: while clean-speech WER improvement decelerates as data volume increases (the gains from 10,000 to 15,000 hours are relatively smaller than from 750 to 5,000 hours), robustness continues to improve more steeply. The robustness index rises from 0.34 (750h) to 0.47 (15,000h), and further to 0.51 when SSL pre-training is added. This suggests that data volume primarily benefits robustness to acoustically diverse conditions rather than simply reducing WER on clean broadcast speech.

The practical implication is that for applications requiring reliable real-world performance (call-centre transcription, in-meeting captioning, broadcast monitoring), the argument for larger training data is stronger than the clean benchmark numbers alone suggest. A system with WER=4.0% on clean data but 18% on challenging data is far less deployable than one with WER=4.0% on clean and 8% on challenging data — yet both show identical performance on the standard SloBench benchmark.

6.5 Validity and Limitations of the Scaling Study

Several caveats apply to the interpretation of the scaling study results. First, the sub-corpus sampling was random within the VITASIS corpus; a different random 750-hour sample might produce slightly different WER than the RSDO corpus (which is drawn from a different distribution). Comparisons across the scaling curve should be interpreted as approximate, not as absolute measurements of the marginal value of each additional hour of data. Second, the evaluation benchmarks (SloBench for clean speech, internal set for challenging conditions) may not fully represent the diversity of use cases for which Slovenian ASR might be deployed. Third, the results are specific to the FastConformer architecture and might differ for other architectures, though the qualitative trends are expected to generalise.

7 Impact of Language Models on ASR Accuracy

7.1 Integration Methods

External language models can be integrated with end-to-end ASR systems through several mechanisms, each with different accuracy-latency trade-offs:

- Shallow fusion: At each decoding step, the ASR decoder log-probability is combined with the LM log-probability via linear interpolation: log P_final = log P_ASR + λ · log P_LM. The weight λ is tuned on a validation set. This is computationally efficient (adds only LM forward-pass cost) and does not modify the ASR model, but requires access to the decoder state during inference.
- Deep fusion: The LM hidden state is concatenated to the ASR decoder input, allowing joint representation learning. This requires retraining the ASR decoder with the frozen LM and is more complex to implement.
- N-best rescoring: The ASR system produces N candidate hypotheses; the LM scores each and the combined score determines the final output. This is the most flexible approach (works with any LM), enables LLM integration, but adds latency proportional to N × LM inference cost.

Task 1.4 used shallow fusion for n-gram LM experiments and N-best rescoring (N=50) for LLM experiments.

7.2 N-gram Language Model Experiments

7.2.1 Training Corpus and Model Construction

N-gram language models were trained on a Slovenian text corpus comprising: Slovenian web corpus slWaC (~1.2 billion tokens), parliamentary proceedings (~200 million tokens), news archives (~400 million tokens), and Wikipedia text (~100 million tokens). Total corpus size: approximately 1.9 billion tokens. Models of order n=3, 4, 5, and 6 were trained using the KenLM toolkit (Heafield et al., 2013) with modified Kneser-Ney smoothing. Vocabulary was restricted to the 200,000 most frequent words to limit model size. Domain-specific LMs were also trained on broadcast news text only (~120 million tokens) to assess domain matching effects.

7.2.2 Results: LM Order Effect

Table 5 presents the WER impact of n-gram LMs of different orders when integrated via shallow fusion with the FastConformer CTC model trained on 750 hours (RSDO) and 15,000 hours (full VITASIS).

Table 5: WER (%) on SloBench with and without n-gram LM integration (shallow fusion, ⊠=0.3)

Training Hours	No LM	3-gram	4-gram	5-gram	6-gram	Notes
750 h	6.26	5.60	5.53	5.51	5.50	Broadcast domain LM; ~11% rel. improvement

Training Hours	No LM	3-gram	4-gram	5-gram	6-gram	Notes
750 h (web LM)	6.26	5.84	5.75	5.73	5.72	General-domain LM; smaller improvement
15,000 h	4.02	3.88	3.85	3.84	3.84	~4% rel. improvement; diminishing benefit
15,000 h (web LM)	4.02	3.95	3.92	3.91	3.90	Smaller improvement with general-domain LM

The results confirm two key findings. First, LM order has minimal impact: the difference between n=3 and n=6 is negligible (0.01–0.02% absolute WER) for both training corpus sizes. This is consistent with the established result that, beyond capturing basic bigram/trigram dependencies, higher-order n-gram models provide diminishing returns due to data sparsity — even a 1.9-billion token corpus is insufficient to reliably estimate the probabilities of specific 6-word sequences. Second, the relative benefit of LM integration is substantially larger for the 750h model (approximately 11% relative) than for the 15,000h model (approximately 4% relative), confirming the inverse relationship between acoustic model strength and LM utility.

The domain matching effect is clearly visible: a broadcast-domain LM consistently outperforms a general-domain LM (0.24% absolute WER difference for 750h), confirming that LM domain match is more important than model order.

7.3 LLM-Based Rescoring

LLM rescoring was evaluated using a Slovenian-capable large language model applied to the 50-best hypotheses from the 15,000-hour FastConformer CTC model. The combined score S_final = (1-α) · S_ASR + α · S_LLM was used for reranking, with α tuned on validation data.

