NEUREM3 Final Research Report

Lukáš Burget and Ondřej Bojar - editors

Brno and Prague, January 2025

Executive summary

Neural Representations in Multi-modal and Multi-lingual Modeling (NEUREM3) was a project funded by the Czech Science Foundation (GAČR) program "Research, Experimental Development and Innovation for the Support of Basic Research Grant Projects" – EXPRO from January 2019 till December 2023. This is the final report summarizing the research work in the whole project, particularly the "delta" achieved from the interim report edited in January 2022 and covering years 2019–2021.

The scientific work was articulated around 5 broad areas: Foundations, Interpretability and task-dependence, Tight integration, Robustness, and Relation of neural representations to multilingual concepts. The description of our research is clustered according to 8 tasks defined in the project proposal, and updated in the interim report.

The project lifetime overlapped with a significant boom of large pre-trained foundation models that nowadays dominate natural language processing (NLP), speech processing and especially automated dialogue systems. The team carefully assessed the capabilities of these models, and worked on (and in many cases pioneered) strategies of efficient use of these models in multilingual, low-resource and multi-modal scenarios.

In several of investigated fields (such as speech translation, speaker recognition and diarization, and target speaker extraction) we reached beyond state-of-the-art results and published papers and software tools with significant international impact.

Foreign cooperation in the project was intensive ranging from hiring foreign specialists, through student interns in respected foreign laboratories, to synergies with EU and US projects. The team of the project was consolidated and has a good balance of top-class PI and co-PI, researchers/post-docs, and both Czech and international PhD students.

The project was highly competitive on the international level, the assessment was done via international technology evaluations (challenges / shared tasks), bibliographic metrics, and organization of top international scientific events. NEUREM3 team efficiently cooperated, was and is at the core of building both Czech and international speech, NLP and machine translation communities.

In total, the project led to 195 publications of which 16 were in peer-reviewed international journals and 174 at conferences and workshops. The project also produced a significant amount of research data and code in open repositories. The project led to two proposals for European Research Council (ERC) grants.

Contents

1 Introduction

Neural Representations in Multi-modal and Multi-lingual Modeling (NEUREM3) was a project funded by the Czech Science Foundation (GAČR) program "Research, Experimental Development and Innovation for the Support of Basic Research Grant Projects" – EXPRO 2019 from January 2019 till December 2023.

NEUREM3 was situated the intersection of two important domains of artificial intelligence (AI) – natural language processing (NLP) and speech processing (SP). It proposed a systematic study of neural representation structures in artificial neural networks (NN) for SP and NLP, with a particular focus on automatic speech recognition (ASR) and translation from text (machine translation, MT) or from speech (speech translation, ST) while addressing the lack of fundamental understanding of neural representations and their interplays in machine learning. NEUREM3 concentrated on granularity/size and scope of neural representations, their interpretability, and interactions between long-span and short-span representations. It also investigated into embeddings and end-to-end systems both in single domains, and in systems combining NLP, SP and partly computer vision (CV).

NEUREM3 lifetime spanned an important transition in the AI field. While at the outset of the project in 2019, it was still the norm to have full control of the entire training data and the whole speech or NLP experiment, the development of foundation models (largely fueled by the investors' appetite and huge media interest after the 2022 release of ChatGPT, initially powered by the GPT-3.5 model) led to widespread adoption of foundation models entering all fields of AI. The research landscape of AI significantly changed in the past five years, and these days, almost any AI-related paper either uses directly a pre-trained model, or at least has to compare the results to one of such models; this holds both for NLP and speech. Despite pessimistic opinions, such as "AI research is over, you can't compete with OpenAI", we believe that it was and remains possible to bring significant contributions in foundation models, specifically to:

- Improve them, adapt, reconfigure and fine-tune for own, especially low-resource tasks.
- Use them as powerful feature generators — several "obsolete" and "forgotten" system architectures were brought back to light with the use of powerful pre-trained model generated features.
- Combine / fuse them with others schemes.
- Criticize them (proprietary, energy and data foot-print, etc.) and come up with lighter / own / better / license free models.

Compared to the interim report written in January 2022, that contained only isolated mentions of pre-trained models, this final report documents the significant shift witnessed by the entire AI community. Importantly, a large portion of our reported results builds upon or examines pre-trained models.

1.1 Purpose and structure of this document

The report covers the full duration of the project (i.e. 2019–2023) with some results published also in 2024 and 2025. While it tries to cover the entire period, we assume that

the reader has access to our interim report[1] — especially in the technical part (Section 2) only the "Δ" is covered.

The report is structured according to the requirements of GAČR. After this introduction, the following sections address (the text in italics is quoted from GAČR requirements):

- Section 2 covers *a) the progress of work and the achievement of the objectives set in comparison with the plan set out in the project proposal and in the interim report.* This section is divided into broad topics in which we were working on and is based on publications that are the main output of our project — for each, a summary, the main results, and the scientific context are given.
- Section 3 addresses *b) foreign cooperation;*
- Section 4 describes *c) the participation of individual members of the research team in the solution and results of the grant project, including the involvement of students and postdoctoral students;*
- Section 5 includes *d) evaluation of outputs within the framework of international excellence;* concentrating especially on publications, bibliographic metrics, international technology evaluations (challenges), and organization of top international scientific events.
- Section 6 outlines *e) personnel, organizational and technical process of team building, cooperation of the beneficiary with the project solver and integration of the team into the organizational structure of the institution, cooperation of the beneficiary with another participant.*

An important requirement of the EXPRO scheme was a submission of at least one proposal to the European Research Council (ERC) grant scheme. Section 7 therefore contains the necessary information and evaluation results of our two ERC proposals submitted within NEUREM3.

The last Section 8 lists all the project publication outputs. It is divided into individual years of project execution, and the citations include hyperlinks either to university repositories or to the public archives so that an interested reader just needs one click. At the very end, a standard section with references (other than project outputs) is appended.

1.2 Global picture

While building upon the state-of-the-art systems and experimental results, we were addressing fundamental issues that are neglected in current research: hierarchy of neural representations, human interpretability, multi-lingual and multi-modal issues, and training under realistic conditions of non-ideal and incoherent data. All this happened in the era of rapid changes and developments in the AI paradigm, transitioning from "fully controlled" models to ones built on large pre-trained models.

Our research in NEUREM3 can be categorized into five broad areas, as defined in the project proposal:

[1] Full text available at https://www.fit.vut.cz/research/publication/12682/.en

- Area 1. Foundations: Setting up baselines, defining a hierarchy of neural representations categorized by granularity/size and scope, studying evaluation of information content.
- Area 2. Interpretability and task-dependence: Studying interpretability of neural representations learned for various tasks, an investigation into task-dependence, portability, and the interplay between long-span and short-span representations. Multi-task training.
- Area 3. Tight integration: Exploring architectures combining SP, NLP, and CV, an investigation into NN embeddings as information carriers among the modalities.
- Area 4. Robustness: Training neural representations on low quality, heterogeneous and non-parallel corpora, end-to-end systems.
- Area 5. Relation of neural representations to multi-lingual concepts.

As will be seen in the technical Section 2, we were addressing all of them. The most important aspect, the hierarchy of neural representations for speech and NLP, can be seen in Figure 1. The horizontal axis plots the granularity of inputs, the span of individual rectangles covers their inputs and outputs (the input and output granularities are often different – we are sorry for the coarseness of this representation). Blue color codes "speech", red "NLP/MT", violet is on the boundary of the two, and green codes "semantics".

1.3 Adherence to the plan

The detailed research plan and methodology set in the project proposal consisted of five tasks. Three others (6, 7, and 8) were added in the interim report.

1.3.1 Task 1. Multi-linguality in ASR, NLP, and MT.

Multi-linguality was at the very core of NEUREM3, as most of the investigated techniques were multi-lingual or language-agnostic — for example in the development of speech recognition (Section 2.5). The whole ST and MT tasks (see below for Task 5 and Sections 2.8 and 2.9) are inherently multi-lingual, but an MT approach making direct use of multi-lingual data for better MT of low-resource languages (Section 2.9.3) is a clear representative of a successful Task 1 work. Language recognition, especially the work we did in evaluations (Section 2.4.1) concentrates on detecting languages. In some tasks however, we attempted to suppress the information on language as a nuisance, such as during training and testing speaker recognition on real-world data (Section 2.2.2) or in the development of speaker recognition training and scoring paradigmata (Section 2.2.3).

1.3.2 Task 2. Multi-modality in ASR, NLP, and MT

The results of Task 2 include "obvious" multi-modal investigations, included primarily in Language and vision (Section 2.10), such as using position information in images (Section 2.10.2) or visual question answering (Section 2.10.3). However, we also extended the notion of multi-modality to smart re-purposing of techniques investigated in speech or

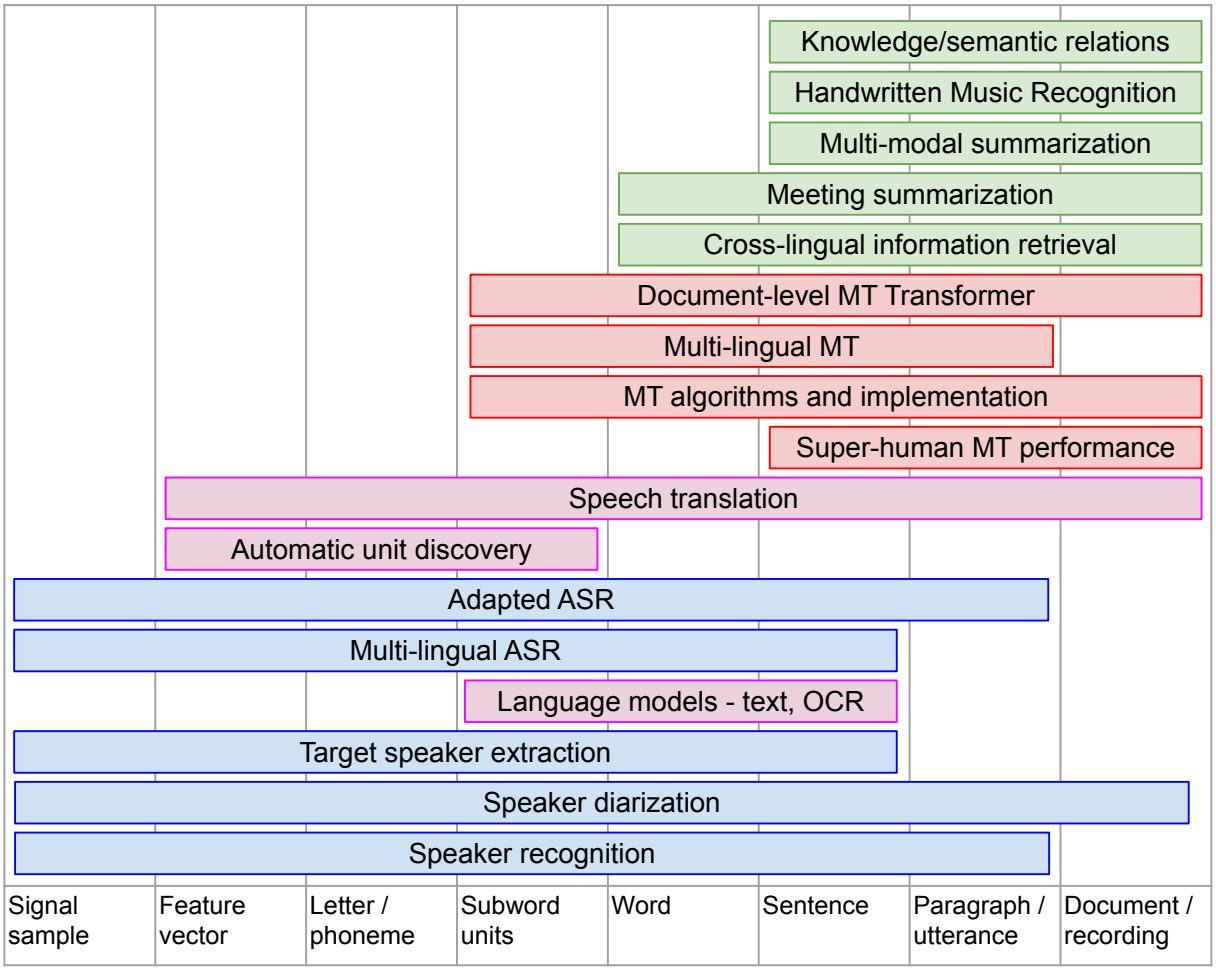

Figure 1: Overall scheme of investigated neural representations across unit granularity - updated version of the same figure from the interim report. Color coding proceeds from speech processing (blue), text translation (red), with violet indicating both of them, up to semantics (green).

NLP on other modalities – historical document recognition (Section 2.7.1) and handwritten music recognition (Section 2.10.5). Note also, that meta-information available on a piece of media can also be considered another modality – see for example Section 2.2.2 on gathering weakly-labeled data for speaker recognition. Finally, in speech processing, 'modality' can also refer to the use of various types of information available or generated for a speech signal or a text. For example, in speech recognition with a particular target-speaker (Section 2.1.2), the information on the speaker is used to clean the signal for ASR either step-by-step or in an end-to-end fashion.

1.3.3 Task 3. Rich input, intermediate, and output representations in neural ASR, NLP, and MT systems.

This (rather broad) task includes our work on a wide range of representations of different granularity ranging from raw speech signal (such as in feature extraction for speaker recognition, Section 2.2.1) to the highest information levels used for our work on semantics (Section 2.11). This task actually underwent the greatest change from our interim

report, as many signal or text representations that were done by models fully under our control, transitioned to the use of large pre-trained models. The whole segments of ASR (Section 2.5) and keyword-spotting (Section 2.6) started to be dominated by pre-trained models, as well as the ST field (Section 2.8). We can proudly state that in some areas, such as in the use and adaptation of large pre-trained models for speaker recognition (Section 2.2.4) or their successful deployment in practical live speech translation applications (Section 2.8.4), the work of NEUREM3 members was truly pioneering. Note that in Figure 1, the "bars" for most speech tasks are extended to the left compared to the figure in the interim report, as most of pre-trained models accept raw waveform as their input.

1.3.4 Task 4. Hierarchies and automatic inference of units

Standard machine learning works with units defined beforehand, however, in some situations, this can be sub-optimal. In section Bayesian models for acoustic unit discovery (Section 2.6.2), we report on our efforts to automatically infer acoustic units that would perform equally or better than hand-crafted representations, especially for unknown languages. Our prior work on end-to-end ST led to ASR with efficient out-of-vocabulary (OOV) word processing (Section 2.5.1), where hierarchy of units was used to represent both known and unknown words. An important work on units was done in the area of keyword spotting (Section 2.6.3) where classical graphemes (for text queries) or signal segments (for querying by example) were replaced by universal embeddings taking into account both modalities. A somewhat special need of processing units arises when we want to handle unlimited inputs, such as in long-form speech recognition or translation (Section 2.8.4). The work on units extends also to the other end of our research — in our "semantic" work we focus on compositionality of the representation in sequence-to-sequence models (Section 2.11.2) and on relations between sentence-level units (our COSTRA corpus of complex sentence transformations and relations between these units, see Section 2.11.1).

1.3.5 Task 5. Text to text and speech to text translation based on non-parallel and heterogeneous training data, robustness towards the noise

Task 5 was populated by a range of topics in speech translation and neural machine translation. In ST (Section 2.8), care was taken of the whole processing pipeline, from preparation of data (Section 2.8.1), through different approaches to ST, such as end-to-end (Section 2.8.2), long-form simultaneous ST (Section 2.8.5) and multi-source simultaneous ST making use of original speech and human interpretation in another language (Section 2.8.6) to proper definition of proper evaluation and organizations of shared task in ST (Sections 2.8.3 and 2.8.7).

This task however included also work on "classical" neural machine translatin (Section 2.9). Here, the work started again from challenging non-parallel (Section 2.9.2) and multilingual (Section 2.9.3) data and included a range of theoretical and practical issues of MT, such as the use of negative lexical constraints (Section 2.9.4), MT decoding using a genetic algorithm (Section 2.9.6), non-autoregressive NMT, (Section 2.9.7), length control in machine translation (Section 2.9.5), or attention regularization and Transformer optimization in NMT (Sections 2.9.8 and 2.9.9). Similarly as for the other Tasks,

properly evaluating MT quality was addressed (Section 2.9.1).