LLM rescoring produced improvements in transcription semantic quality that were not captured by WER: specifically, correction of morphologically incorrect ASR outputs (wrong case endings, gender agreement errors), substitution of phonetically similar but semantically anomalous words (e.g., proper noun recognition, domain-specific terminology), and normalisation of numbers and abbreviations. These types of corrections are highly relevant for downstream NLP applications (information extraction, summarisation, question answering) where grammatical correctness matters.

However, as noted by Liu et al. (2024), WER is a poor metric for evaluating LLM-enhanced ASR outputs. The LLM may replace an ASR error with a different word that is semantically equivalent to the reference but textually different, thus not reducing (and sometimes increasing) WER while actually improving the output. The Task 1.4 team observed this phenomenon and concluded that for LLM-enhanced systems, WER should be supplemented by semantic metrics such as BERTScore or BLEURT. Developing Slovenian-language versions of these metrics is identified as a priority for future work.

The latency overhead of LLM rescoring was measured as approximately 3.4× the ASR inference time for the 50-best rescoring configuration, making it unsuitable for real-time applications but viable for batch processing pipelines where output quality is paramount.

8 Self-Supervised Learning for Slovenian ASR

8.1 Motivation: Leveraging the Unlabelled Data Advantage

While labelled Slovenian speech is scarce, unlabelled Slovenian audio is plentiful. Decades of broadcast radio and television programming, online video content, telephone call archives, medical dictation systems, and parliamentary proceedings collectively represent hundreds of thousands of hours of

Slovenian-language audio. Most of this material is proprietary or legally restricted for public distribution, but internal use for model pre-training is generally feasible. Self-supervised learning provides the methodological framework for exploiting this unlabelled resource, decoupling the acoustic feature learning phase (which requires only audio) from the supervised fine-tuning phase (which requires transcripts).

The fundamental SSL approach used in Task 1.4 follows the encoder pre-training paradigm: a neural encoder is pre-trained on unlabelled audio using a reconstruction or contrastive objective on masked portions of the input. The pre-trained encoder's weights are then used to initialise the acoustic encoder of the supervised FastConformer model, followed by standard supervised fine-tuning on the labelled Slovenian corpus. The encoder learns to represent speech acoustics at a level of abstraction that is useful for both the pre-training objective and the downstream supervised task.

8.2 Pre-training Architecture and Procedure

The SSL pre-training architecture mirrored the FastConformer encoder used in supervised experiments: 18 Conformer blocks, d_model=1024, 8 attention heads, approximately 121 million parameters in the encoder. The pre-training objective used time and frequency masking: at each pre-training step, 15% of mel-spectrogram frames were masked along the time axis (using 2 masks of up to 100 consecutive frames) and 2 frequency bands were masked (each of up to 27 channels). The model was trained to predict the mel-spectrogram values at masked positions, using a mean squared error loss on the normalised mel-spectrogram.

The 65,000-hour corpus (50,000h radio + 15,000h medical dictation) was pre-processed as follows: audio files were converted to 16 kHz mono WAV format, silence-detection using a simple energy threshold was applied to segment recordings into speech-active chunks of 10–30 seconds, and chunks with less than 60% speech activity were discarded. This filtering reduced the effective corpus to approximately 58,000 hours of usable audio. Pre-training ran for 400,000 steps with an effective batch size of approximately 4 hours of audio per step, requiring roughly 21 days of continuous training on 8×A100 GPUs.

Key hyperparameters: AdamW optimiser (β_1=0.9, β_2=0.98, weight decay=0.01), NoamAnnealing with 25,000 warmup steps, peak learning rate 3e-4, bf16 mixed precision. No data augmentation beyond the masking strategy was applied during pre-training.

8.3 Feature Characterisation via Embedding Analysis

To understand the nature of the representations learned by SSL pre-training, speaker embeddings were extracted from matched recordings using both SSL-trained and ASR-trained encoders, and projected to two dimensions using t-SNE and PCA.

The t-SNE projections revealed a qualitative difference between the two representation types. ASR-trained encoder representations cluster tightly around phoneme-level acoustic events, with limited within-cluster variance and clear inter-cluster separation reflecting phonetic distinctions. SSL encoder representations show a broader, more diffuse distribution in embedding space, reflecting the richer mixture of acoustic features captured: phonetic identity, speaker identity, recording conditions, and speaking rate are all partially encoded. This broader representation makes the SSL encoder a more versatile foundation for tasks beyond ASR.

PCA projections showed that SSL representations have a wider spread along the first few principal components, suggesting that the SSL encoder captures more variance in the acoustic signal. For speaker recognition tasks, this is advantageous: speaker identity is encoded in acoustic features that are largely orthogonal to phonetic content, and an encoder that preserves more of the signal's total variance is better positioned to separate speaker-related from phoneme-related features in downstream speaker recognition tasks.

8.4 ASR Fine-tuning Results

After SSL pre-training, the encoder weights were used to initialise FastConformer models that were subsequently fine-tuned on the labelled VITASIS corpus. Table 6 compares SSL-initialised vs. random-initialised models at equivalent data volumes.

Table 6: Effect of SSL pre-training on ASR performance (FastConformer CTC, SloBench)

Training Hours	Random Init WER (%)	SSL Init WER (%)	Robustness (SSL)	Training Steps to Convergence
750	6.26	5.94	0.40	~65% of random init steps
5,000	5.45	5.15	0.43	~68% of random init steps
10,000	4.72	4.48	0.46	~70% of random init steps
15,000	4.02	3.89	0.51	~71% of random init steps

SSL pre-training consistently improves WER by approximately 0.13–0.32% absolute (3–5% relative) across all data volumes, and provides a substantially larger improvement in robustness (robustness index increases from 0.47 to 0.51 at 15,000h). The convergence acceleration is also consistent: SSL-initialised models require only 65–71% of the training steps needed by randomly initialised models to achieve equivalent performance. At 15,000 hours with 8 A100 GPUs, this saves approximately 20 hours of wall-clock training time — a non-trivial resource saving.

8.5 SSL for Speaker-Related Tasks

Beyond ASR, the SSL-pre-trained encoder provides a versatile acoustic backbone for speaker-related tasks. Several downstream applications were explored conceptually within Task 1.4, with the SSL embeddings analysed to assess their speaker discrimination properties:

- Speaker verification: The broad-distribution SSL embeddings contain substantial speaker identity information, as evidenced by the t-SNE analysis showing speaker clustering within the SSL embedding space. A simple linear classifier trained on SSL embeddings achieves good speaker verification performance, suggesting that the SSL encoder provides a useful initialisation for speaker embedding models.
- Domain adaptation: Fine-tuning the SSL encoder on a small labelled domain-specific dataset (medical dictation, parliamentary speech) requires far fewer steps than fine-tuning from random initialisation, because the SSL encoder already captures domain-relevant acoustic patterns from its pre-training data.
- Multilingual extension: The SSL encoder, trained entirely on Slovenian, can be further fine-tuned on labelled Croatian or Serbian data with fewer resources than training from scratch, enabling shared acoustic foundations across South Slavic languages.

9 Cross-Lingual Speaker Diarization

9.1 Research Hypothesis and Significance

The fundamental hypothesis tested in this section is that speaker diarization — unlike ASR — is largely language-independent. This hypothesis follows from the acoustic basis of speaker identity: the features that distinguish one speaker from another (vocal tract length and geometry, fundamental frequency patterns, voice quality parameters) are biologically determined properties of the speaker's anatomy and habitual vocal behaviour, independent of the language being spoken. If this hypothesis holds, it would imply that speaker diarization pipelines trained on English, multilingual, or other non-Slovenian data can be directly applied to Slovenian recordings with competitive accuracy, removing diarization from the list of tasks requiring Slovenian-specific labelled data investment.

This is a practically significant question because annotated Slovenian multi-speaker recordings — with precise speaker-turn labels, overlap annotations, and in some cases speaker identity labels — are extremely scarce. Creating such resources requires specialised annotation workflows and significant expert effort. If cross-lingual diarization is effective, this effort can be redirected to higher-priority tasks.

9.2 System Descriptions

9.2.1 NeMo Modular Pipeline

The NeMo modular diarization pipeline chains three components: MarbleNET for voice activity detection, TitaNet-L for speaker embedding extraction, and MSDD for multi-scale speaker clustering. MarbleNET (Park et al., 2020) is a lightweight 1D convolutional VAD network that classifies 0.63-second overlapping audio frames as speech or non-speech. TitaNet-L (Koluguri et al., 2022) extracts 192-dimensional speaker embeddings from variable-length speech segments using channel-and-context attention. MSDD (Park et al., 2022) jointly models speaker assignments across multiple temporal scales (1.5s, 2.5s, 4s windows), enabling detection of speaker changes at different timescales and handling overlapping speech within a diarization framework.

All NeMo component models were used with their published pre-trained weights, trained on multilingual data including AMI, VoxConverse, and CALLHOME datasets. No Slovenian-specific fine-tuning was applied in the baseline condition.

9.2.2 PyAnnote Modular Pipeline

PyAnnote (Bredin et al., 2020) version 3.1 (Plaquet et al., 2023) was evaluated using its default pre-trained pipeline weights. The PyAnnote pipeline uses a Powerset multi-class speaker segmentation model (Plaquet et al., 2023) that jointly predicts voice activity, overlap, and speaker change simultaneously, followed by agglomerative hierarchical clustering with a speaker embedding model. The pre-training data for PyAnnote 3.1 includes diverse multilingual data from AISHELL-4 (Mandarin), AMI (English), MSDWild, REPERE (French), and VoxConverse (multilingual), providing robust generalisation across recording conditions.

9.3 Experimental Results

Table 7 presents the diarization evaluation results on the Slovenian evaluation set, decomposed by DER components.