1.3.6 Task 6. Personality and individual adaptation

This task was newly introduced to cover our work in individual adaptation and relation to the personality of author/writer/speaker. It was naturally addressed by target speaker speech extraction (Section 2.1.2) and the whole broad areas of speaker recognition (Section 2.2) and speaker diarization (Section 2.3). Here, the highlights include the advances in large pre-trained models in speaker recognition (Section 2.2.4), and our highly cited work on Bayesian HMM Speaker Diarization (VBx, Section 2.3.1) with follow-ups in end-to-end speaker diarization (Section 2.3.3). Another example are the OCR systems for historical hand-written text (Section 2.7.1), with of strong adaptation to writing style of the author. Beyond the individual person level, we experimented with genre adaptation when preparing synthetic training data (Section 2.8.1).

1.3.7 Task 7. Semantic processing

The task operating at a high level of language structures hierarchy, namely semantics, was newly introduced to cover our work going beyond speech, NLP and MT. It is mainly aligned with research topic "Towards Semantics" (Section 2.11) and closely linked to Task 8 on human factors.

Taking some manually created annotations for aspects of semantics,[2] we are interested in finding these aspects in the continuous representations learned automatically by the models as well as in their operations.

We proceed both bottom-up, from individual elements, as well as top-down, directly targetting a complex goal. For the former, our work analyzed the knowledge of semantic relations in pre-trained models (Section 2.11.1) and compositionality in sequence-to-sequence models (Section 2.11.2). For the latter, we focussed on the task of automatic meeting summarization where the transcript from genuine project meetings contains many pieces of information spread across not always coherent speeches of multiple speakers and the goal is to automatically assemble them into concise minutes (Section 2.11.3).

The work on Task 7, on meaning representation and automatic language understanding was also the inspirational source and basis for the ERC Synergy application by Ondřej Bojar (Section 7.1).

1.3.8 Task 8. Human performance and human factors

Natural languages are a product of the human brain and the human society. The whole field of NLP attempts to put artificial agents in the same communication situations and tries to design them so that they mimic the behaviour of humans. In Task 8, we focus specifically on this similarity between human and machine processing of language. Given the field of study and the data we created in NEUREM3, this comparison targets primarily the translation process and translation outcomes.

Starting with low-level and easy-to-observe features of text translation (Section 2.12.1) and comparing human interpreting and speech translation (Section 2.12.2) we move on

[2] We could have tried to start with some complex manual representation of the meaning of the sentences, such as Abstract Meaning Representation, but we saw such representations hard to achieve for a start.

to close collaboration with translation theoreticians. The problem in current machine translation research for high-resource languages is the very good achievable performance. When we matched human translation quality in English-to-Czech translation [Pop+20], we knew that navigating research further will be difficult. In order to prepare the space for this exploration (comparing human and machine translation at quality levels higher than "standard professional translation"), we designed a methodology for creating very good reference translations (Section 2.12.3), making a solid stepping stone for further research.

As the last element of our work on relating humans and machines when processing the language, we focus on objectively observable quantities when humans are carrying language processing tasks. We start with their gaze patterns when translating or reading but in the long term beyond NEUREM3, we would like to explore also e.g. EEG (Section 2.12.4).

Again, the work on Task 8 and the desire for more human-like processing in artificial language models, was the basis of our grant proposal; a joint CUNI-BUT GAČR EXPRO application in the 2024 call which was unfortunately not selected for funding.

2 Research topics in detail

The work in the project was articulated around several main research topics, that intersect with the broad areas defined in Section 1.2 and with the tasks specified in the original proposal and updated in interim research report as described in Section 1.3 here:

- Neural signal processing
- Speaker recognition
- Speaker diarization
- Language and emotion recognition
- Automatic speech recognition
- Automatic speech unit discovery and keyword search
- Between speech and NLP
- Speech translation
- Neural machine translation
- Language and vision
- Towards semantics
- Human performance and human interfaces

The details are in the following sections.

2.1 Neural signal processing

2.1.1 Speech separation and extraction

In the past few yeas, a number of time-domain speech separation methods have been proposed. However, most of them are very sensitive to the environments and wide domain coverage tasks. In [Han+22], from the time-frequency domain perspective, we propose a densely-connected pyramid complex convolutional network, termed DPCCN, to improve the robustness of speech separation under complicated conditions. Furthermore, we generalize the DPCCN to target speech extraction (TSE) by integrating a new specially designed speaker encoder. Moreover, we also investigate the robustness of DPCCN to unsupervised cross-domain TSE tasks. A Mixture-Remix approach is proposed to adapt the target domain acoustic characteristics for fine-tuning the source model.

We evaluate the proposed methods not only under noisy and reverberant in-domain condition, but also in clean but cross-domain conditions. Results show that for both speech separation and extraction, the DPCCN-based systems achieve significantly better performance and robustness than the currently dominating time-domain methods, especially for the crossdomain tasks. Particularly, we find that the Mixture-Remix finetuning with DPCCN significantly outperforms the time-domain baseline for unsupervised cross-domain TSE, with around 3.5 dB performance improvement on target domain test set, without any source domain performance degradation.

Recently designed time-domain single-channel separation models have achieved remarkable output quality in clean-audio conditions. A significant drop in performance

is observed when deploying such neural networks in challenging acoustic conditions (including noise and reverberation, common phenomena occurring in recording enclosures, such as rooms). In speech processing, the use of multiple microphones generally helps alleviate the difficulty of far-field data. Inspired by this, we designed an approach to building multi-channel source separation models by extending current strong single-channel time-domain separators in [Moš+23]. Our approach allows reusing pre-trained models by inserting a designed lightweight reference channel attention (RCA) combiner, the only trained module. It comprises two blocks: 1) The former allows attending to different parts of other channels w.r.t. the reference one. It was motivated by the potential ability to align information from various channels and, eventually, recover the information occluded in the reference channel from other channels. 2) The latter provides an attention-based combination of channels.

The proposed approach is compatible with various time-domain models. Moreover, it allows us to conveniently estimate beamforming weights and perform a combination of channels through spatial filtering. In our experiments, we emphasize the synergy of multi-channel signal processing (beamforming) and network-based separation: the final models combining the two yields the best performance (especially when using more microphones).

2.1.2 Target speaker speech extraction

Speech technologies often suffer from problems when presented with an overlapping speech of multiple speakers. The work in this area followed on our successful introduction of SpeakerBeam — a neural network, which takes the mixture and the adaptation utterance at the input and outputs the extracted target speech. The IEEE Journal of Selected Topics in Signal Processing paper [Žmo+19] is actually the second most cited paper of the whole project (see data in Section 5.1.1). It also served as a basis for Katerina's PhD thesis defended in 2022 (see Section 5.4).

In [Pen+24b], we advanced the idea of target speaker extraction into the era of self-supervised learning (SSL) models. These models have shown remarkable success across various speech tasks. However, their application to target speech extraction (TSE) — isolating a target speaker's voice from mixtures — remained underexplored. This work, developed in cooperation with NTT, introduced TSE as a downstream task for SSL models, leveraging pre-trained features for mixture processing and speaker embedding extraction. We propose an enhanced TSE system with two novel modules: the adaptive input enhancer (AIE) for capturing hierarchical features and a speaker encoder for robust embedding generation. Experimental results on LibriMix datasets revealed significant gains over existing systems, with a 14.0 dB SI-SDR improvement. Our contributions included integrating CNN and Transformer layers of SSL models, enhancing performance through multi-scale feature representation, and achieving further gains by fine-tuning SSL parameters. This demonstrated the importance of strategic SSL feature utilization in TSE frameworks, offering insights for improved real-world multi-talker speech processing systems.

The second paper in the series is [Pen+24a], concentrated on probing SSL models with target speech extraction, as none of previous evaluation schemes (for example the

known SUPERB challenge[3] evaluated the SLL models' performance simultaneously on speaker recognition and signal processing tasks. Experiments on Libri2Mix revealed that TSE performance could not be directly inferred from related tasks like speaker verification or separation, highlighting its unique evaluation value. We benchmarked nine SSL models and showed that WavLM Base Plus achieved superior results for TSE and speaker verification, emphasizing the importance of robust pre-training. While the proposed model trains faster than advanced systems like TD-SpeakerBeam, it shows room for improvement, particularly in SSL architectures and temporal resolution. These findings underscore TSE as a critical benchmark for probing SSL models in real-world multi-speaker scenarios.

2.2 Speaker recognition

2.2.1 Feature extraction for speaker recognition

Recent advances in learnable feature extraction for speaker verification systems have highlighted the potential of filters learned in both time and frequency domains. However, existing approaches often result in filters that closely resemble their initialization (e.g., Mel filterbank) or are constrained by rigid assumptions. To address these limitations, we proposed LearnSF [Pen+22a], a novel learnable sparse filter-bank designed to optimize sparsity without enforcing predefined distributions. Our approach builds upon a standard pre-processing pipeline, including STFT and magnitude spectrum calculation. LearnSF employs a sparsity-driven optimization framework that selectively activates critical frequency components relevant to speaker identity, minimizing redundancy across filters. These features are then processed with logarithmic compression and mean-variance normalization before being passed to a neural network for speaker classification. LearnSF achieved state-of-the-art results, particularly in challenging cross-language scenarios, demonstrating its potential as a robust learnable front-end for speaker verification systems. This work was evaluated on VoxCeleb and CNCeleb datasets, showing consistent improvements over traditional features (e.g., Mel-FBank, MFCC) and existing learnable filters.

2.2.2 Training and testing speaker recognition on real-world data

Real world applications rarely come with large volumes of data with precise speaker labeling. In [Sta+22], we demonstrated a method for training speaker embedding extractors using weak annotation. More specifically, we used the full VoxCeleb[4] recordings and the name of the celebrity appearing on each video without knowledge of the time intervals the celebrity appears in the video. We showed that by combining a baseline speaker diarization algorithm that requires no training or parameter tuning, a modified loss with aggregation over segments, and a two-stage training approach, we were able to train a competitive ResNet-based embedding extractor. Finally, we experimented with two different aggregation functions and analyze their behaviour in terms of their gradients. The paper was a nice example of cooperation between BUT and Omilia R&D (Greece) headed by Dr. Themos Stafylakis.

[3] https://superbbenchmark.github.io/

[4] https://www.robots.ox.ac.uk/~vgg/data/voxceleb/

In [Sil+22], we analyzed speaker verification embedding extractors and back-ends under language and channel mismatch. This paper was inspired by the novel test scenario introduced in NIST SRE 2021. Compared to previous editions of NIST speaker recognition evaluations, this time the task included cross-language and cross-domain target trials. Under such conditions, both front- and backend speaker verification models are required to be robust under domain and language mismatch to prevent an increased amount of "miss" errors. In the paper, we analyze the behavior and performance of speaker embeddings and the back-end scoring model and suggest the possible directions of improving them under such conditions. We present our findings regarding ResNet-based speaker embedding architectures and show that reduced temporal stride yields improved performance. We then consider a PLDA back-end and show how a combination of small speaker subspace, language-dependent PLDA mixture, and nuisance-attribute projection can have a drastic impact on the performance of the system. Besides, we present an efficient way of scoring and fusing class posterior logit vectors recently shown to perform well on speaker verification task.

2.2.3 Speaker recognition training and scoring paradigmata

In [Pen+22b], we introduced Progressive contrastive learning (PCL) for self-supervised text-independent speaker verification. Self-supervised learning (SSL) methods have gained attention for speaker verification but are limited by reliance on pre-defined cluster numbers, which are hard to estimate for large-scale unlabeled data. We proposed PCL to dynamically estimate clusters based on data statistics at each step, progressively approaching the true speaker count. PCL utilizes eigendecomposition to estimate cluster numbers and K-means clustering to assign data, optimizing speaker representations through a redesigned contrastive loss. Our method outperforms existing SSL frameworks on VoxCeleb1, achieving superior performance with improved training stability and generalization. By clustering only memory queue data rather than the entire dataset, PCL offers a time-efficient, scalable approach for speaker verification, addressing the limitations of static clustering-based SSL systems.

Paper [Bru+22] is a contribution to scoring of speaker embeddings using Probabilistic Spherical Discriminant Analysis as an alternative to PLDA for length-normalized embeddings. In speaker recognition, where speech segments are mapped to embeddings on the unit hypersphere, two scoring backends are commonly used, namely cosine scoring or Probabilistic Linear Discriminant Analysis (PLDA). Both have advantages and disadvantages, depending on the context. Cosine scoring follows naturally from the spherical geometry, but for PLDA the blessing is mixed—length normalization Gaussianizes the between-speaker distribution, but violates the assumption of a speaker-independent within-speaker distribution. In this paper, we proposed a new scoring backend called Probabilistic Spherical Discriminant Analysis (PSDA), an analog to PLDA that uses Von Mises-Fisher (VMF) distributions on the hypersphere for both within and between-class distributions. We demonstrated how the self-conjugacy of VMF gives closed-form likelihood-ratio scores, making it a drop-in replacement for PLDA at scoring time. All kinds of trials can be scored, including single-enroll and multi-enroll verification, as well as more complex likelihood-ratios that could be used in clustering and diarization. We also described training EM-algorithm with closed-form updates.

The follow-up [Sil+23b] extends the previous PSDA work to toroidal PSDA (T-

PSDA). Compared to the original PSDA, T-PSDA has the ability to model within and between-speaker variabilities in toroidal submanifolds of the hypersphere providing greater flexibility when modeling the embedding space - something that we found lacking in the original PSDA. Like PLDA and PSDA, the model still allows closed-form scoring and closed-form EM updates for training. We tested T-PSDA on two datasets: VoxCeleb and NIST SRE 2021, typically, cosine scoring performs well on the first one, while trained backend model as PLDA typically outperforms cosine scoring on NIST scenarios. In our experiments, we find T-PSDA accuracy on par with cosine scoring for VoxCeleb, while PLDA accuracy is inferior to both of them (confrning the previous experience). On NIST SRE'21 we find that T-PSDA gives large accuracy gains compared to both cosine scoring and PLDA.

Previous work on probabilistic speaker representations in speaker recognition and diarization (for example [Sil+20]) led to successfully defended PhD thesis of Anna Silnova (see Section 5.4) in which Gaussian Probabilistic Linear Discriminant Analysis (G-PLDA), heavy-tailed PLDA (HT-PLDA) and neural approaches to derive not point embeddings, but thesir distributions, were examined.

2.2.4 Large pre-trained models in speaker recognition

Given by the success of SSL models in other areas of speech processing, we naturally turned to their use in speaker verification. These efforts led to a series of important papers, concentrating on making he SSL models fit for the task.

In [Pen+23a], we investigated fine-tuning strategies for adapting pre-trained Transformer models to speaker verification. We proposed a lightweight, convolution-free backend based on multi-head factorized attentive pooling (MHFA) to efficiently extract speaker representations. The method aggregates frame-level features into phonetic clusters and combines them into utterance-level embeddings. Additional techniques, including layer-wise learning rate decay (LLRD) and fine-tuning regularization, stabilize the training process and improve performance. The proposed method achieves state-of-the-art results on VoxCeleb1, with 0.59% EER on Vox1-O while significantly reducing training time to 4 hours. Our approach outperforms traditional TDNN-based systems when integrated with pre-trained models such as Wav2Vec2.0, HuBERT, and WavLM, demonstrating the effectiveness of attention mechanisms for speaker verification tasks.

In [Pen+23c], we investigated parameter-efficient transfer learning (PETL) methods for speaker verification, addressing the challenges of full fine-tuning large pre-trained Transformer models. We proposed using lightweight adapters integrated into each Transformer block while freezing the pre-trained parameters. Our experiments compared three PETL methods: bottleneck adapter, prefix tuning, and mix-and-match adapter (MAM). The MAM adapter achieves the best performance by balancing efficiency and accuracy, requiring fewer than 4% of parameters to be updated during fine-tuning. To improve performance in low-resource and cross-language scenarios, we introduce a two-step tuning strategy that pre-trains the model on an intermediate dataset before fine-tuning it on a small target dataset. Results show that this approach significantly enhances robustness and achieves state-of-the-art performance on VoxCeleb and CNCeleb datasets, demonstrating its effectiveness for adapting large-scale pre-trained models to speaker verification tasks.

The next paper in the series, [Pen+23b], addressed improving speaker verification

with self-pretrained Transformer models. Obviously, pre-trained Transformer models have demonstrated success in speech processing, yet challenges remained in adapting them effectively for speaker verification tasks. We proposed self-pretraining, a hierarchical training approach where models are pre-trained and fine-tuned on the same dataset to reduce domain shifts and improve speaker representation discriminability. Experiments on VoxCeleb and CNCeleb datasets using three architectures—HuBERT, Conformer, and WavLM—show significant performance gains compared to traditional generalist pretraining on Librispeech. Key findings reveal that self-pretrained models achieve comparable results to models pre-trained on 94k hours of data using only one-third of the dataset size. The Conformer architecture outperforms others, especially with the lightweight MHFA backend. These results validate the effectiveness of self-pretraining in capturing speaker-specific information, enhancing both accuracy and generalization for downstream tasks.