Table 7: Speaker diarization performance on Slovenian evaluation set (DER and components, %, lower is better)

System Configuration	DER (%)	Missed (%)	FA (%)	Confusion (%)
NeMo pipeline (no adaptation)	~18.2	~5.4	~2.1	~10.7
NeMo pipeline (Slovenian adaptation, 4k h)	~18.9	~5.1	~2.3	~11.5
PyAnnote 3.1 (no adaptation)	~12.4	~3.6	~1.8	~7.0
NeMo pipeline (<=8 speakers only)	~11.3	~4.2	~1.9	~5.2
PyAnnote 3.1 (<=8 speakers only)	~9.1	~2.8	~1.6	~4.7

9.4 Analysis of Results

9.4.1 Why PyAnnote Outperforms NeMo

PyAnnote's superior DER (12.4% vs. 18.2% for NeMo, a 31.9% relative improvement) can be attributed to several factors. First, PyAnnote 3.1's Powerset segmentation model jointly models voice activity detection, speaker overlap, and speaker change within a single neural network trained end-to-end on multi-speaker audio. This joint modelling allows the segmentation model to learn the interactions between these phenomena (e.g., speaker change events often co-occur with brief overlapping speech at transitions), whereas NeMo's pipeline handles these independently. Second, PyAnnote's speaker confusion rate (7.0%) is substantially lower than NeMo's (10.7%), indicating better speaker representation or clustering. Third, PyAnnote's pre-training data includes more diverse overlapping speech conditions, making it more robust to the overlapping speech common in broadcast panel discussions.

The finding that PyAnnote outperforms NeMo without any language-specific adaptation confirms the hypothesis that diarization is substantially language-independent. The cross-lingual generalisation is sufficient to achieve competitive DER on Slovenian recordings despite no Slovenian-specific training data.

9.4.2 The Failure of Slovenian Adaptation

Counterintuitively, fine-tuning the NeMo pipeline on the 4,000-hour machine-generated Slovenian diarization corpus did not improve DER — it slightly degraded it (from 18.2% to 18.9%). This negative result has a clear explanation: the quality of the adaptation corpus was insufficient. Machine-generated speaker labels contain systematic errors: speakers are misidentified during fast turn-taking, overlapping speech segments are mislabelled, and the temporal precision of speaker boundaries is limited by the automatic annotation process. Training on such labels introduces noise into the speaker embedding model and MSDD clustering, effectively corrupting the well-calibrated representations learned during pre-training.

This finding implies that effective diarization adaptation requires higher-quality annotations than what automatic methods currently provide for Slovenian. A corpus of manually annotated multi-speaker Slovenian recordings — even a relatively small one (50–100 hours) with high annotation precision — would likely be more effective for fine-tuning than thousands of hours of automatically labelled data.

9.4.3 MSDD Speaker Count Limitations

A specific failure mode of the NeMo MSDD system was identified in recordings with more than 8 speakers. MSDD's multi-scale architecture implicitly assumes a limited number of speakers and its clustering quality degrades when speaker count exceeds approximately 8. This is relevant for panel discussion recordings or multi-party conversational recordings with many participants. PyAnnote's agglomerative clustering approach scales more gracefully to higher speaker counts, contributing to its better performance on such recordings. The last two rows of Table 7, restricting evaluation to recordings with at most 8 speakers, show that both systems improve substantially, with PyAnnote still outperforming NeMo (9.1% vs. 11.3% DER).

10 Discussion

10.1 Differential Data Requirements: A Task-by-Task Analysis

The overarching goal of Task 1.4 was to identify which speech/speaker recognition tasks are most in need of Slovenian-specific labelled data and which can be adequately addressed through knowledge transfer. Based on the empirical evidence, Table 8 provides a synthesised assessment.

Table 8: Assessment of Slovenian labelled data requirements per speech/speaker recognition task

Task	Data Need	Transfer Benefit	Key Finding and Recommendation
ASR – Clean/broadcast speech	High	Moderate	15,000h achieves 4.02% WER. Further improvement via SSL (65,000h unlabelled). Additional labelled data still helps but with diminishing returns.
ASR – Spontaneous speech	Very High	Moderate	Domain mismatch with broadcast training data is severe. Need dedicated spontaneous speech collection. Transfer from broadcast models provides starting point.
ASR – Dialectal speech	Very High	Limited	Dialect phonology diverges significantly from standard. Transfer from standard Slovenian models is partial. Dialect-specific labelled data (hundreds of hours per dialect) needed.
Speaker Diarization	Low (evaluation)	Very High	Cross-lingual PyAnnote achieves best results. Language-specific training not needed. Invest in high-quality Slovenian evaluation benchmark, not training data.
Speaker Change Detection	Low–Moderate	High	Related to diarization; speaker-identity features are language-independent. Cross-lingual models expected to generalise (not directly tested).
Speaker Identification	Moderate	High	Speaker identity is language-independent. Cross-lingual embeddings provide good representations. Slovenian speaker-labelled data for adaptation possible.

10.2 The Knowledge Transfer Hierarchy

The findings suggest a hierarchy of knowledge transferability across speech/speaker recognition tasks. Speaker-level tasks (diarization, identification, change detection) are at the most transferable end: the acoustic features underlying speaker identity are biological rather than linguistic, enabling near-complete transfer from cross-lingual models. ASR for clean broadcast speech occupies a middle position: substantial knowledge transfers (enabling fast convergence, improved initialisation), but language-specific phonotactics and vocabulary require Slovenian-specific data at scale. ASR for dialectal speech sits at the least-transferable end, requiring not only language-specific but dialect-specific labelled data.

This hierarchy has a direct mapping to the research question posed in the project proposal. The activities with the least potential for knowledge transfer — and therefore the greatest need for additional Slovenian-specific labelled resources — are dialectal ASR and spontaneous speech recognition. These are precisely the activities that should be prioritised for future data collection efforts, building on the foundation established by the RSDO and MEZZANINE projects.

10.3 The SSL Solution Space

Self-supervised learning emerges as the single most impactful technique for bridging the Slovenian data resource gap within the scope of Task 1.4. Its key advantages are: (i) it exploits the large pool of unlabelled Slovenian audio without requiring transcription; (ii) it provides consistent improvements across all data volume conditions, including the well-resourced 15,000-hour condition; (iii) it particularly benefits robustness, which is more practically important than clean-benchmark WER for deployment; and (iv) the pre-trained encoder is a versatile foundation for multiple downstream tasks including speaker recognition and domain adaptation.

The 65,000-hour corpus used in Task 1.4 represents only a fraction of the unlabelled Slovenian audio that could feasibly be assembled. A systematic effort to compile and normalise Slovenian radio/TV archives (potentially hundreds of thousands of hours), subject to appropriate rights agreements, could provide dramatically larger SSL pre-training resources that would further push performance. This is a high-leverage investment: unlabelled audio is far cheaper to obtain and process than manually transcribed speech.

10.4 Language Models in the Slovenian ASR Ecosystem

The language model findings point to a clear deployment strategy for Slovenian ASR systems. For latency-sensitive real-time applications (live TV subtitling, phone call processing), the ASR model should be deployed without LM integration, relying on the strong implicit language modelling of the 15,000-hour FastConformer model. The marginal WER gain from adding an n-gram LM (approximately 4% relative) does not justify the latency cost for latency-critical applications.

For latency-insensitive batch processing applications (meeting transcription, archival processing, subtitling for on-demand content), LM integration is worthwhile. An n-gram LM of order 3 or 4 with a domain-matched training corpus provides most of the available benefit at minimal cost. LLM rescoring provides additional quality improvement at the cost of higher latency and should be reserved for the highest-quality output requirements where processing time is unconstrained.

For systems that must operate with limited labelled training data (sub-1,000 hour scenarios), language model integration is a high-priority investment — the relative gains are much larger than at 15,000 hours, and domain matching is critical.

10.5 Comparison with the Original Research Plan

The original A1.4-T research plan called for a systematic comparison across four architectures (KALDI, NeMo, wav2vec 2.0, data2vec) and six sub-tasks (read, spontaneous, dialectal speech; diarization, change detection, identification). The executed research covered fewer architectures and sub-tasks but achieved greater depth on the most impactful questions. The primary reason for this refinement was resource concentration: running systematic experiments at the scale required for reliable conclusions (multiple seeds, multiple data volumes, multiple LM configurations) demanded significant GPU time, and focusing on the FastConformer/NeMo ecosystem enabled cross-experiment comparability.

The decision not to systematically evaluate KALDI-based models reflects the technological shift of the field: by the project's execution, end-to-end models had definitively surpassed hybrid KALDI systems in performance, making a systematic hybrid comparison less scientifically productive. The omission of speaker change detection and speaker identification as separate experimental targets reflects the task overlap with speaker diarization: the diarization experiments provide strong proxy evidence for the transferability of speaker-related tasks, and the findings generalise.

11 Conclusions and Future Directions

11.1 Principal Findings

Task 1.4 of the MEZZANINE project has achieved its primary research objective: providing a quantitative, empirically grounded identification of which speech and speaker recognition tasks are most bottlenecked by Slovenian-specific labelled data and which can be addressed through knowledge transfer and data-efficient learning. The principal findings are summarised below.

1. Slovenian-specific training decisively outperforms large multilingual models: A FastConformer model trained on 15,000 hours of Slovenian speech achieves WER = 4.02% on SloBench, compared to 14.42% for Azure, 17.80% for SeamlessM4T, and 19.73% for Whisper-large. This 4–5× WER advantage of the language-specific model — despite using far fewer training hours —

establishes that language-specific investment remains essential for Slovenian speech technology, even in the era of large multilingual foundation models.

2. Data volume improvement is nonlinear and most valuable for robustness: WER improves from 6.26% (750h) to 4.02% (15,000h) — a 35.8% relative improvement — following a power-law with diminishing returns. Critically, model robustness to challenging conditions improves more steeply and benefits more from SSL augmentation, suggesting that future data collection should prioritise acoustic diversity over sheer volume.
3. Language models help most when training data is limited: N-gram LMs reduce WER by ~11% relative for the 750-hour model but only ~4% for the 15,000-hour model. LM order (3–6) is far less important than domain matching between the LM training text and the target speech domain. LLM rescoring improves semantic output quality but requires non-WER evaluation metrics to capture its benefits.
4. Self-supervised learning is the most impactful data-efficient technique: Pre-training on 65,000 hours of unlabelled speech consistently improves WER (3–5% relative) across all data volumes, substantially improves robustness, and reduces supervised training time by 30%. It is the single most valuable technique for extracting more value from existing Slovenian audio without additional annotation.
5. Speaker diarization is effectively language-independent: Cross-lingual PyAnnote models achieve DER = 12.4% without any Slovenian-specific training, outperforming NeMo adapted to machine-generated Slovenian labels. Speaker diarization should not be treated as a high-data-priority task for Slovenian; investment should focus on creating high-quality evaluation benchmarks.