Additionally, MHFA has proven versatile and effective in other speech-related tasks, including dialect recognition in the NAKI project JARIN (see Section 3.5) and language identification (LID, see Section 5.5.3), further demonstrating its robustness and adaptability across diverse speech processing applications.

2.2.5 Multi-channel speaker recognition

Single-channel speaker verification has a longstanding history in the research community. Due to its potential applications, such as the personalization of smart devices, our goal is to foster advances in multi-channel speaker verification. Over the last year, studies on the topic have used different (often self-designed) datasets. It hinders fair comparison, making it challenging to identify promising techniques. Motivated by the unconsolidated data situation and the lack of a standard benchmark in the field, we complemented our previous efforts by designing a comprehensive corpus for training and evaluating text-independent multi-channel speaker verification systems in the presence of reverberation and an interfering source [Moš+22b]. It can also be readily used for de-reverberation, de-noising, and speech enhancement experiments since we provide clean and reverberant targets of the present sources.

Since collecting a realistic multi-channel training corpus satisfying the data demand of the current speaker embedding extractors is both expensive and resource-demanding, we opted for a compromise — data simulation on top of clean parts of a large-scale single-channel dataset. The development and evaluation trials were created from publicly available data that has been retransmitted. We published full recipes that create the dataset from public sources as the MultiSV corpus,[5] and we provided results with two baseline systems.

Our previous work and studies of other authors evidenced gains in far-field speaker verification from employing multi-channel pre-processing that precedes speaker embedding extraction. The type of pre-processing is, however, crucial: while various multi-channel enhancement and source-separating networks directly predicting sources have been developed, the non-linear distortions and artifacts may hurt the performance of the downstream speaker verification model. Linear filtering performed by a beamformer represents a plausible alternative.

[5] https://github.com/Lamomal/MultiSV

In order to estimate beamforming weights, spatial covariance matrices (SCMs) for speech and noise are required. A well-known approach to estimating them utilizes neural networks trained to predict time-frequency (TF) masks representing a prevalence of speech or noise in the TF bins. Our main contribution in [Moš+22a] was introducing an alternative way of predicting SCMs for a beamformer from the waveform (motivated by the power of recent time-domain enhancement/separation models). Finally, due to the differentiability of all the modules in the beamformer and embedding extractor, we were able to fine-tune our multi-channel frontend w.r.t. speaker verification objective. Our experiments on the MultiSV data revealed consistent improvements with a 2.7 times smaller model compared to the described baseline.

2.2.6 Speaker recognition Evaluation Systems

Organization and participation in international challenges is a key activity in speech, NLP and generally AI research. Preparing, tuning and training challenge systems is however also a serious research endeavor. This chapter therefore contains several sections on these systems, while a broader view on challenges and evaluations is presented in Section 5.5,

In [Ala+22], we provide a description of the ABC team's collaborative efforts toward the development of speaker verification systems for the NIST Speaker Recognition Evaluation 2021 (NIST SRE 2021). The evaluation included three distinct tasks: audio-only, audio-visual, and visual-only verification. While [Sil+22] is also based on the data and task introduced in NIST SRE2021, in [Ala+22], we concentrated on analyzing the performance of audio-based speaker verification system only. This publication, on the other hand, is supposed to provide a complete and detailed description of our contribution to all three tasks without focusing on any particular one.

2.2.7 High-level summary of speaker and language recognition

In 2023, the PI of NEUREM3 Lukáš Burget and the head of BUT speech lab, Prof. Černocký, had the honor to be invited to contribute sections on speaker and language recognition and speaker diarization for a special edition of Signal Processing Magazine issued at the occasion of 75th anniversary of IEEE Signal Processing Society. Paper [Yu+23] coordinated by Dong Yu (Tencent) and co-authored by several highly influential and highly cited speech processing scientists covered the dramatic developments in the last 25 years of Speech and Language Processing R&D.

2.3 Speaker diarization

The speaker diarizarion saw a significant increase of activity in the 2nd part of the project and was an important part of Task 6. Personality and individual adaptation newly introduced in our interim report.

2.3.1 Bayesian HMM speaker diarization (VBx)

The key journal paper [Lan+22a] summarized our work of the previous period — we presented the derivation and update formulae for the variational Bayes x-vector (VBx)

model, focusing on the efficiency and simplicity of this model. Besides, we carried out an extensive comparison of performance of the VBx diarization with other approaches in the literature, showing that VBx achieves superior performance on three of the most popular datasets for evaluating diarization: CALLHOME, AMI and DIHARDII. The paper and the associated diarization recipe[6] became extremely popular in the community, [Lan+22a] is the 3rd most cited paper of the whole project (see Section 5.1.1), and it laid the base for extensice follow-up research both at BUT and world-wide.

Building on the VBx foundation, and in collaboration with NTT, we recently proposed Discriminative VBx (DVBx) [Kle+24], an enhanced framework that incorporates discriminative training to fine-tune VBx parameters in an end-to-end manner. Unlike the original VBx, which relies on exhaustive grid searches for hyperparameter tuning, DVBx optimizes parameters directly by minimizing a novel Expected Detection Error (EDE) loss. This loss function better aligns with diarization error rates compared to the conventional binary cross-entropy loss, ensuring improved performance consistency.

The proposed method significantly simplifies hyperparameter tuning, eliminating the need for labor-intensive grid searches, while achieving comparable or better performance than baseline VBx models across datasets such as AMI, CALLHOME, and DIHARD II. Additionally, the framework demonstrates the benefits of fine-tuning PLDA parameters, particularly in addressing domain mismatches and enhancing system robustness. By introducing a structured approach to optimizing VBx parameters, DVBx bridges the gap between clustering-based and end-to-end diarization methodologies.

2.3.2 Data for end-to-end speaker diarization

While modular diarization systems have attained very strong performance in certain scenarios, they struggle when dealing with conversations with high levels of overlap as meetings or casual conversations. Besides that, modular pipelines tend to be cumbersome and slow. To address these shortcomings, we also worked on end-to-end systems. We first focused on an important aspect, training data. These models require large amounts of training data with speaker-segment annotations. In academic settings, data with manual diarization annotations are scarce so the compromise solution consists in generating training data automatically. A common approach has been to utilize simulated mixtures, as in source separation. However, simulated mixtures do not resemble real conversations in many ways. In [Lan+22b] we presented an alternative method for creating synthetic conversations that resemble real ones by using statistics about distributions of pauses and overlaps estimated on genuine conversations. Furthermore, we analyzed the effect of the source of the statistics, different augmentations and amounts of data. We demonstrated that our approach performs substantially better than the original one, while reducing the dependence on the fine-tuning stage. Experiments were carried out on 2-speaker telephone conversations of Callhome and DIHARD 3.

In [Lan+23] we extended the approach to more than two speakers per conversations and showed similar advantages as before. We also created simulated conversations with wide-band public audio sources and presented an analysis on several evaluation sets. Together with these publications, we released our implementations of EEND[7] and the

[6] https://github.com/BUTSpeechFIT/VBx
[7] https://github.com/BUTSpeechFIT/EEND

method for creating simulated conversations[8].

2.3.3 End-to-end speaker diarization and approaches to correct diarization outputs

Besides the training data used for end-to-end models, we also focused on the architectural choices related to them. End-to-end neural diarization with encoder-decoder attractors (EEND-EDA) has been the predominant line of research. This approach uses the notion of attractor to model a speaker. While EEND-EDA can handle a variable number of speakers in a conversation (by decoding a number of attractors dependent on the input), the performance on recordings with more than three speakers is considerably worse. To address this, in [Lan+24] we replaced the EDA module with a Perceiver-based one and showed its advantages over EEND-EDA; namely obtaining better performance on the largely studied Callhome dataset, finding the quantity of speakers in a conversation more accurately, and faster inference time. Furthermore, when exhaustively compared with other methods, our model, DiaPer, reached remarkable performance with a very lightweight design. Together with this publication, we released the code of DiaPer as well as models trained on public and free data.[9] This work was also the final contribution in Federico Landini's PhD thesis (see Section 5.4).

Despite error correction techniques being common in automatic speech recognition (ASR), there are only a few related studies in speaker diarization. In [Han+24], we explored to refine diarization results similar as the error correction in ASR. To this end, we proposed an error correction framework for speaker diarization which we name DiaCorrect.[10] It automatically refines the diarization results provided by an initial diarization system, by exploiting the interactions between the input acoustic features and the initial speaker activity predictions (SAPs) with two parallel convolutional encoders and a transformer-based decoder. To reduce over-fitting in the error correction process, we adopt data pruning to select hard samples from the simulated training set. Moreover, by analyzing the distribution of initial SAPs, we calibrate them for the inference. We analyzed DiaCorrect's performance under two scenarios. If only simulated data are available, DiaCorrect can be trained on a small set of hard samples. If some limited target domain samples are provided, DiaCorrect can be fine-tuned or even trained from scratch on such data. Experiments on 2-speaker telephony data show that the proposed DiaCorrect can effectively improve the initial model's performance.

2.3.4 End-to-end models combined with Bayesian HMM diarization

Modular and end-to-end speaker diarization systems provide advantages in different scenarios. For example, modular systems are usually more accurate finding the right number of speakers in a conversation while end-to-end systems handle overlaps more accurately. Combining end-to-end neural speaker diarization (EEND) with vector clustering (VC), known as EEND-VC, has gained interest for leveraging the strengths of both methods. EEND-VC estimates activities and speaker embeddings for all speakers within an audio chunk and uses VC to associate these activities with speaker identities across different

[8] https://github.com/BUTSpeechFIT/EEND_dataprep
[9] https://github.com/BUTSpeechFIT/DiaPer
[10] https://github.com/BUTSpeechFIT/diacorrect

chunks. EEND-VC generates thus multiple streams of embeddings, one for each speaker in a chunk. We can cluster these embeddings using constrained agglomerative hierarchical clustering (cAHC), ensuring embeddings from the same chunk belong to different clusters. In [Del+23] we introduced an alternative clustering approach, a multi-stream extension of the successful Bayesian HMM clustering of x-vectors (VBx), called MS-VBx. Experiments on three datasets demonstrated that MS-VBx outperforms cAHC in diarization and speaker counting performance.

2.4 Language and emotion recognition

Language identification (LID) and speech emotion recognition (SER) tasks were not in the corner-stone of NEUREM3 research, but they enormously benefited from our speaker recognition and speaker diarization research presented above. Especially multi-head factorized attentive pooling (MHFA) developed in the context of speaker recognition [Pen+23a] was shown to bring strong results in LID and SER.

2.4.1 Emotion recognition building on speaker embeddings

Speech emotion recognition (SER) is a challenging task in human-computer interaction. In a series of two papers written with our Greek partners [Kak+23; Sta+23], we presented a novel approach to SER using self-supervised learning (SSL) models and attentive correlation pooling with label smoothing. The method addresses three main challenges in SER: (i) Capturing emotion-relevant information from speech, (ii) Modeling long temporal context of emotions, and (iii) Dealing with noisy and ambiguous emotion labels.

The proposed method leverages pre-trained SSL models like HuBERT, Wav2vec 2.0, and WavLM as feature extractors. Inspired by our previous work in speaker verification, we introduce the use of an attentive correlation pooling method for SER. This method combines channel-wise correlations with an attention mechanism to better capture emotion-relevant information across time. Label smoothing is applied to address the issue of noisy labels.

In our experiments, we keep the SSL models frozen and follow a SUPERB style evaluation with a light classification head on top of these models, which leads to good computational efficiency. Results on the IEMOCAP dataset show state-of-the-art performance, with the best result of 75.60% unweighted accuracy using WavLM. This outperforms previous approaches using both frozen and fine-tuned SSL models.

2.4.2 Language recognition evaluation systems

In 2022, we participated in the latest edition of the NIST language recognition evaluation (LRE). The particular task of NIST LRE 2022 consisted of classifying audio recordings into 14 language classes. The focus was mainly on low-resource languages from Africa. The evaluation had two tracks: the Fixed condition, where participants are required to use only predefined data sets, and the Open condition allowing to use any data to train the systems. Our system description [Sil+23a] summarizes the main components of our submission for both evaluation conditions and motivates for the particular design choices that were made during the development of our solution.

Paper [Mat+23] is an extended version of the system description of our submission for NIST LRE 2022. First, we introduce the submitted system: we describe what data we have used to train our models, and what embedding extractors and backend classifiers were used. After covering the architecture, we concentrate on post-evaluation analysis. In particular, we compare different topologies of DNNs, different backend classifiers, and the impact of the data used to train them. We also report results with XLS-R pre-trained models that could have been used in the Open condition. The results indicate that the performance on language recognition task can largely benefit from both using the additional data and the pre-trained model.

Similarly to other "challenge sections" in this chapter, a broader view on NIST LRE 2022 can be found in Section 5.5.

2.5 Automatic speech recognition

2.5.1 Spelling-aware ASR

Based on the prior work on speech-machine translation conducted by NEUREM3 team-member Hari Vydana [Vyd+21b], we proposed an end-to-end architecture for automatic speech recognition that expands the listen, attend and spell (LAS) paradigm [Ego+22]. While the main word-predicting network is trained to predict words, the secondary, speller network, is optimized to predict word spellings from inner representations of the main network (e.g. word embeddings or context vectors from the attention module). We have shown that this joint training improves the word error rate of a word-based system and enables solving additional tasks, such as out-of-vocabulary word detection and recovery. The tests were conducted on LibriSpeech dataset consisting of 1000h of read speech.

This journal paper contributed to the successfully defended PhD thesis by Ekaterina Egorova, see Section 5.4.

2.5.2 Adapting large pre-trained models for ASR of under-resourced scenarios

Dysarthric speech recognition has posed major challenges due to lack of training data and heavy mismatch in speaker characteristics. Recent ASR systems have benefited from readily available pretrained models such as wav2vec2 to improve the recognition performance. Speaker adaptation using fMLLR and x-vectors have provided major gains for dysarthric speech with very little adaptation data. In [Bas+22], we did one of early efforts to adapt large SSL models for a low-resource task. We proposed a simple adaptation network for fine-tuning wav2vec2 using fMLLR features. The adaptation network is also flexible to handle other speaker adaptive features such as x-vectors. Experimental analysis show steady improvements using our proposed approach across all impairment severity levels and attains 57.72% WER for high severity in UASpeech dataset. We also performed experiments on German dataset to substantiate the consistency of our proposed approach across diverse domains. The work was performed with colleagues from the Quality and Usability Labs of TU Berlin.

This and previous work of M.K. Baskar (a NEUREM3 team member in 2019 and 2020) (such as [Bas+21]) led to his successfully defended PhD thesis, see Section 5.4.

2.5.3 Fusion of end-to-end ASR systems

End-to-end systems have recently gained wide popularity in automatic speech recognition. However, these systems do generally not provide well-calibrated word-level confidences. To address this issue, we proposed Hystoc, a simple method for obtaining word-level confidences from hypothesis-level scores [BKB24]. Hystoc is an iterative alignment procedure which turns hypotheses from an n-best output of the ASR system into a confusion network. Eventually, word-level confidences are obtained as posterior probabilities in the individual bins of the confusion network. We showed that Hystoc provides confidences that correlate well with the accuracy of the ASR hypothesis, showing that the final level output representation of various ASR systems actually carries this information in fairly accessible form. Furthermore, we show that utilizing Hystoc in fusion of multiple end-to-end ASR systems increases the gains from the fusion by up to 1 % WER absolute on Spanish RTVE2020 dataset. Finally, we experimented with using Hystoc for direct fusion of n-best outputs from multiple systems, but the results in this case were mixed.

2.5.4 ASR evaluation systems / shared task systems

In [Koc+22b], we presented the collaborative efforts of BUT and Telefónica Research in developing Spanish ASR systems for the Albayzin 2022 Challenge. We train and evaluate both hybrid systems and those based on end-to-end models. We also investigate the use of self-supervised learning speech representations from pre-trained models and their impact on ASR performance (as opposed to training models directly from scratch). Additionally, we also apply the Whisper model in a zero-shot fashion, post-processing its output to fit the required transcription format. On top of tuning the model architectures and overall training schemes, we improve the robustness of our models by augmenting the training data with noises extracted from the target domain. Moreover, we apply rescoring with an external LM on top of N-best hypotheses to adjust each sentence score and pick the single best hypothesis. All these efforts lead to a significant WER reduction. Our single best system and the fusion of selected systems achieved 16.3 % and 13.7 % WER respectively on RTVE2020 test partition, i.e. the official evaluation partition from the previous Albayzin challenge.

A note on Albayzin 2022 Challenge is also included in Section 5.5.