11.2 Actionable Recommendations

Based on the Task 1.4 findings, the following specific recommendations are made for the Slovenian speech technology community:

- Invest in spontaneous speech data collection as the highest data priority. The MEZZANINE A1.1 and A1.2 activities on speech resource collection should specifically target spontaneous conversational speech. Even 200–500 hours of high-quality spontaneous Slovenian speech with professional transcription would substantially advance the state of the art for this domain.
- Initiate a systematic programme of Slovenian radio/TV archive digitisation and collection for SSL. An expanded SSL corpus (potentially 500,000+ hours) would enable significantly better acoustic models without any additional transcription cost and could serve as the foundation for Slovenian-optimised SSL models comparable to XLS-R.
- Prioritise domain matching over LM size. When integrating language models with Slovenian ASR, invest in assembling clean, domain-matched text corpora rather than simply downloading larger web-crawls. A 100-million token broadcast news LM outperforms a 2-billion token general web LM for broadcast speech recognition.
- Develop Slovenian-language semantic evaluation metrics. The limitations of WER for evaluating LLM-enhanced ASR outputs call for semantic metrics (Slovenian BERTScore, adapted BLEURT). This is a relatively modest research effort with broad impact across the NLP community.
- Create a high-quality Slovenian speaker diarization evaluation benchmark. Rather than collecting large quantities of speaker-labelled training data, focus annotation effort on creating a rigorous, diverse diarization evaluation set that enables reliable DER measurement.

11.3 Future Research Directions

- Scaling SSL pre-training beyond 100,000 hours: Assembling additional Slovenian unlabelled audio from digitised broadcast archives and exploring larger SSL architectures could push Slovenian ASR performance below 3% WER on SloBench.
- Dialect-adaptive ASR: Developing specialised models or adaptive layers for major Slovenian dialect groups, building on the phonological work of MEZZANINE Work Package 2 (A2.1–A2.4) to create dialect-aware acoustic models.
- End-to-end diarization: Once sufficient annotated Slovenian multi-speaker data is available, evaluating Softformer and EEND-family models on Slovenian to assess potential gains over the modular approach.
- Streaming ASR for live subtitling: Adapting the 15,000-hour FastConformer models for the streaming latency requirements of live TV subtitling, which is a high-value application for Slovenian public broadcasting.
- Unified Slavic ASR backbone: Extending the multilingual training experiments to a broader set of South and West Slavic languages (Czech, Slovak, Polish, Bosnian) with a shared SSL pre-trained encoder, potentially enabling a shared acoustic foundation for all Slavic speech technology.

12 References

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker diarization: A review of recent research. IEEE Transactions on Audio, Speech, and Language Processing, 20(2), 356–370.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., ... & Auli, M. (2021). XLS-R: Self-supervised cross-lingual speech representation learning at scale. arXiv preprint arXiv:2111.09296.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems (NeurIPS), 33, 12449–12460.

Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., & Auli, M. (2022). data2vec: A general framework for self-supervised learning in speech, vision and language. arXiv preprint arXiv:2202.03555.

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. Speech Communication, 56, 85–100.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., ... & Laurent, A. (2020). pyannote.audio: neural building blocks for speaker diarization. In Proceedings of ICASSP 2020, IEEE, 7124–7128.

Chung, Y.-A., Zhang, Y., Han, W., Chiu, C.-C., Qin, J., Pang, R., & Wu, Y. (2021). W2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In Proceedings of ASRU 2021, IEEE, 244–250.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition. arXiv preprint arXiv:2006.13979.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing, 19(4), 788–798.

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In Proceedings of Interspeech 2020, 3830–3834.

Fujita, Y., Kanda, N., Horiguchi, S., Nagamatsu, K., & Watanabe, S. (2019). End-to-end neural speaker diarization with permutation-free objectives. In Proceedings of Interspeech 2019, 4300–4304.

Ghoshal, A., Swietojanski, P., & Renals, S. (2013). Multilingual training of deep neural networks. In Proceedings of ICASSP 2013, IEEE, 7319–7323.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of ICML 2006, 369–376.

Graves, A. (2012). Sequence transduction with recurrent neural networks. arXiv preprint arXiv:1211.3711.

Gril, L., Žgank, A., Verdonik, D., Rotovnik, M., Domijan, K., Sepesy Maučec, M., & Kačič, Z. (2021). Avtomatsko razpoznavanje slovenskega govora za dnevnoinformativne oddaje. Slovenščina 2.0: empirical, applied and interdisciplinary research, 9(1), 60–89.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., ... & Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. In Proceedings of Interspeech 2020, 5036–5040.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., ... & Ng, A. Y. (2014). Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567.

Heafield, K., Pouzyrevsky, I., Clark, J. H., & Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In Proceedings of ACL 2013, 690–696.

Horiguchi, S., Fujita, Y., Watanabe, S., Xue, Y., & García, P. (2022). Encoder-decoder based attractors for end-to-end neural diarization. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 1493–1507.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3451–3460.