2.6 Automatic speech unit discovery and keyword search

2.6.1 Resources and benchmarks for keyword spotting

Keyword search (KWS) is the task of automatically searching, detecting, and retrieving a set of user-defined keywords (usually in text form) from a spoken audio corpus. The technologies developed for KWS have various applications including, but not limited to indexing, searching in multimedia archives, video lectures, or voice based human-computer interfaces.

Even though several efforts have been made for KWS in low-resource languages (e.g. IARPA Babel program), there is a lack of a carefully crafted set of keywords and benchmarks for datasets that are freely available to the public. Moreover there exists no recipe or guidelines on how to create a suitable set of candidate keywords from the existing speech resources.

Bearing these limitations in mind, we published [Nad+22], where we presented the procedure of creating a suitable set of candidate keywords for keyword search from spoken audio, while using existing freely available speech datasets. We re-processed Microsoft Indic speech dataset 2018 and Multilingual and code-switching 2021 datasets comprising six low-resource Indian languages (Gujarati, Hindi, Marathi, Odia, Tamil, Telugu); and created keywords by taking into account the three properties (keyword occurrence, length and confusability distance), and their effect on the term-weighted value (TWV), the KWS evaluation metric. We trained four ASR-based KWS systems to benchmark KWS results on the created resources. We provided an in-depth analysis of keyword properties and their effect on TWV across all the languages and KWS systems. The keywords and other necessary resources (including Kaldi recipes) for replicating our experiments are made public.

2.6.2 Bayesian models for acoustic unit discovery

Acoustic unit discovery (AUD) entails discovering discrete phone-like representations from untranscribed speech, with the goal of making language processing systems such as translation and keyword spotting—which typically require discretized speech representations as from speech recognition—viable for the vast majority of languages in the world that lack annotated resources typical of modern speech processing systems.

We published a journal paper [Ond+22] consolidating our work on nonparametric AUD models. The models are a pair of AUD models which were introduced in the interim report, the Subspace Hidden Markov Model (SHMM) and the Hierarchical Subspace Hidden Markov Model (H-SHMM). Both are generative models which assume the speech in a language is generated from a number of discrete units each of which can be modeled by a Hidden Markov Model (HMM), and reduce the AUD task into one of inferring the number of units required to model the speech in a particular language along with the parameters of each unit. The SHMM constrains the parameters of each unit to dwell in a low-dimensional subspace spanned by the columns of a low-rank matrix whose parameters are learned from other languages with phonetically-transcribed data, thereby ensuring that parameters of each unit learned from untranscribed data define phone-like units. The H-SHMM allows the subspace parameters to be adapted for each language without completely removing the constraint on the parameters. We achieve this by introducing a second hierarchy of subspaces that constrains the parameters of the subspaces for each language.

In [Ond+22], we provided hitherto unpublished details of the model definitions and derivations of the inference schemes and compared with various cross-lingual and neural-network-based. Furthermore, we experimented with utilizing features obtained from large-scale self-supervised transformers instead of the prior spectral representations in our model, showing a symbiosis between the two: our approach, based on Gaussian mixture models, benefits significantly from having better input representations, while the transformer features benefit from the structure provided by our constrained models which significantly outperform across languages the various flavors of k-means which are typically used in speech discretization literature.

The SHMM was one of the key contributions of Lucas Ondel's Ph.D. thesis, alongside the variational Bayesian HMM approach to AUD and its extension into phonotactic language models (see also Section 5.4).

2.6.3 End-to-end architectures for keyword spotting

Conventionally, keyword search involves transcribing the speech using an ASR system into a graph of hypotheses (called a lattice) and text-based on the resulting graph.

In [YČS23], we proposed a novel end-to-end neural-network-based KWS paradigm that allows us to avoid the complexity, potential error propagation, and costly decoding and search algorithms in ASR-based systems. In our method, we replace the graph decoding and symbol matching of ASR-based systems with a simpler search method based on dot products on the outputs of a pair of encoding neural networks. Documents are transformed into sequences in the output space of one encoder, and queries are projected into the same vector space by another encoder. The search is then conducted by comparing, using dot-products, the query vector with the vectors that represent frames of the document. Locations in the document with high dot product similarity to the query are returned as hits, and locations with low dot products are ignored as background. We also adopt a multilingual training strategy that increases the viability of the proposed framework for KWS in languages for which low amounts of training data are available. We show experimentally the utility of our approach across several languages as a replacement for or complement to ASR-based KWS, and we conduct thorough analyses of its various strengths and weaknesses.

This work formed the core of Bolaji Yusuf's recently defended Ph.D. thesis (see also Section 5.4). Furthermore, it has been deployed in software delivered to the Police of the Czech Republic as part of the ROZKAZ project supported by the Czech Ministry of the Interior.

2.7 Between speech and NLP

2.7.1 Historical document recognition

In [Kiš+22], we looked at several facets of document analysis beyond pure transcription. Specifically, we developed systems for identification of used font/script, likely location of origin as well as the year of production. Submitting our systems to the ICDAR 2021 Competition on Historical Document Classification, we secured the first place in all tasks.

We proposed a system for visual document classification fusing patch level and textline level approaches. Our experiments show that textline processing significantly outperforms patch processing in classification of font and script types and in dating. On the other hand, the patch level approach yields better localization results. Fusion of the two brought significant improvements only in localization, showing that different levels of representation are indeed beneficial, but when of them fits the task very well, the other can be omitted, e.g. when classifying the font, the system simply has to look at the letters while the overall layout of the page and possible artifacts in its backgrounds are largely irrelevant.

This work has combined our NEUREM3 efforts with experience in OCR and document analysis in general obtained in the NAKI PERO project (see also Section 3.5). It also contirbuted to Karel Benes's (NEUREM3 team member) PhD thesis submitted in the end of 2024, with expected defense in early spring 2025 (see also Section 4.3).

2.8 Speech translation

In this section, we outline our contributions to speech translation. First, we mention our contribution to the dataset for the task (Section 2.8.1). Then we discuss end-to-end approaches for off-line use (i.e. whole input is available upfront), see Section 2.8.2. In the area of *simultaneous* speech translation (SST), we worked on manual and automatic evaluation (Section 2.8.3) and decoding strategies (Section 2.8.4). Next, we explored long-form inputs for SST (Section 2.8.5) and handling multiple sources concurrently for a better robustness (Section 2.8.6). Finally, we mention the organization of IWSLT shared tasks in Section 2.8.7 but refer the reader to more details on these evaluation challenges to Section 5.5.

2.8.1 Data preparation

In order to support multilingual research (Task 1) as well as multimodal research (Task 2), we created Khan Academy Corpus [Ďur+24] by curating the collection of subtitles available on the Khan Academy web site. The corpus covers 137 languages in subtitle text, starting with speech in 29 source languages. The dataset can serve in creation or evaluation of multilingual speech recognition or translation systems, featuring a diverse set of subject domains.

We also experimented with creating synthetic data for speech translation using genre transfer [KB22a]. The idea is to see text-based parallel corpus as the seed data for speech translation systems and use text style transfer systems to modify the data to better reflect the spoken genre (think textbook vs. an informal chat on a given topic). While the results were not fully satisfactory, the idea is worth further exploration and intermediate outputs (a genre classifier and a set of rules to inject spoken features) are useful.

2.8.2 End-to-end spoken language translation

We revisited several strategies for improving low-resource speech translation [Kes+23b]. We primarily relied on transfer learning and a simple end-to-end architecture for direct speech-to-text translation. We combined recent findings from joint-training and decoding in ASR and direct speech translation techniques and studied them with-respect-to various initializations in low-resource scenarios. Our experiments reconfirmed prior works that target-language ASR acts as good initialization for downstream speech translation. In addition, we found that pre-trained multilingual ASR is a viable alternative and performs better than the monolingual ASR in a majority of the settings. We showed that joint training and decoding with CTC objective helps even in low-resource setups. Finally, with only 300 hours of pre-training, our approaches achieved 7.3 BLEU score on low-resource Tamasheq-French dataset, outperforming prior works from IWSLT 2022. These findings helped us in building speech-to-text translation system for Marathi - Hindi as part of IWSLT 2023 shared task.

2.8.3 Evaluation of simultaneous speech translation

In past years, the evaluation of simultaneous speech translation usually did not involve humans who would rate SST end-to-end, in simulated SST session. The end-to-end eval-

uation is necessary to capture all aspects of SST, including the limited time in which the users receive the outputs, the paralinguistics in original audio, the presentation options and user interface, and others. Therefore, we created Continuous Rating [JMB22], an evaluation framework in which human evaluators watch video with SST-generated subtitles, and continuously click rating buttons to express their rating. Our results show that the bilingual evaluators are capable to rate the translation quality reliably.

Later, Continuous Rating had been collected in the evaluation of the IWSLT 2022 Simultaneous Speech Translation shared task on one language pair. In [MBD23], we analyzed the question whether the standard machine translation metrics, such as COMET, BertScore, ChrF and BLEU, can reliably assess the translation quality of SST, although they were not designed for the simultaneous task. We report a high correlation of the metrics and Continuous Rating of bilingual evaluators, which can be used as evidence that MT metrics can be used safely for SST quality evaluation.

2.8.4 Simultaneous decoding for simultaneous speech translation

Simultaneous decoding is crucial in SST systems, as it determines how much of the input can be safely translated, balancing quality and latency.

In [Pol+23a], we proposed an online beam search for a vanilla Transformer and a blockwise architecture for simultaneous speech inputs [TKW21]. The blockwise architecture uses beam search with hypothesis reliability scoring to decide when to wait for more input before translating. However, it maintains multiple hypotheses until the full input is processed, preventing real-time incremental translation. Additionally, it lacks mechanisms to control the quality-latency tradeoff.

To address these issues, we developed an improved beam search algorithm. Like standard beam search, it expands multiple beams per step. To handle growing input, we heuristically detect hallucinations — cases where a beam has translated all available input — and stop such beams to prevent quality degradation. Once all beams finish, the one with the highest score is selected, and existing incremental policies (e.g., local agreement [Pol+22b]) are applied to manage the quality-latency tradeoff.

Experiments with models trained for online and offline translation on MuST-C show 0.6–3.6 BLEU gains without increasing latency or 0.8–1.4 second latency reductions without affecting quality.

In [Pol+23b], we continued our work on developing a decoding algorithm with an optimal quality-latency tradeoff and introduced a new policy leveraging auxiliary speech translation CTC. Using the CTC prefix probability, we derived a score for each beam that represents the likelihood of the beam covering the entire speech input. The likelihood score and a tunable threshold value control the quality-latency tradeoff. This approach allowed us to replace the hallucination detection and policy steps from the previous paper, resulting in significant improvements in quality (up to 1.1 BLEU) and computational efficiency (up to 45% reduction in relative RTF).

Additionally, this paper described our submission to the IWSLT 2023 Simultaneous Speech Translation Track [Aga+23]. Based on a human evaluation study, the submitted system placed first in English-to-German and second in English-to-Japanese.

Last but not least, we implemented Whisper-Streaming [MDB23], a tool for simultaneous decoding of Whisper-like speech-to-text translation and transcription models.

2.8.5 Long-form simultaneous speech translation

A key limitation of most end-to-end SST approaches in the current literature is the assumption that the source speech is pre-segmented into sentences, which poses a significant challenge for practical, real-world applications. In a thesis proposal [Pol23], we reviewed existing research and analyzed the main challenges in SST. Based on this analysis, we proposed a new direction for future exploration in the long-form regime.

2.8.6 Multi-source simultaneous speech translation

Machine speech translation can exploit multiple language sources, such as original and simultaneous interpreting in an intermediate language. We continue with research in this direction in [Mac+23] by analyzing multisourcing in simplified conditions: with optimally aligned and segmented text translations and with various levels of artificial ASR errors. Our results show that there is an area where multi-sourcing achieves a higher BLEU score, which motivates for further research.

We summarize our research in simultaneous speech translation from multiple sources and relevant underlying topics (background, data, SST evaluation, simultaneous ASR baseline) in the dissertation thesis [Mac24].

2.8.7 Shared task organization in speech translation

We contributed to the organization of the simultaneous speech translation shared task in 2022 [Ana+22], 2023 [Aga+23], and 2024 [Ahm+24]. Comprehensive details regarding the task setup and findings are provided in Section 5.5.2.

2.9 Neural machine translation

In the area of neural machine translation, we worked on evaluation, training from non-parallel data, inference and basic research, and we took part in several translation shared tasks. We detail our publications in these areas in the following.

2.9.1 Evaluating machine translation quality

We see evaluation as the driving force behind research in MT, as we described in the chapter "Evaluation Techniques, Experiment Design and Common Misconceptions in Machine Translation Research" (to appear as [Boj25]). Without sound evaluation, model deployment decisions can easily become misaligned, leading to suboptimal service. For definitive evaluation, human annotators are often hired to assess the quality of machine translation outputs. Who the specific annotator is, however, can play a big role.

In Section 2.12.3, we describe our methodology for a novel method for creating reference translations and evaluating the resulting reference quality. The same dataset collection also produced a set of diverse post-edited translations, also known as references. Human evaluation does not scale well, and oftentimes reference-based automated metrics are used to drive decisions. In [ZB24] we contrast the quality of the human reference, coming from multiple sources, against its usability for machine translation evaluation. Surprisingly, we find that on top of being the most expensive, the translations with the highest quality are not the ones that are most useful (leading to highest correlations

for automated metrics). This informs future practitioners, such as at the Conference of Machine Translation, to choose an adequate quality of references.

In [TTB23], we relate the power of machine translation (MT) systems and the power of quality estimation (QE) systems trained on their outputs. One would naturally expect that a QE system trained only on good MT outputs would miss the "proper scale" and would not be the best at assessing quality of MT outputs in general. This is indeed the case and bad MT systems are better source of synthetic data for QE training.

2.9.2 Training MT on non-parallel data

Unsupervised machine translation, i.e. methods for training translation systems on non-parallel data, are important for languages suffering from parallel data shortage, but they can also be important for less-resourced specific domains in well-covered language pairs.

In [KB23a], we evaluate a surprisingly powerful technique for unsupervised MT which combines parallel data discovery with unsupervised MT training. As a whole, the technique does not need parallel data, but the trick to create synthetic parallel corpus in a separate step significantly helps.

The main result of our work in this area is the encompassing look at unsupervised MT in the book [Kva25] based on Ivana Kvapilíková's dissertation thesis defended in 2024 [Kva24]. While updating the book contents to reflect large language models (LLMs), we realized that their "emergent ability" of translation can be explained by the similarity of their training data and training methods to methods successfully used in unsupervised MT.

Additionally, we explored unsupervised MT techniques in shared task submissions: In our shared task submission [Pož+22], we experiment with methods for extracting term counterparts (i.e. a translation dictionary) from comparable corpora, trying out static word embeddings, contextualized multilingual embeddings and also the full pipeline of unsupervised phrase-based machine translation, which produces a phrase translation dictionary as a side-effect. The best of our methods reach a precision of 88%.

In two shared task submissions on translation, we explore techniques of unsupervised MT for genuinely low-resource language pairs, namely translation from English to Ukrainian, Kazakh and Georgian [KB22b] and translation between English and Assamese, Khasi, Mizo and Manipuri [KB23b].

2.9.3 Multilinguality for better MT of low-resource languages

We further researched the use of character-level NMT for related, low-resource languages, started in the previous phase of the project. [Jon+21a].

In [JB23b] we investigated how character-level neural machine translation (NMT) with the Transformer architecture performs for pairs of languages at varying levels of similarity—namely Czech paired with Slovak, Croatian, Hungarian, German, and Spanish. Using different training data sizes and both subword- and character-level segmentations, we measured translation quality with standard automatic metrics. Our results show that for closely related languages (like Czech-Slovak), character-level segmentation can outperform subword-level segmentation. However, for less related pairs, subword segmentation usually yields better quality, especially when larger datasets are available.

We also confirmed that finetuning from a trained subword-level model to character-level can boost performance for certain language pairs — most notably Czech-Croatian — where training a purely character-level system from scratch was less effective. Additionally, we tested deeper Transformers and observed that character-level models do benefit from added model capacity, though in many cases the gains depended on having sufficient training data. Overall, our findings suggest that character-level processing can be advantageous for closely related languages or in scenarios where finetuning from a subword base model is feasible.

2.9.4 Negative lexical constraints for NMT

We also followed up on our work on lexically constraining translation systems [Jon+21c; BVB21].

In [Jon+23] we investigate how to prevent a neural machine translation (NMT) model from using specific "forbidden" words or phrases in its output, a technique known as negative lexical constraining. We compare a range of methods for enforcing these constraints, such as modifying the model's decoding process (e.g., beam filtering or score penalization) and adapting the training data with special markers. Our experiments address two main tasks: paraphrasing—where constraints help force the model to produce alternative word choices—and feedback-based translation refinement—where constraints reflect user feedback about incorrect tokens in the model's output.

While we show that prohibiting certain words or expressions is feasible, our findings also reveal limitations. In morphologically rich Czech, for example, the model often "evades" constraints by generating slightly altered forms (inflections) of the forbidden words. Training with stemmed constraints helps to mitigate this problem, but harms overall translation quality.

2.9.5 Length control in machine translation

Controlling the output length of machine translation (MT) is essential in scenarios such as dubbing, subtitling, and structured document translation with limited space constraints. We address this challenge from two perspectives: (1) identifying semantically important words to omit less meaningful content, and (2) developing end-to-end MT systems trained on the Paraphrase Database (PPDB) to produce shorter target sequences through paraphrasing.

Identifying important words is a key step in shortening MT outputs. We explored two approaches. In [JBY23], we proposed a method that uses a selected attribution technique to assign importance scores to each input token. The resulting scores exhibited desirable properties, including: (a) content words being rated as more important than function words; (b) scores reflecting contextual dependencies; and (c) minimal changes in sentence meaning when low-score words were removed. In a complementary study, [OJB24], we introduced a novel self-supervised task that modifies the masked language modeling objective. Instead of predicting masked words, our approach involves learning to identify which words were inserted into a sentence. We evaluated this method on a small-scale human-labeled dataset, finding its performance competitive with the attribution-based approach of [JBY23]. These importance scores serve as a foundational step for developing fully-fledged MT models optimized for shorter outputs.

End-to-end methods offer an alternative pathway to shortening MT outputs. In [Per+23], we explored this approach, leveraging the property that certain paraphrasing rules inherently produce shorter target sequences. By fine-tuning an MT model using paraphrasing rules that shorten outputs, we achieved a slight bias towards generating shorter translations and we leave room for improvement in future work.

2.9.6 MT decoding using a genetic algorithm

In [JB23a], we present a novel use of genetic algorithm (GA) for manipulating n-best translation hypotheses. Starting with a set of candidate outputs from a machine translation (MT) model, we apply standard GA operations–mutation (token insertions, deletions, or replacements) and crossover (swapping portions of sentences)–to optimize translation quality under a chosen "fitness function" (an MT evaluation metric). By combining multiple metrics (e.g., COMET-based measures and traditional lexical metrics like BLEU/ChrF), we show that GA-based optimization can yield higher-quality translations than simple reranking alone, as demonstrated by improvements in scores on held-out metrics.

We also show how the same approach can reveal weaknesses or "blind spots" in automatic MT metrics themselves [JB24]. When using a single metric as the fitness function, the GA is effectively able to overfit that metric, producing adversarial examples with high scores in the target metric but poor quality as judged by other held-out metrics. We illustrate that reference-free COMET, for instance, is especially vulnerable to certain artificially constructed outputs (e.g., hallucinated details, changes in named entities or numbers). Our study indicates that GA offer a promising new angle both for improving MT outputs via consensus across multiple metrics and for stress-testing the robustness of individual metrics.

We further used this approach in our WMT 23 submission [JPB23]. We applied the GA on n-best lists produced by the English-Czech NMT models from the previous years of the competition. The resulting translations ranked the first in the constrained category, outperforming the translations from the base system without the use of the GA.

2.9.7 Non-Autoregressive NMT

We continued our research of non-autoregressive models. These models promise fast sequential decoding because they impose a conditional independence assumption on the output tokens, allowing the use of parallel decoding. In [HHB22] we develop non-autoregressive machine translation models and compare them with the state-of-the-art models optimized for inference speed using conventional methods such as knowledge distillation, model pruning and quantization. We find that at the time that under realistic conditions, autoregressive models are superior to non-autoregressive models, both in terms of translation quality and speed. The publication is intended as a call for more realistic and extensive evaluation of non-autoregressive models in future work.

The research in the aforementioned paper formed a chapter in Jindřich Helcl's PhD thesis defended earlier in 2022 [Hel22]. The thesis features an extensive overview of papers published about non-autoregressive MT and highlights selected flaws in evaluation methodology present in almost all of them, giving grounds to call for fair evaluation.

2.9.8 On difficulties of attention regularization

As part of our basic research on the Transformer architecture, we examined long input handling. Previous literature has suggested variations of the Transformer model that added shared memory to the Transformer, aiming to reduce the memory complexity of the attention operation from $O(n^2)$ to linear time. We however discovered that the size of this memory component does not affect the performance of the model, and specifically, that all the memory cells converge to the same values. This surprising finding was presented in two student research / tiny papers at ECAI and ICRL [YBH23; Yor+24] and serves as the motivation for our further research on efficient long-form Transformers.

2.9.9 Limitations of Transformer optimization in NMT

Following up on work described in the interim report, namely Sections 2.5.7 and 2.5.8, we further explored the limits of the contemporary Transformer architecture and its optimization process. [Var23] provides an in-depth description of our experiments and their results. The described research was split into three sections each investigating Transformer generalization capability, Transformer issues with catastrophic forgetting and options for smarter Transformer modularization, respectively.

Generalization abilities The experiments studying Transformer generalization expanded on the results described in [VB21]. Besides the experiments studying the length-generalization ability of Transformers, we also investigated their ability to transcribe previously unseen terms (e.g. named entities not seen during training) and how the train-test dataset vocabulary overlap skews the translation evaluation. The former experiments showed that Transformers still have issues with input copying (a common fallback method when translating unseen words) and, based on the results in the latter experiments, we were unable to conclude whether a significant train-test vocabulary distribution overlap negatively affects translation evaluation.

We also attempted to propose a methodology for studying generalization abilities of the Transformer in [TSB23]. The underlying experiment focuses on compositional expressions such as "green leaves". The idea was to downsample the training data so that the whole phrase is never available in any training item but its components ("green" and "leaves") are. The Transformer model should be able to compose these components and correctly translate even the unseen phrase. The experiment confirmed that Transformer exhibits a reasonable generalization power in this respect but a pressing issue remains. Some phases (e.g. "United States"), even if compositional, are too domain specific. Filtering out sentences with these phrases thus damages the performance for all topics in the domain (politics), so automatic assessment of this ability is not possible. Note also our more general work on generalization of compositional structures mentioned in Section 2.11.2 below.

Catastrophic forgetting Furthermore, we followed up the experiments described in [VB19a]. We introduced modifications to the existing Elastic Weight Consolidation (EWC) regularization, proposing new regularizations to the Fisher Information matrices used to determine the importance of the individual parameters with respect to the original task. We expanded the previous evaluation by validating the regularization method

on simple string editing tasks and machine translation. We concluded that while the EWC regularization can help tackling the catastrophic forgetting, it does not remove it completely.

Modularization Lastly, we proposed a novel method of modularization of the transformer multi-head attention and feed-forward blocks and self-learned selection of subset of these blocks given a specific task input. Such module allocation should, in theory, help with better allocation of Transformer parameters between multiple tasks during training. The experiment results showed promising results when applying module selection solely to the feed-forward blocks, similar to the results reported by the previous research on adapters in Transformer networks.

2.10 Language and Vision

2.10.1 Datasets for multimodal research

We contributed to the survey of multimodal datasets [Gar+22].

Continuing our work from the first half of the project [PBD19], we prepared three more datasets for multimodal translation from English into low-resourced languages (Bengali [Sen+21] and Hausa [Abd+22]), i.e. translation with the support of a related image. For Hausa, we also created a dataset for visual question answering [Par+23]. These datasets serve in the WAT multimodal tasks and are also a basis of evaluation of multimodal abilities of future models.

2.10.2 Position information in images

In [RL22], we examine the level to which models of blending language and visual information (vision-language models, VL) are capable of understanding the position information of image parts. While probing experiments document that the position information is available in the representation and this can be improved with additional training strategies (including the addition of the "depth" axis), the studied model LXMERT is not capable of utilizing the position information in image-text matching task. Our results thus highlight an important issue of multimodal modeling: the mere presence of information detectable by a probing classifier is not a guarantee that the information is available in a cross-modal setup.

2.10.3 Visual Question Answering

Within our research on pre-trained encoders combining language and vision, we studied the effect of positional information about objects in the images on the downstream performance. In our case study of LXMERT [RL22], at that time a state-of-the-art VL model, we probe the use of the PI in the representation and study its effect on Visual Question Answering. We show that the model that can use the position information for the image-text matching task on a challenge set where only position differs. Our probing experiments confirm that the position information is indeed present in the representation. We introduce two strategies to tackle this: (i) Positional Information Pre-training and (ii) Contrastive Learning on position information using Cross-Modality Matching.

By doing so, the model can correctly classify if images with detailed position information statements match. Additionally to the 2D information from bounding boxes, we introduce the object's depth as a new feature for better object localization in the space. Even though we were able to improve the model properties as defined by our probes, it only has a negligible effect on the downstream performance. Our results thus highlight an important issue of multimodal modeling: the mere presence of information detectable by a probing classifier is not a guarantee that the information is available in a cross-modal setup.

2.10.4 Multimodal Summarization

In the area of Multimodal Summarization, we continued our work by developing an innovative metric [KP22] aimed at evaluating the quality of textual summaries. This metric was trained on historical data to closely emulate human annotators' scoring patterns. In the realm of Multimodal Summarization, we identified several key limitations in prior research, one of the most significant being the absence of publicly available datasets and baseline models, which posed a challenge for new researchers entering the field. To bridge this gap, we curated and released a unique dataset [KP23a] compiled from various news sources, making it accessible to the scientific community. In our related study [KP23b], we investigated the benefits of pre-training on the simpler task of text-only summarization and introduced a novel loss function that incorporates inner-video similarity across consecutive frames. Additionally, we developed a framework for collecting human annotations to evaluate the adequacy and effectiveness of pictorial summaries.

Building on this work, our follow-up research [KP24] uncovered another drawback in commonly adopted methodologies: highly intricate model architectures often limit the adaptability of models to specific data modalities. To address this, we proposed a unified approach to Multimodal Summarization that enables a single sequence-to-sequence model to learn jointly across multiple modalities (text+video→text+image, text+images→text+image, text→text) using a multi-task training paradigm. Our findings demonstrated that this strategy serves as a viable alternative to text-only pre-training while retaining the model's inherent capability to process purely textual inputs. Moreover, we delved into generative modeling in the visual domain, presenting a pipeline solution leveraging Stable Diffusion to enhance multimodal content generation.

2.10.5 Handwritten music recognition

We continued our research focusing on the visual modality of interpreted data in the task of Handwritten Music Recognition (HMR). In [May+24], we explain our motivations for building a dedicated handwritten music notation synthesizer that was used in [MP21] and the possibility of extending the synthesizer to polyphonic music and adding support for more musical concepts discussed in [MP22]. We compare this approach to alternatives, such as modifying existing music notation engravers and base our decision on the visual broadness and variability of handwritten music. We outline the top-level architecture of such a music notation synthesizer and list all the pain points and risks to be aware of when attempting such synthesis. This work was presented at the annual WoRMS workshop, attended by most of the optical music recognition researchers from Europe

and the western world. The paper served as a guideline for the architecture of the next-generation synthesizer Smashcima, developed at the end of 2024 [11].

2.11 Towards semantics

Our work towards handling the meaning of the text can be organized into two broad areas: On the one hand, we analyze the continuous representations learned by the models at the sentence level and relate them to annotated features of semantics (Sections 2.11.1 and 2.11.2), and on the other hand, we focus on longer inputs and their summarization in our meeting summarization task AutoMin, for which certain necessary research steps had to be taken: creating an annotation tool and preparing a training and evaluation dataset (Section 2.11.3). Finally, our AutoMin shared task submission is also reported on in this section.

2.11.1 Analyzing knowledge of semantic relations

In [HM21], we analyze BERT, from a lexical semantic perspective. We focus particularly on its knowledge of hypernymy, the relationship between a noun (e.g. "car") and its superordinate category ("vehicle"). Encouragingly, we find that BERT is able to retrieve noun hypernyms much better than prior systems, even in challenging scenarios where a word might not have had a clear hypernym. However, BERT achieved peak performance when retrieving the hypernym of frequently-encountered nouns, and struggled with more abstract hypernyms.

Building upon our COSTRA dataset of sentence transformations created in the first half of the project [BB20a], we examined the popular SBERT and LaBSE sentence embeddings, evaluating to what extent they reflect 13 semantic attributes exemplified in COSTRA. Some features, such as sentence tense representation can be extracted from the embeddings but features of meaning beyond the level of individual words remains difficult for these models. See [Zha+24] for more details.

In [BZB22] we study where (in terms of layer depth) and how well large pre-trained models like BERT store linguistic traits of sentences, namely sentence ambiguity, grammaticality and sentence complexity. We note that some of the existing datasets which mark these phenomena, do so using too overt features, i.e. the feature is apparent already from the presence of a trigger word in the sentence. We observe sentence complexity to be much more accessible than sentence ambiguity, indicating that the models are not aware of sentence ambiguity much. Finally, we study the popular t-SNE projection which is often used in this sequence of implications: If we see a clear clustering in t-SNE projection, a classifier for this feature will have a high accuracy. If a classifier for a feature has a high accuracy, the feature is present in the representation. The problem is that the lack of clear cluster separation in t-SNE cannot be used as evidence for low representation of that feature.

2.11.2 Compositionality in sequence-to-sequence models

In [AP22], we focus on recurrent encoder-decoder architectures with discrete bottleneck. In the research area of emergent languages, it has been disputed whether producing

[11] https://github.com/OMR-Research/Smashcima

structured (compositional) representations leads to better generalization performance. We showed that compositionality is indeed a necessary condition for better generalization provided the generalization problem is hard enough, such as when the test data come from a different data-generating distribution. Our results were later corroborated by other research.

2.11.3 AutoMin – Automatic meeting summarization

Thanks to NEUREM3 support, we were able to run two instances of our shared task on automatic meeting summarization, AutoMin 2021 and 2023.

In [Pol+22a], we introduce our annotation tool designed for meeting annotation, alignment, and evaluation. The tool features an efficient and user-friendly interface that facilitates fast annotation of multi-party meetings while reducing the likelihood of errors. The tool was used to create our minuting corpus [Ned+22]. Additionally, the tool includes an evaluation mode that supports a comprehensive quality evaluation of meeting minutes. This evaluation mode was utilized for the manual evaluation of the AutoMin 2021 [Gho+21] and AutoMin 2023 [Gho+23] shared tasks. The tool is available both as source code[12] and as a Python package.[13]

Our master students Kristýna Klesnilová and Michelle Elizabeth also took part in the AutoMin shared task [KE23] with BART-based models.

2.12 Human performance and human factors

2.12.1 Comparison of human and machine text translation

In [NB23], we carry out a targeted study on the particular phenomenon of German compounds in English-to-German translation. While this language pair has reached overall translation quality comparable to humans, we discover that the number of compounds produced by even the best NMT systems is much lower than produced by human translators. The reason for this difference has yet to be explained in full (we suspect it is related to the beam search and different inference algorithms may help), but it clearly documents a remaining substantial qualitative difference between human and machine output and offers a complementary translation quality measure.

2.12.2 Comparison of human and machine performance in simultaneous interpreting

In [Klo+24], we analyze translation departures in human simultaneous interpreting (professional and trainee interpreters) and automatic simultaneous speech translation systems. Our findings reveal that trainee interpreters outperformed the professional one in multiple areas, such as avoiding over-reliance on original structure or excessive use of demonstrative pronouns. In comparison, the automatic simultaneous speech translation model translated more literally, with outputs having fewer translation departures than human interpretations (the processing capacity is technically higher than that of humans – given the limited task of translating words, not delving into the actual meaning of the

[12] https://github.com/ELITR/alignmeet
[13] https://pypi.org/project/alignmeet/

message). However, the automatic model struggles with difficult sound, pronunciation or syntactic conditions, leading to unnatural errors.

2.12.3 Optimal reference translations

In [Klo+23], we propose a novel methodology for constructing high-quality reference translations, called "optimal reference translation", in an attempt to evaluate generally very good translation systems. In the paper, we explore the quality of translation by analyzing two different Czech translations of English newspaper articles, done by professionals hired by a translation company. The primary focus is on the formal aspect of the existing translations and the phenomenon known as translationese. The problems identified in the Czech translations fall into a wide range of linguistic categories such as spelling, morphosyntax, grammar, lexicon, and word formation. Special attention is paid to the presence of source-language interference; having reviewed existing theoretical discussions of interference, the authors drafted a typology which was then expanded to include several other types of errors recurrent in the translations analyzed.