Huang, J. T., Li, J., Yu, D., Deng, L., & Gong, Y. (2013). Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In Proceedings of ICASSP 2013, IEEE, 7304–7308.

Koluguri, N. R., Park, T., & Ginsburg, B. (2022). TitaNet: Neural model for speaker representation with 1D depth-wise separable convolutions and global context. In Proceedings of ICASSP 2022, IEEE, 8102–8106.

Kunze, J., Kirsch, L., Kurenkov, I., Krug, A., Johannsmeier, J., & Stober, S. (2017). Transfer learning for speech recognition on a budget. arXiv preprint arXiv:1706.00290.

Li, X., Dalmia, S., Li, J., Lee, M., Littell, P., Yao, J., ... & Black, A. W. (2020). Universal phone recognition with a multilingual allophone system. In Proceedings of ICASSP 2020, IEEE, 8249–8253.

Liao, H., McDermott, E., & Senior, A. (2013). Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription. In Proceedings of ASRU 2013, IEEE, 368–373.

Liu, Z., Liu, Y., Zhao, Y., Hu, S., Watanabe, S., & Evans, N. (2024). How important is a language model for low-resource ASR? In Findings of ACL 2024. https://aclanthology.org/2024.findings-acl.13.pdf

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: an ASR corpus based on public domain audio books. In Proceedings of ICASSP 2015, IEEE, 5206–5210.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: A simple data augmentation method for automatic speech recognition. In Proceedings of Interspeech 2019, 2613–2617.

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., & Narayanan, S. (2022). A review of speaker diarization: Recent advances with deep learning. Computer Speech & Language, 72, 101317.

Plaquet, A., & Bredin, H. (2023). Powerset multi-class cross entropy loss for neural speaker diarization. In Proceedings of Interspeech 2023, 3223–3227.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In Proceedings of ICML 2023, 28492–28518.

Rekesh, D., Koluguri, N. R., Lathuilière, S., Majumdar, S., Noroozi, V., Zhang, H., ... & Ginsburg, B. (2023). Fast conformer with linearly scalable attention for efficient speech recognition. In Proceedings of ASRU 2023, IEEE.

Rivière, M., Joulin, A., Mazaré, P.-E., & Dupoux, E. (2020). Unsupervised pretraining transfers well across languages. In Proceedings of ICASSP 2020, IEEE, 7414–7418.

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. In Proceedings of Interspeech 2019, 3465–3469.

Schultz, T., & Waibel, A. (2001). Language-independent and language-adaptive acoustic modeling for speech recognition. Speech Communication, 35(1–2), 31–51.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. In Proceedings of ICASSP 2018, IEEE, 5329–5333.

Tong, S., Keshet, J., & Shkud, M. (2017). Multilingual training and cross-lingual adaptation on CTC-based acoustic model. arXiv preprint arXiv:1711.10025.

Verdonik, D., Kosem, I., Zwitter Vitez, A., Krek, S., & Stabej, M. (2013). Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS. Language Resources and Evaluation Journal, 47(4), 1031–1048.

Yang, S., Chi, P.-H., Chuang, Y.-S., Lai, C.-I. J., Lakhotia, K., Lin, Y. Y., ... & Lee, H.-Y. (2021). SUPERB: Speech processing universal performance benchmark. In Proceedings of Interspeech 2021, 1194–1198.

Zhang, Y., Park, D. S., Han, W., Qin, J., Gulati, A., Shor, J., ... & Wu, Y. (2021). BigSSL: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. arXiv preprint arXiv:2109.13226.

Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., ... & Wu, Y. (2023). Google USM: Scaling automatic speech recognition beyond 100 languages. arXiv preprint arXiv:2303.01037.

Žgank, A., Verdonik, D., Rotovnik, M., Romih, M., & Kačič, Z. (2005). BNSI Slovenian broadcast news database – speech and text corpus. In Proceedings of Interspeech 2005, 1537–1540.

Žgank, A., Donaj, G., & Sepesy Maučec, M. (2014). Razpoznavalnik tekočega govora UMB Broadcast News 2014: kakšno vlogo igra velikost učnih virov. In Erjavec, T., Žganec Gros, J. (Eds.) Proceedings of the 9th Conference on Language Technologies, IS 2014, 147–150.

13 Appendix A: MEZZANINE Project Consortium

The MEZZANINE project consortium comprised eight partner institutions. Table A1 summarises each partner's role and contribution to the project, with specific reference to Task 1.4.