In [Zou+24] we evaluate the quality of our optimal reference translations in order to validate the approach. We collect a novel dataset of human annotations of machine translation outputs with annotators ranging from students of translation studies to professional translators. We find that the perception of translation quality largely depends on the professional expertise of the given annotator. The lower the expertise, the fewer problems are spotted in the translation and vice versa.

2.12.4 Towards relating human and machine processing of languages

In [WZB24], we build upon our EMMT dataset developed earlier in the project [Bha+22a] and containing eye-tracking recordings where human subjects are translating English sentences to Czech with or without a supporting image. In the study, we compare human reading times (fixation durations) and model surprisal values when processing the source sentences and find out that ambiguous sentences indeed pose a slightly bigger challenge to both. In a separate part of the analysis, we observe that text-only machine translation outputs are more similar to human translations produced without the visual information compared to the case when the human saw a related image. We also found that a context-relevant image has a significant influence on the decision of the human translator to update the translation.

In line with the goals of our Tasks 2 (multimodality) and 8 (human processing), we proposed a multimodal variant of the "Shannon Game", an interactive exercise in which the subject (human or machine) predicts the next word of a sentence. The multimodality lies in priming human respondents with variants of visual information associated with the sentence and prompting a (text-only) LLM with a textual form of this information, names of objects appearing in the picture. We relate human and machine processing in terms of confidence by (1) asking people to provide their confidence when making the prediction, (2) recording the model score associated with the highest-scoring prediction. Similarly, we relate humans and machines in terms of accuracy: self-reported by humans when seeing the true word and measured as an exact match with the true word for machines. We observed that the presence of any visual information positively influenced the confidence and accuracy of the next-word prediction, with the full image configuration

yielding the best results. GPT-2 also benefited from incorporating an additional modality, although with more variability. The correlation of these metrics between the human and the machine decreased with the additional modality, suggesting differences in how humans and the model process visual inputs. The work was presented in [ZBB24].

In [BKB22], we experimented with "predictive" style of modeling human processing of language. The paper describes our submission to the CMCL 2022 Shared Task of Multilingual and Cross-lingual prediction of human reading behavior, where the goal was to predict Total Reading Time and First Fixation Duration and their respective standard deviations. We used BERT and XLM, observing that XLM delivers a better performance and proposing several tricks to improve the prediction (e.g. including more context, including derived linguistically motivated features or choosing the best pooling method to extract the information).

3 Foreign cooperation

Among the Czech research teams, the BUT SpeechFIT group at BUT and the Institute of Formal and Applied Linguistics at CUNI are probably the most international, both from the point of view of its staff and PhD students, and international relations. In this section, we include the facts rather than simply stating "we cooperate with Prof. xx from University yy". Please, note also the number of foreign co-authors in NEUREM3-sponsored publications.

3.1 Hosting foreign students and co-workers

This list covers only staff and students funded by NEUREM3 during the whole lifetime of the project (i.e. covering years 2019-2023). Please see Section 4 to see the details on their track record, activity on NEUREM3, etc.

- Dr. Hari Krishna Vydana, senior, India
- Dr. Santosh Kesiraju, senior, India
- Dr. Johan Rohdin, senior, Sweden
- Anna Silnova, PhD student and later senior, Russia
- Shuai Wang, PhD student, China
- Murali Karthick Baskar, PhD student, India
- Junyi Peng, PhD student, China

3.2 Self-funded co-workers and visitors

We are regularly receiving colleagues and students coming with their own funds. The benefit for the project is obvious – we obtain quality research results and can save project's budget for those that do not have personal funding. During NEUREM3, we hosted:

- Dr. Alicia Lozano[14] (UAM Madrid, Audias group) spent almost two years at BUT, first funded by the H2020 Marie Curie project No. 843627, then by UAM-Santander post-doctoral grant. She significantly contributed to our efforts in end-to-end speaker recognition and diarization and keeps working with us even while back in Spain [Ala+19; Lan+21b; Ala+20; Bur+20; Loz+20; Lan+22b; Lan+23].
- Ebrahim Ansari (Assistant Professor at Department of Computer Science and Information Technology Institute for Advanced Studies in Basic Sciences (IASBS) Zanjan, Iran) was a post-doc at CUNI and then spent further year funded by our EU project ELITR. Ebrahim worked with us primarily on the evaluation of spoken language translation and also supervised a student of his, Mohammad Mahmoudi on the topic [Ans+21; Ans+20]. He also created a dataset for processing colloquial Persian as a low-resource language [KAB20a].

[14] http://audias.ii.uam.es/staff/

- Shantipriya Parida (Researcher at Idiap Research Institute, Martigny, Switzerland) spent his post-doc in 2018 at CUNI under the supervision of Ondřej Bojar. During that time, we developed Hindi Visual Genome which was later finalized and published within NEUREM3 [PBD19], our first dataset for multi-modal translation. During NEUREM3, our collaboration continued. Jointly, we also organized shared tasks on multi-modal translation [Nak+19; Nak+20; Nak+21] and contributed our systems [PMB19a; Par+20; Par+21].
- Prof. S Umesh[15] (Indian Institute of Technology Madras, India) is the esteemed leader of Indian speech and language processing activities. He visited BUT in 2022 and 2023, for 2 months in total, supported by project "Multilinguality in speech technologies" (joint funding of the Czech Ministry of Education and Indian DSF). Besides bringing his significant expertise to the NEUREM3 team, he also brought his daughter, Jahnavi Umesh, that was instrumental in the development of ASR system for the Albayzin 2022 challenge (see Section 5.5.4 and [Koc+22b]).
- Lin Zhang from Hitosubashi University (Japan) came in 2023 for a PhD intern funded by Japanese Government. She continued her PhD work on spoofing diarization but also contributed the NEUREM3 team in several speaker recognition activities [Pen+22b]. Lin returned to BUT in 2024 as a post-doc.

3.3 Organization and participation in international workshops, challenges and evaluations

This important international activity is covered in Section 5.

3.4 Short-term internships

Short-term internships in excellent academic and industrial laboratories are important for professional and personal development of our PhD students. We encourage these internships and actively help our students to get to the best possible teams. The list also covers the whole lifetime of the project (i.e. 2019-2023), please mind the significant gap in these activities in the COVID years (2020–2021):

- From March till June of 2019, Anna Silnova was at the internship in STAR lab of Stanford Research Institute (SRI). There, Anna was working on the robustness of speaker verification back-end models to language mismatch between training and test data as well as between enrollment and test speech segments.
- From October 2018 till April 2019, Karel Beneš was at an internship at RWTH Aachen, at the department of prof. Hermann Ney. He was working on improving ASR by means of self-adaptation with low-dimensional topic-summarizing embeddings of transcriptions.
- In the autumn of 2019, Ivana Kvapilíková did a 2-months research stay at the University of the Basque Country where she co-operated with the local NLP team on a

[15] https://www.ee.iitm.ac.in/~umeshs/

project investigating unsupervised learning of sentence representations. The findings of the project were published at the Student Research Workshop of the ACL 2020 (see [Kva+20]).

- From June till October 2021, Murali Karthick Baskar was an intern with Google, Inc. He worked on Masked Speech Models (MSM) such as wav2vec2 or w2v-BERT. This work is named as Ask2mask (ATM) as it masks only high confident samples instead of random masking during MSM pre-training.

- From December 2023 till February 2024, Jan Brukner interned at University of Stuttgart (Germany). He worked on speaker anonymization, focusing on low real-time-factor and regularization by prosody cloning. He collaborated with with Sarina Meyer and was supervised by Ngoc Thang Vu.

- From September till October 2023, Ladislav Mošner interned at LORIA, INRIA (France) to Work on extending SSL pre-trained models (such as WavLM) to their multi-channel versions in the context of multi-channel speaker verification. He collaborated with with Romain Serizel and Emmanuel Vincent. The stay was co-funded by the French Embassy in the Czech Republic as part of the Joseph Fourier prize Ladislav won, and led to joint publication [Moš+24].

- In July - August 2023, Junyi Peng spent summer intern at NTT (Japan). He worked target speech extraction with Marc Delcroix, Tsubasa Ochiai, Shoko Araki and Takanori Ashihara. This work led to two papers [Pen+24b; Pen+24a].

- Dominik Macháček spent 3 months in NICT in Japan in 9-12/2022. The stay was funded by NICT and lead to a fruitful collaboration afterwards, primarily the remote collaboration with Raj Dabre between 2/2022 and 5/2024, leading to the joint publications [MBD23; Mac+23; MDB23].

- From January to July 2022, Peter Polák visited Karlsruhe Institute of Technology (Germany), where he worked together with professors Alex Waibel and Jan Niehues on simultaneous speech translation. This cooperation led to two first places in the Simultaneous Speech Translation Track in IWSLT 2022 and IWSLT 2023 [Pol+22b; Pol+23b].

  From October 2022 to March 2023, Peter Polák also visited Carnegie Mellon University (USA), where he worked together with professors Alex Waibel and Shinji Watanabe. This cooperation let to two joint publications [Pol+23a; Yan+23].

- From March 2023 to August 2023, Dávid Javorský interned the University of Paris-Saclay in France. He worked on methods for shortening machine translation under the supervision of François Yvon. This collaboration resulted in a joint publication on deriving word importance scores from models trained on semantic tasks [JBY23]. The collaboration continues to date, with one more internship from October 2024 till December 2024 and further publications in the pipeline.

- From March to May 2023, Josef Jon stayed at Google Research in Berlin supervised by Felix Stahlberg and Shankar Kumar, working on alternative training objectives for machine translation. The work on publishing the outcomes is in progress and will serve as a chapter in Josef's PhD thesis.

- In 2023, Věra Kloudová stayed at the Leipzig University, to present the newly published [Klo+23] to the students and colleagues of translation studies branch and to discuss and rework the chapter [Klo+24].

NEUREM3 students were actively interning in our partner institutions also in 2024, Junyi Peng has notably visited NTT Japan again (February - April) and Carnegie Mellon University (Watanabe lab) from September till December.

3.5 Synergetic international and national projects

NEUREM3 does not exist in isolation — both institutions are involved in a number of international and local projects that are interacting with NEUREM3 and exploiting its results mostly in more application-oriented scenarios, In addition to IARPA MATERIAL, Horizon 2020 ROXANNE, Horizon 2020 ATCO, Horizon 2020 HAWAAII, Horizon 2020 ELITR and Horizon 2020 WELCOME, CELSA CELL and Czech MoI ROZKAZ projects, already reported in the interim report, the team members were and are involved in the following international and national projects:

- "Language memory of the regions of the Czech Republic. Machine learning methods for preservation, documentation and presentation of the dialects of the Czech language" (JARIN)[16] aims to adapt existing language technologies and develop new procedures for automatic processing, storage, documentation and presentation of Czech language dialects. The project is funded by the Czech Ministry of Culture under the NAKI program and is coordinated by Dr. Martin Karafiat (ex-senior member of the NEUREM3 team).
- EU - Answer to EC Tender CNECT/LUX/2022/OP/0030 - LANGUAGE TECHNOLOGY SOLUTIONS Lot 2 (2023-2024) is not a research project but a delivery contract awarded by the EC to a BUT-led consortium of three Euroepan partners. It aims at delivering ASR data and production ASR systems (including training recipes) to the EC to boost the efficiency of European administration and industry. Dr. Karel Vesely (ex-senior member of the NEUREM3 team) is key ASR developer on the CNECT team.
- Horizon Europe project Multilingual and Cross-cultural interactions for context-aware, and bias-controlled dialogue systems for safety-critical applications (ELOQUENCE)[17] aims to research and develop new technologies supporting collaborative voice/chat bots for both low secure (low risk) and highly secure (high risk) applications. Key objective is to understand unstructured dialogues and conduct them in an explainable, safe, knowledge-grounded, trustworthy and unbiased way. Dr. Santosh Kesiraju (NEUREM3 senior team member) is a key contributor to Eloquence (links of speech and language technologies).
- From January 2025, CUNI and BUT are the core of consortium of OP JAK MSV project "Linguistics, Artificial Intelligence and Language and Speech Technologies: from Research to Applications" funded by the Czech Ministry of Education. The

[16] https://www.jamap.cz/
[17] https://eloquenceai.eu/

OP JAK initiative[18] has several sub-programs; MSV concentrates on approaching the basic research to the industry. Three innovative Czech companies – Phonexia, Mamma.AI and Phrase – are in the consortium, and the PI and coPI of NEUREM3 lead two crucial workl-packages: WP2 Multilingual systems and machine translation of written and spoken text (Bojar), and WP3 Person identification and identity assignment (Burget). The project will run for 4 years, building in many ways upon the research carried out in NEUREM3, and supports several key members of the NEUREM3 team.

- CUNI has been co-ordinating the EU project HPLT[19] (High Performance Language Technologes, September 2022–August 2025) which creates extremely large text data for the training of large language models and machine translation models for all EU languages. Multilinguality and efficiency (Section 2.9) in machine translation as studied in NEUREM3 has been strongly connected to HPLT goals and Jindřich Helcl from our team moved on to HPLT some time after leaving NEUREM3.
- RES-Q Plus[20] (Comprehensive solutions of healthcare improvement based on the global Registry of Stroke Care Quality – HORIZON-HLTH-2021-TOOL-06 project) aims to improve the quality of care for stroke patients and reduce healthcare costs through creating a fully automated tool for obtaining and processing data on the provided health care, providing feedback to care providers through artificial intelligence-based virtual assistants, and by creating virtual assistants for stroke patients that will monitor their condition. CUNI is a key project partner providing solution for automated information extraction from clinical documents.
- MEMORISE[21] (Virtualisation and Multimodal Exploration of Heritage on Nazi Persecution – HORIZON-CL2-2021-HERITAGE-01 project) aims to develop a framework to preserve and enhance materials on the Heritage of Nazi Persecution by virtualising and linking multimodal data and by developing and offering novel digital technologies to the general public for accessing, exploring and engaging such materials. CUNI is a core partner in the project contributing with its strong experience in processing of Holocaust survivors' testimonies gained in the last 20 years.

3.6 Networking and International research infrastructure projects

- H2020 ICT-48 Humane AI Net[22] (European network of human-centered artificial intelligence, coordinated by DFKI) was a large networking project built around ethics values and trust (Responsible AI). BUT and CUNI were both involved in this project, including participation in its μ-projects. Several publications stemmed from these μ-projects, notably [Bas+22] showcasing the use od neural adapters in dysarthric ASR.

[18] Named after Jan Amos Komenský (Johann Amos Comenius) — philosopher, pedagogue and theologian who is considered the father of modern education.

[19] https://hplt-project.org/

[20] https://www.resqplus.eu/

[21] https://memorise.sdu.dk/

[22] https://www.humane-ai.eu/

- H2020-MSCA-RISE ESPERANTO[23] (Exchanges for SPEech ReseArch aNd TechnOlogies, coordinated by the University of Le Mans) is a Marie-Curie project aiming at exchange of staff and students among European academic and industrial laboratories. The project heavily builds on BUT's expertise and international relations in speaker recognition and diarization. ESPERANTO supported a work-group at the Jelinek Summer Workshop on Speech and Language Technology (JSALT) in Le Mans in 2023 and will support JSALT 2025 in Brno, see section 5.8 for details).
- i-AIDA[24] (International AI Doctoral Academy) is a joint initiative of four ICT-48 networks (AI4Media, ELISE, HumanE-AI NET, TAILOR) and the VISION project to support a world-level AI education and research programe. CUNI is i-AIDA's founding member and BUT joined shortly after as a full member.
- CLARIN[25] (Common Language Resources and Technology Infrastructure) is an ERIC research infrastructure with the goal to support the sharing, use and sustainability of language data and tools for research in the humanities and social sciences.
- ELG[26] (European Language Grid, 2019–2022) was an industrial leadership EU project that developed the ELG platform listing, i.a., 8k corpora, almost 4k tools and services, and 2k organizations to enable easy access to language technologies for 24 official languages of EU, supporting a digital single market. CUNI was one of the 8 participants in the project.
- ELE and ELE2[27] are two consecutive EU actions to develop and maintain a strategic agenda and roadmap for making digital language equality a reality in Europe by 2030.

[23] http://esperanto.univ-lemans.fr/en/index.html
[24] https://www.i-aida.org/
[25] https://www.clarin.eu/
[26] https://live.european-language-grid.eu/
[27] https://ecspm.org/multilingualism-and-digital-technology/european-language-equality/

4 Involvement of team members

The team relies on two strong research leaders at both participating institutions, a group of researchers/post-docs, several PhD students and a dedicated research support staff.

4.1 Team leaders

Lukáš Burget is associate professor at BUT and scientific director of the BUT Speech@FIT research group. He is one of the most cited Czech researchers in computer science field and PI or co-PI or several Czech-, EU- and US-funded projects. In NEUREM3, he is the main scientific coordinator, he oversees all scientific work, directly coordinates the research done at BUT and supervises the BUT-CUNI cooperation. He is also the supervisor of several involved Ph.D. students and oversees BUT's hiring covered by the project. Technically, he is involved in most of the work done at BUT (therefore, we do not list all his NEUREM3-related publications in this section), he is especially involved in speaker recognition and diarization.

Ondřej Bojar is associate professor at CUNI, with full professorship process underway (the successful first defense at the faculty level happened on January 8, 2025). After early work around syntactic analysis and automatic extraction of lexico-syntactic information from plain texts, he has been working in the area of machine translation since 2006 when he participated the JHU workshop that developed the Moses translation toolkit. In 2016, he switched to neural MT. Since 2018, observing the unprecedented rise in translation quality, Ondřej Bojar has been broadening his research scope to several areas: towards speech recognition on the input side, towards tasks requiring deeper understanding of text than MT (e.g. summarization) and towards mapping and relation of human vs. machine language processing. which immediately brings in the subfield of multi-modality. He is involved in the vast majority of NEUREM3 research, and he is the most passionate about the analysis and relation of human language performance in various specific tasks and its simulation by deep neural networks.

4.2 Post-docs / researchers

- Karel Veselý, Ph.D. (BUT) is specializing in ASR and its low-resource, supervised and semi-supervised training [Kar+21; Koc+21]. He has been one of the main contributors to the KALDI toolkit. Moreover, he is a specialist in voice activity detection, an inevitable block of any speech processing system, and thus supports also speaker recognition efforts [Bur+20; Loz+20]. He is also involved in ASR applications, especially in conjuction with the H2020 ATCO^2 project [Zul+20]. Karel left the NEUREM3 team in 2022 to concentrate on more application-oriented projects, especially the EC-ASR contract CNECT.
- Martin Karafiát, Ph.D. is BUT's ASR "guru" and is behind almost all ASR work, especially low-resource [Kar+21; Koc+21]. Martin officially left the NEUREM3 team in October 2020 to lead a Ministry of Interior funded project (see Section 3.5) but has continuously supports NEUREM3.
- Hari K. Vydana, Ph.D. joined the project team shortly after its start in April 2019 (he was hired from one of the major Indian speech laboratories at IIIT Hyderabad)

before defending his Ph.D. Hari is an expert in end-to-end neural ASR models and in NEUREM3, he significantly advanced the topic of speech translation [Vyd+21c]. His e2e models however also significantly contributed to the success of OOV detection and processing work [Ego+21] and he contributed to automatical discovery of speech units [Ond+19]. Hari left the team in July 2021 for Huawei, Finland, followed by a position at the University of Cambridge, UK.

- Santosh Kesiraju, Ph.D. graduated at IIIT Hyderabad in January 2021. Santosh is responsible for probabilistic text and speech representations. He joined the NEUREM3 team in 2021 and was instrumental for BUT's speech translation work [Kes+23a; Kes+23b] and for cooperation with Indian partners on topics related to low-resource ASR and keyword spotting [Nad+22]. He supervised Marek Sarvas (an MSc student, see below).

- Johan Rohdin, Ph.D. received M.Sc. in Eng. (Civilingenjör) in Engineering Physics, and M.Sc. degree in Mathematics from Chalmers University of Technology in 2008, and Ph.D. degree in Computer Science from Tokyo Institute of Technology in 2015. He was with NEUREM3 in 2023, concentrating on speaker recognition [Ala+22; Sta+22] and co-supervising PhD work on speaker diarization [Han+24].

- Oldřich Plchot, Ph.D. obtained Ing. [MS]. Brno University of Technology in 2007 and Ph.D. from Brno University of Technology in 2014. He is renowned specialist in speaker and language recognition. He was member of NEUREM3 team between 2022 and 2023 and led research works on large models in speaker verification (with Junyi Peng, for example [Pen+23a]), language recognition [Sil+23a] and multichanel approaches to different speech processing tasks (with Ladislav Mošner, for example [Moš+23]).

- Pavel Pecina, Ph.D. is an associate professor working in the area of machine translation , cross-lingual information retrieval, and multimodal-data processing. He has been working on multimodal summarization (see Section 2.10.4) and optical music recognition (Section 2.10.5). He has also been supervising Michal Auersperger and Jiří Mayer.

- Věra Kloudová, Ph.D. joined the project starting from 2021, her expertise in translation studies and her experience in interpreting are an essential component for our analyses relating human and machine performance in the tasks. In particular, she was one of the main authors of the design of methodology for creating reference translation of a very good quality [Klo+23] and a study carefully comparing human and machine outputs in the simultaneous interpreting task [Klo+24].

4.3 PhD and MSc students

- Karel Beneš is MSc. graduate of BUT and was working in the area of language model adaptation [BB21], fusion of ASR systems [BKB24], speech translation [Kes+23a] and applying ASR and machine learning techniques in optical character recognition (for example [Kiš+22]). Karel submitted his PhD thesis "Recurrent neural networks as language models in recognition systems" in the end of 2024 is expected to graduate in early spring 2025.

- Murali Karthick Baskar is MSc. graduate of IIT Madras, India. He was working in end-to-end ASR and its training on disjoint speech and text resources using an ASR-TTS loop [Bas+21]. He also published a significant paper on ASR of dysarthria patients [Bas+22]. Karthick graduated in 2023 and is with Google Research N.Y.
- Anna Silnova received Specialist degree in applied mathematics from Saint-Petersburg State University, Russia in 2013 and Master degree in computer science from University of Eastern Finland in 2015. She is a key person in BUT's speaker recognition team and has been at all recent speaker, language and spoofing recognition challenges (from 2019 till now, for example [Ala+19; Roh+24]) and she also significantly contributed to the diarization work at BUT [Lan+21a; Lan+21b] and with the NTT Japan team [Del+23]. Anna defended her PhD thesis "Exploiting Uncertainty Information in Speaker Verification and Diarization" in 2022 and stayed with BUT as post-doc researcher.
- Jan Brukner is MSc. graduate of BUT. He started his PhD in 2020 and works on synthesis and voice modification with neural architectures. He also contributed to spoofing detection work [Roh+24].
- Junyi Peng received Bachelor degree in electronic information from Northeastern University, China in 2017 and Master degree in compute science from Peking University, China in 2020. He started his Ph.D. in fall 2021. During NEUREM3, he produced a significant volume of work in speaker recognition, from signal processing front-ends [Pen+21a; Pen+21b] to a series of papers on speaker recognition using large pre-trained models (for example [Pen+23a; Pen+23c; Pen+23b]). Junyi is also instrumental in international cooperation with Tencent (China/USA), NTT (Japan) and CMU (USA). He plans to submit his PhD thesis in 2025.
- Shuai Wang visited BUT when doing his PhD at Shanghai Jiao Tong University (SJTU), China. He was briefly on NEUREM3 (October 2019). At BUT, Shuai did a significant amount of work in speaker recognition and diarization, including significant contribution to several evaluation systems [Ala+20; Wan+20; Ala+19; Die+19; Zei+19; Wan+19]
- Jiangyu Han obtained B.E. degree in Electronic Information Engineering from Shandong University of Science and Technology and M.E. in Electronics and Communication Engineering from Shanghai Normal University, He ha senrolled into the BUT Ph.D. programme in 2022. He joined the NEUREM3 team in 2023. After his work on signal processing for speaker recognition [Han+22], he concentrated on speaker diarization [Han+24] and currently combines diarization and ASR in an end-to-end system that was successful in the 2024 CHiME-8 challenge.
- Marek Sarvaš was an MSc. student under the supervision of Santosh Kesiraju. With the support of NEUREM3, he worked on his Bc. thesis "Interpretability of Neural Networks in Speech Processing" and MSc. thesis "Detection of key information in emergency calls" and contributed to speech translation and dialogue processing activities [Kes+23b]. Marek currently pursues an industrial career in Mama.AI (a Czech conversational intelligence company).

CUNI:

- Michal Auersperger is a PhD student at CUNI. His research interests are representation learning, specifically at the concepts of disentanglement and compositionality. His work focused on solving the SCAN tasks [AP21]. and currently he has been studying the concept of compositionality in neural-network-generated representations [AP22] and preparing his dissertation to be defended in 2025.
- Sunit Bhattacharya is a M.Sc graduate of Central University of Rajasthan, India. He started his PhD in 2019 and is currently finishing the write-up of this dissertation thesis "Multimodal machines from a perspective of humans" on multilingual and multimodal learning of representations and relating machine and human processing in language tasks.
- Mgr. Dominik Macháček, Ph.D. was a PhD student at CUNI. He specializes in multi-source machine translation and simultaneous speech translation. He collaborated with experts on multi-source MT from NICT, Japan, which, among others, resulted in the well-received tool Whisper-Streaming [MDB23]. Dominik successfully completed his PhD in June 2024, thesis title "Multi-Source Simultaneous Speech Translation" [Mac24].
- Ivana Kvapilíková, Ph.D. was a doctoral student specializing in unsupervised machine translation and multilingual language modeling. Collaborating with researchers from the University of the Basque Country, she contributed to advancements in learning language-neutral sentence representations and enhancing their multilingual alignment using synthetic data. Her work on the project led to the successful completion of her PhD in February 2024 and the publication of a book based on her thesis [Kva25].
- Peter Polák is a PhD student at CUNI. He specializes in simultaneous speech translation. Part of his research was also supported by the START Programme[28] of Charles University. His main achievements within NEUREM3 include the best-performing system in simultaneous speech translation task at IWSLT 2023 [Pol+23b], or advancing the simultaneous speech translation decoding done while research visit at Carnegie Mellon University, USA [Pol+23a]. Together with Dávid Javorský, he also co-organizes the IWSLT Simultaneous Speech Translation Shared Tasks [Aga+23; Ahm+24].
- Josef Jon is a third year PhD student at CUNI. He specializes in machine translation and in how the properties of a text are changed by processing it or generating it by NLP models. He has experience with enforcing use of proper terminology in NMT [Jon+21c; Jon+23], developing quality-aware NMT decoding scheme [JB24; JB23a], low-resource languages [JB23b] and general machine translation [Jon+21a; JB23b; JPB23].
- Dávid Javorský is a PhD student at CUNI, currently in his fourth year. His research focuses on length control in machine translation, particularly by analyzing word importance and exploring methods to incorporate word importance scores into length control mechanisms. His work has resulted in several publications on this topic [JBY23; OJB24]. In addition to his research, he is actively involved in

[28] https://cuni.cz/UKEN-1340.html

spoken speech translation. Together with Peter Polák, he co-organizes the IWSLT Simultaneous Speech Translation Shared Tasks [Ana+22; Aga+23; Ahm+24].

4.4 Support staff

At BUT, NEUREM3 was supported by Renata Kohlová that was instrumental in contractual, reporting and management work. BUT part of the team was also largely helped by Šárka Nesvedová (payroll) and Sylva Otáhalová (financial management).

At CUNI, project administration was carried out by Libuše Kaprálová who was later replaced by Hana Kubištová. Both were supported by Tereza Vojtěchová (responsible for managing annotators and student workers on short-term contracts). Stanislava Gráf has joined the team for a small portion of time, to help aligning the research across projects solved at CUNI and more importantly, to help CUNI NEUREM3 researchers with their applications for ERC grants.

Both teams are supported by the standard project and administrative support personnel of their respective faculties.

5 Outcomes of the project and international excellence

5.1 Project outputs

The following table shows the project outputs as specified in the proposal versus the reality:

type of results	promised	done
J - article in an impacted journal (Jimp - WoS)	4	16
J - article in journal included in the Scopus database (Jsc)	2	0
J - article in other journals (Jost)	2	2
B - book (B)	1	1
C - chapter in a book (C)	2	2
D - article in conference proceedings (D)	30	174
O - Other - ERC project proposal	1	2

Note that we did not reach the planned numbers of "weak" journal publications "Jsc". We believe that this is sufficiently compensated by significantly higher number of papers in impacted journals, than planned: 16 vs. 4. The required numbers have been reached in other categories. We note that the publication type "D" (article in conference proceedings) is somewhat peculiar with respect to reporting. The field of natural language processing and computational linguistics departs somewhat from general standards and publishes work *primarily* at conferences organized by colleagues in the field, including its own indexing service (ACL Anthology). Many of excellent conferences from the field thus do not apply for indexing by Scopus, let alone WoS. Similar situation occurs in the speech processing area, where the main focus is on two primary conferences: IEEE ICASSP and ISCA Interspeech, but smaller workshop papers (such as Speaker Odyssey in the areas of speaker and language recognition and speaker diarizarion) are highly valued as well.

In order to reach our audience, we often decided to present our work at conferences that will never make it to Scopus. We report our publications as "D", although we are aware that they may not be recognized as such when formally reviewed by GACR. This is reflected, e.g. in the protocol of our report after the year 2023 where a significant number of our "D" publications were not approved. We believe that the big surplus (we planned to publish 30 "D" outputs) allows us to maintain this strategy.

5.1.1 Impact of publications

According to GACR requirements, we select five top-cited publications supported by the project. The citation analysis is performed using Google Scholar. In the speaker diarization category, more papers would enter this list, but we omit them to make space for other topics. It is not surprising, that papers published in the early phases of the project managed to gather more citations.

1. "Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals" [Pop+20] published in Nature Communications in 2020 gathered 364 citations.

2. "Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures" [Žmo+19] published in journal IEEE Journal of Selected Topics in Signal Processing in 2019 gathered 246 citations.

3. "Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks" [Lan+22a] published in journal Computer Speech & Language in 2022 gathered 231 citations. In 2023, according by the statistics published by the journal, the VBx paper was 12th on the list of journal's most cited papers dating from 2020 and later.

4. "Analysis of DNN speech signal enhancement for robust speaker recognition" [Nov+19a] published in journal Computer Speech & Language in 2019 gathered 58 citations.

5. "CUNI-KIT system for simultaneous speech translation task at IWSLT 2022" [Pol+23b], our system description for the shared task gathered 23 citations.

Shared task results papers are also highly cited, e.g. IWSLT 2023 [Aga+23] gathered already 52 citations.

In general, the team leaders and senior researchers on the team have a significant citation track, the following table presents their h-indices (SCOPUS and Google Scholar) retrieved in January 2025 during the preparation of this final report. We are proud that some of the our freshly graduated PhD students actually have excellent bibliometric results, too:

person	SCOPUS	G. scholar
Lukáš Burget	46	63
Ondřej Bojar	38	44
Martin Karafiát	30	39
Pavel Pecina	20	30
Karel Veselý	18	26
Anna Silnova (PhD defended 2022)	12	17
Murali Karthick Baskar (PhD defended 2023)	12	13
Jindřich Helcl (PhD defended 2022)	11	12
Santosh Kesiraju (PhD defended 2020)	8	11
Dominik Macháček (PhD defended 2024)	6	8
Peter Polák	5	6
Juni Peng	4	6
Josef Jon	4	6
Ivana Kvapilíková (PhD defended 2024)	3	5
Dávid Javorský	3	4
Karel Beneš (soon to defend PhD)	4	4
Věra Kloudová	2	4
Michal Auersperger (soon to defend PhD)	2	3
Sunit Bhattacharya (soon to defend PhD)	1	3

5.2 Software

For projects dealing with speech, NLP and AI in general, releasing software is of crucial importance. Samples of the important software releases during the lifetime of the project,

are below (we do not mention small codebases, teaching demos, etc):

VBHMM x-vectors Diarization (aka VBx) was released in February 2020[29]. Since then, it has received more than 261 stars in GitHub, and more than 57 forks. The model and software have been widely adopted by the community: not only we have used it with great success in challenges but also several other teams have used VBx as part of their systems in VoxCeleb Challenge 2020[30], Third DIHARD Challenge 2020[31] and VoxCeleb Challenge 2021[32]. One of the highest-cited papers of NEUREM3, [Lan+22a] (see above) is strongly related to this software.

DiaPer: End-to-End Neural Diarization with Perceiver-Based Attractors is a neural end-to-end follow-up of VBx and was released in February 2024. So far, it gathered 57 stars and 3 forks. It is complementing the DiaPer journal paper [Lan+24].

SLTev is a tool for evaluating (simultaneous) speech recognition and translation. It received the **outstanding demo award** at EACL 2020. [Ans+21].

ASR and MT transformer models The code for training models related to our work in end-to-end ASR and ST [Vyd+21a] were released:

- Transformer_ASR/Transformer_E2E-ST [33]
- Transformer_MT code [34]

Mashcima a library that produces synthetic images of monophonic handwritten music[35] presented in [MP21].