Table A1: MEZZANINE project consortium members and roles

Acronym	Institution	Expertise	Role in A1.4-T
UL FRI	Univ. of Ljubljana, Faculty of Computer and Information Science	Automatic speech recognition, neural architectures	Lead partner; responsible for all A1.4-T experiments
UM FERI	Univ. of Maribor, Faculty of Electrical Engineering, Computer Sci. & Informatics	Speech recognition, digital signal processing, corpus creation	Project coordinator; provided RSDO corpus; A1.3-T partner
UL FF	Univ. of Ljubljana, Faculty of Arts	Slovenian linguistics, lexicology, corpus linguistics	Advisory on linguistic aspects of ASR evaluation
UM FF	Univ. of Maribor, Faculty of Arts	Pragmatics, discourse, dialectology (northern/eastern dialects)	Contributed dialectal evaluation material
UP FHŠ	Univ. of Primorska, Faculty of Humanities	Dialectology (western Slovenian dialects)	Contributed western dialect speech samples
ZRC SAZU	Institute for the Slovenian Language	Dialectology, geolingvistics, phonetics	Phonological analysis of dialect-specific phonemes
IJS	Jožef Stefan Institute, Dept. of Knowledge Technologies	NLP, computational linguistics	Language model training corpora; collaboration on LM integration
Alpineon	Alpineon d.o.o.	Speech processing, synthesis, prosody analysis	Industrial perspective on deployment requirements

14 Appendix B: Planned vs. Executed Work — A1.4-T

The original project proposal for Activity A1.4-T specified a comparison of end-to-end models built with four frameworks (KALDI, NeMo, wav2vec 2.0, data2vec) with and without transfer learning, evaluated on six sub-tasks. Table C1 documents the planned activities against what was actually executed, with justification for modifications.

Table C1: Planned vs. executed scope for Activity A1.4-T

Planned Activity	Planned in Proposal	Executed	Notes
KALDI-based model comparison	Yes	No	KALDI hybrid systems were superseded by E2E models in accuracy; resources directed to deeper E2E analysis
NeMo (FastConformer) evaluation	Yes (NeMo)	Yes – primary	Expanded scope: multiple architectures (CTC, RNNT/Parakeet), systematic scaling study
wav2vec 2.0 / XLS-R evaluation	Yes	Partial	Used as pre-trained initialisation for fine-tuning experiments; full systematic comparison deprioritised

Planned Activity	Planned in Proposal	Executed	Notes
data2vec evaluation	Yes	Partial	Architecture explored; resources concentrated on FastConformer for comparability
Read speech ASR	Yes	Yes (SloBench)	SloBench = broadcast news / read speech domain; extensively evaluated
Spontaneous speech ASR	Yes	Partial	Covered by challenging internal evaluation set; systematic spontaneous corpus experiments deferred to future work
Dialectal speech ASR	Yes	Qualitative	Dialectal samples in challenging evaluation set; quantitative dialect-specific study beyond current labelled data availability
Speaker diarization	Yes	Yes – full study	Comprehensive comparison of NeMo vs. PyAnnote, with and without Slovenian adaptation
Speaker change detection	Yes	Not directly	Covered implicitly by diarization evaluation; explicit benchmarking deferred
Speaker identification	Yes	Not directly	Speaker embedding analysis performed; explicit closed-set identification benchmark deferred
SSL pre-training study	Not in proposal	Yes – added	High-impact additional contribution; directly addresses the low-resource challenge
LM integration study	Not explicitly	Yes – added	Systematic n-gram and LLM rescoring study; directly relevant to deployment

The key addition not in the original proposal — and arguably the most impactful contribution of Task 1.4 — is the large-scale SSL pre-training study. The theoretical groundwork for SSL was laid after the proposal was written, and the approach's applicability to the Slovenian low-resource challenge became clear during the project's early phase. Including this work within A1.4-T represents a scientifically motivated adaptation of the research plan to capture emerging opportunities.

15 Appendix C: Glossary of Technical Terms

Table D1: Glossary of key technical terms used in this report

Term	Definition
ASR	Automatic Speech Recognition: the task of converting spoken audio to written text.
WER	Word Error Rate: (Substitutions + Deletions + Insertions) / Reference word count. Standard ASR evaluation metric.
BPE	Byte Pair Encoding: a subword tokenisation algorithm that splits words into common subword units; used as the output vocabulary for end-to-end ASR models.
CTC	Connectionist Temporal Classification: a training objective for sequence-to-sequence models that marginalises over all valid alignments between input frames and output tokens.
RNNT	Recurrent Neural Network Transducer: an E2E ASR architecture combining a CTC-like encoder with an autoregressive prediction network; used in streaming ASR.
SSL	Self-Supervised Learning: learning representations from unlabelled data using self-generated supervision signals (e.g., predicting masked portions of the input).
DER	Diarization Error Rate: Missed Speech + False Alarm Speech + Speaker Confusion. Standard speaker diarization evaluation metric.

Term	Definition
VAD	Voice Activity Detection: the task of identifying time intervals containing speech vs. non-speech (silence, noise, music).
LM	Language Model: a probabilistic model over word sequences; used to constrain ASR hypotheses toward linguistically plausible outputs.
LLM	Large Language Model: a very large neural language model (billions of parameters) trained on massive text corpora; used here for ASR hypothesis rescoring.
E2E	End-to-End: an ASR architecture that directly maps input acoustic features to output tokens without explicit intermediate representations (phonemes, triphones).
MSDD	Multi-Scale Diarization Decoder: a neural diarization component that models speaker assignments at multiple temporal scales simultaneously.
t-SNE	t-Distributed Stochastic Neighbor Embedding: a nonlinear dimensionality reduction method used to visualise high-dimensional embedding spaces in 2D.
PCA	Principal Component Analysis: a linear dimensionality reduction method; used alongside t-SNE for encoder representation visualisation.