ALIGNMEET is a tool for manual alignment between meeting transcript and meeting minutes, and for the evaluation of meeting minutes created automatically. The tool is available both as source code[36] and as a Python package.[37]

Target Speaker ASR with Whisper [38] contains the end-to-end model containing speaker diarization and ASR that was used in our successful participation in CHiME-8 challenge (see below). Although released only recently (December 2024), the code gathered 32 stars and 3 forks. It accompanies the challenge papers [Pol+24b; Pol+25] and will serve as basis for our work in JSALT 2025 (workgroup led by Lukas Burget and Samuelle Cornel from CMU, USA).

[29] https://github.com/BUTSpeechFIT/VBx
[30] https://www.robots.ox.ac.uk/~vgg/data/voxceleb/competition2020.html
[31] https://dihardchallenge.github.io/dihard3/
[32] https://www.robots.ox.ac.uk/~vgg/data/voxceleb/interspeech2021.html
[33] https://github.com/BUTSpeechFIT/ASR_Transformer.git
[34] https://github.com/BUTSpeechFIT/MT_Transformer.git
[35] https://github.com/Jirka-Mayer/Mashcima
[36] https://github.com/ELITR/alignmeet
[37] https://pypi.org/project/alignmeet/
[38] https://github.com/BUTSpeechFIT/TS-ASR-Whisper

Whisper-Streaming [39] presents our technique for running off-the-shelf speech recognition systems in online form, i.e. while the input is still arriving. Whisper-Streaming demonstrates state-of-the-art SST, reaching extremely high quality and robustness in many languages. The code on GitHub received large positive feedback, documented by more than 2 300 stars in January 2025. (The corresponding paper [MDB23] had 21 citations on Google Scholar.)

5.3 Data

An important part in any AI-related project is collection, generation and consolidation of data. Although this is not the primary goal of NEUREM3, we were active also in this domain. Several data resources (ASR data for Czech: Large Speech Corpus for Czech and Speech test set with additional relevant texts, Synthesized Training Data for Handwritten Music Recognition, COSTRA: Corpus of Complex Sentence Transformations, and Data for Meeting Summarization shared task), were covered already in our interim report. In period 2022-2024, the following data releases were done:

- MultiSV: dataset for far-field multi-channel speaker verification[40] is a corpus designed for training and evaluating text-independent multi-channel speaker verification systems [Moš+22b]. The training multi-channel data is prepared by simulation on top of clean parts of the Voxceleb dataset. The development and evaluation trials are based on either retransmitted or simulated VOiCES dataset, which we modified to provide multi-channel trials. The corpus gathered 27 stars and 3 forks.
- ELITR Minuting Corpus[41] [Ned+22] is a corpus of meeting transcripts and summaries in the form meeting minutes. The corpus served in two instances of the AutoMin shared task (2021 and 2023) and the third instance is in preparation for 2025. As of January 2025, the corresponding paper has 32 citations on Google Scholar.
- Eyetracked Multi-Modal Translation (EMMT) dataset[42] already mentioned in the interim report [Bha+22a] served in one published paper [WZB24] (see Section 2.12.4) and one PsyArXiv paper [Bha+22b].

5.3.1 Patents

Patenting has not been the primary focus of the NEUREM3 team but several Japanese patents were filed thanks to our cooperation with NTT Communications Science Laboratories. Of these one patent (JP7112348B2)[43] is already awarded.

CUNI managed to patent one aspect of multi-source speech translation as explored in the PhD thesis by Dominik Macháček [Mac24], receiving US patent: O. Bojar, D. Macháček: Computer-implemented method of real time speech translation and a computer system for carrying out the method (US 12,056,457 B2)[44]

[39] https://github.com/ufal/whisper_streaming/
[40] https://github.com/BUTSpeechFIT/MultiSV
[41] https://ufal.mff.cuni.cz/elitr-minuting-corpus
[42] https://ufal.mff.cuni.cz/eyetracked-multi-modal-translation
[43] https://patents.google.com/patent/JP7112348B2
[44] https://patentsgazette.uspto.gov/week32/OG/html/1525-1/US12056457-20240806.html

5.4 PhD theses

PhD students have been extremely important and valuable for the project. Several PhD students (see section 4.3) were members of the NEUREM3 team while others were supervised by NEUREM3 senior staff and/or cooperated with NEUREM3 team while being supported by other projects. We are proud of post-PhD careers of our PhD students. We highlight the following ones.

- Jindřich Libovický, 2019, "Multimodality in Machine Translation". While Jindřich did not formally connect with NEUREM3 for any extended period of time, he is our very active colleague. Among other activities, Jindřich is (jointly with Jindřich Helcl) co-organizing the course NPFL140 on Large Language Models, as a research associate.

- Santosh Kesiraju, 2020, "Generative models for learning document representations along with their uncertainties" (while being co-supervised by Lukas Burget, he defended his thesis defended at IIIT Hyderabad, India). Santosh was a senior researcher in NEUREM3 team.

- Lucas Ondel, 2021, "Discovering Acoustic Units from Speech: a Bayesian Approach", Lucas holds a French CNRS position and is a researcher at LISN laboratory at Université Paris-Saclay (France).

- Ondřej Novotný, 2021, "Improving Robustness of Speaker Recognition using Discriminative Techniques". Ondra is the chief scientist at ThreatMark, a Czech AI company.

- Jindřich Helcl, 2022, "Non-Autoregressive Neural Machine Translation" continues with his career at CUNI and is now a research associate.

- Kateřina Žmolíková, 2022, "Neural Target Speech Extraction". Katka is an AI Research Scientist with Meta AI in London.

- Ekaterina Egorova, 2022, "Out-of-Vocabulary Words Detection and Recovery". Katia is an NLP researcher at Seznam.cz

- Anna Silnova, 2022, "Exploiting Uncertainty Information in Speaker Verification and Diarization". Anna is a post-doc researcher at BUT. She was NEUREM3 PhD and then senior team member.

- Murali Karthick Baskar, 2023, "Semi-Supervised Speech-to-Text Recognition with Text-to-Speech Critic", Karthick is researcher with Google Research N.Y. He was NEUREM3 PhD team member

- Federico Nicolás Landini, 2024, "From Modular to End-to-End Speaker Diarization", Fede holds a position of Research Scientist at Deepgram in London, UK.

- Dušan Variš, 2023, "Learning capabilities in Transformer Neural Networks". Dušan is now researcher at CUNI, actively involved in EU and other research projects at the department.

- Bolaji Yusuf, 2024, "End-to-End Open Vocabulary Keyword Search". Bolaji defended at Bogazici University in Istanbul (Turkey) but during his PhD, he spent most of the time at BUT and significantly contributed to NEUREM3 efforts [Ond+22; YČS23]. Bolaji is a post-doc researcher at BUT, and works also with Eva Corporation, a US-Turkish start-up aiming at analog computations in AI.

- Dominik Macháček, 2024, "Multi-Source Simultaneous Speech Translation". Dominik's PhD thesis was strongly in line with NEUREM3 research goals. Dominik is currently a researcher at CUNI, planning for his post-doc position probably at the University of Edinburgh.

- Ivana Kvapilíková, 2024, "Towards Machine Translation Based on Monolingual Texts". Ivana's research formed one of the central topic directions of NEUREM3. Ivana now works part-time at a Czech start-up and remains connected with the department in two research projects.

5.5 International Evaluations and Challenges

Open evaluations and challenges (also dubbed benchmark campaigns, shared tasks etc.) are an important tool to assess where we stand compared to the state of the art (SotA). The challenges are organized both by established national or international organizations (such as NIST[45]) and by the scientific community. They allow for objective and quantitative comparison of results among research labs and companies by specifying defined data-sets with their division to training / development / test, evaluation metrics, time plan, evaluation protocol (for example allowing or forbidding using additional data, split of evaluation into conditions, etc.), and technical evaluation framework based either on submission of data to the evaluator at a fixed deadline and/or by running a "leader-board" accessible through well defined interfaces. The organizers often provide also a baseline system or a set of baseline results.

Challenges are taken very seriously in the speech/NLP/AI communities as they allow for objective comparison between research approaches, systems, and laboratories. NEUREM3 team members were active both as challenge organizers and participants. Challenges happening in 2019-2021 were extensively covered in our interim report, the following subsections therefore concentrate on challenges organized in 2022-24.

5.5.1 Machine translation challenges (WMT, WAT)

We continued our long-term involvement in WMT (Conference on Machine Translation) and its General Translation Shared Task [Koc+23; Koc+22a]. With NEUREM3 support, we were also involved in the multimodal translation task at WAT (Workshop on Asian Translation) [Nak+22; Nak+23].

5.5.2 Speech translation challenges (IWSLT)

We co-organized a challenge aimed at evaluating systems that translate text or audio from a source language into text in a target language, focusing on both translation quality

[45] US National Institute of Standards and Technology, https://www.nist.gov/itl/iad/mig

and latency. The challenge included two tracks: text-to-text and speech-to-text. Our contribution was specific to the speech-to-text track. Each year featured updates and modifications to the setup, including novel elements. The details for each year are as follows:

2022 In 2022 [Ana+22], the language directions were English-to-German and English-to-Japanese. Our work focused exclusively on English-to-German. Two key innovations were introduced that year: (1) A manual evaluation for the English-to-German speech-to-text track and (2) a human interpretation benchmark.

The first innovation was the human evaluation, which employed our Continuous Rating method [JMB22] (see Section 2.8.3 for details). In this approach, bilingual speakers listened to the audio and rated the quality of subtitles generated by the submitted systems. This evaluation was used to assess both the translation quality of system submissions and interpreting quality—our second innovation that year. The human interpretation benchmark was conducted remotely by a native German speaker who is a certified German conference interpreter as well as a sworn translator and interpreter. The interpreting was transcribed by students of German for Intercultural Communication at the Institute of Translation Studies, Charles University, Faculty of Arts.[46]

The English-to-German task included two parts in the test set: TED Talks and recordings of Non-Native students that presented their mock companies.

The human evaluation showed that the interpreter outperformed the systems on the Non-Native test set (difficult, out-of-domain). However, on the TED Talks test set (in-domain), human evaluators rated some system translations higher than the interpreter's. While this might suggest that SST models can match or even surpass human performance, we interpret these results cautiously, attributing them to the translation and evaluation setup of the shared task. First, the systems were provided with a golden segmentation of speech into sentences, a significant advantage given that simultaneous speech segmentation is a challenging and under-researched area [Pol23]. Second, the interpreter was performing "simultaneous interpreting," a task distinct from translation, while human evaluators judged translation quality rather than interpretation quality. This likely led to unfair penalties against the interpreter. Overall, these results suggest that under optimal conditions (in-domain speech, high-quality segmentation), automatic systems for simultaneous speech translation can be a viable alternative to human interpreting.

2023 In 2023 [Aga+23], the task covered three language directions: English-to-German, English-to-Chinese, and English-to-Japanese. As in the previous year, we focused exclusively on the English-to-German direction. We used a similar evaluation setup as in the previous year, with one notable difference: the human interpretation was omitted. The Continuous Rating scores [JMB22] of this year reinforced that the Non-Native portion of the test set was substantially more challenging, primarily due to less favorable sound conditions: We strikingly observed that the top-performing systems in the human evaluation for the TED Talks subset reported performance drop on the Non-Native portion. This may indicate a degree of overfitting to the clean input of the Common (TED Talks) subset.

[46] http://utrl.ff.cuni.cz/en

We also participated in IWSLT shared task on low-resource speech-to-text translation from Marathi → Hindi [Kes+23a]. We made use of transfer learning from bilingual ASR systems and built two systems. The primary one is an end-to-end system, where as the second one is a cascade of ASR and MT systems. The primary system achieved the top BLEU score of 39.6, and chrF2 of 63.3 on the unseen test set, outperforming our cascade pipeline by 11 BLEU points, and second best submission by 8 BLEU points absolute [Aga+23]. These result confirmed that transfer learning can be used to train direct speech-to-text translation in low-resource scenarios. Our implementation and training recipes are based on ESPnet and are available for public[47].

2024 In 2024 [Ahm+24], the task covered the same language directions as the previous year (English-to-German, English-to-Chinese, and English-to-Japanese) with the addition of a new direction: English-to-Czech. For this novel language pair, we supplied a dedicated evaluation test set. The test set was sourced from mock conferences conducted as part of the interpreting curriculum at the Faculty of Arts, Charles University. In these mock conferences, speakers impersonated celebrities or notable figures, delivering fictional speeches on predetermined topics. Reference translations for the English-to-Czech test set were provided by professional translators.

We also introduced a baseline system based on the Whisper model [Rad+23], specifically the large-v2 version. To adapt this offline model for simultaneous translation, we applied our onlinization technique [Pol+22b], enabling it to operate in a real-time regime. Additionally, we employed prompting to incorporate translation history from previous segments, leveraging context for improved performance.

Unlike previous years, the evaluation in 2024 relied solely on automatic metrics, without employing the Continuous Rating method. The results indicated that at the segment level, the baseline system underperformed compared to participants' systems. However, when provided with one or two of its previous translation outputs, the baseline outperformed the competitors. This highlights the critical importance of context in speech translation tasks and suggests that the optimal utilization of large language models (LLMs) for this domain remains an open area of exploration.

5.5.3 Speaker, language and deepfake recognition

NIST Speaker recognition evaluations The ABC team's submission to the NIST Speaker Recognition Evaluation (SRE) 2024, which was led by BUT, was a collaborative effort involving multiple other institutions and companies, including Politecnico di Torino (Polito), Phonexia, Omilia, Universidad Autonoma de Madrid (UAM), Computer Research Institute of Montreal (CRIM). The team participated in all evaluation tracks: audio-only, visual-only, and audio-visual—under both fixed and open conditions, utilizing a wide range of innovative approaches for embedding extraction, backend scoring, calibration, and fusion.

For the audio-only systems, we developed multiple frontends based on ResNet architectures, and we also experimented with recent models such as ReDimNet-B3. These models were trained to focus mostly on the telephone domain, for which we utilized data

[47] https://github.com/BUTSpeechFIT/espnet/tree/main/egs2/iwslt23_low_resource

provided by NIST. The frontends employed advanced state-of-the-art Additive Angular Margin (AAM) Softmax to obtain utterance-level fixed-length vectors (speaker embeddings). These speaker embeddings were later subject to different scoring methods that included cosine similarity, Probabilistic Linear Discriminant Analysis (PLDA), and Pairwise Support Vector Machines (PSVM). We also experimented with multiple embedding preprocessing techniques like Linear Discriminant Analysis (LDA) and nuisance attribute projection to handle mismatches in language, gender, and channel. Calibration and fusion relied on logistic regression.

To create the visual-only systems, we used pre-trained face recognition models such as ResNet100 and MagFace, trained on datasets like MS1MV2. Preprocessing steps included face detection (RetinaFace) and alignment (MT-CNN). Similarly to audio, we explored multiple scoring strategies. Calibration and fusion were performed again using logistic regression or by simple score averaging.

The audio-visual systems combined scores from audio-only and visual-only systems through logistic regression-based fusion. This approach leveraged the strengths of both modalities to improve overall performance.

Under the open training condition, we leveraged larger ResNet models, incorporated new training datasets such as VoxBlink2 and VoxCeleb2, and experimented with fine-tuning pre-trained foundation models like XLS-R with multi-head factorized attention backends. These enhancements allowed for better generalization across diverse conditions.

Overall, our submission demonstrated robust performance across all tracks by integrating state-of-the-art techniques in speaker recognition for audio, visual, and audio-visual modalities.

NIST Language recognition evaluations In 2022, we participated in the latest edition of the NIST language recognition evaluation (LRE). The particular task of NIST LRE 2022 consisted of classifying audio recordings into 14 language classes with the focus on low-resource languages from Africa. The evaluation had two tracks: the Fixed condition, where participants are required to use only predefined data sets, and the Open condition allowing to use any data to train the systems.

The ABC team, led by BUT, developed and submitted multiple systems for both fixed and open conditions. In the fixed condition, we implemented several classifiers based on ResNet, RepVGG, and ECAPA-TDNN models and different acoustic features.

For the open condition, we scaled up the efforts with a ResNet101 model and two systems based on the pre-trained XLS-R-1B model. These XLSR models were fine-tuned to be language classifiers with a TDNN classification head. We also experimented with different backends on top of utterance embeddings. These included a Gaussian Linear Classifier (GLC) for most systems and PLDA for the XLS-R based models.

During the efforts, we compared different topologies of DNNs, different backend classifiers, and the impact of the data used to train them. Results with XLS-R pre-trained models that could have been used in the Open condition indicate that the task of language recognition task can largely benefit from the capabilities of pre-trained models, especially if they are combined with a large amount of task-specific (language-labeled) data for fine-tuning.

AVSpoof 5 We participated in the fifth ASVspoof challenge (ASVspoof5) held in 2024. The objective in this challenge is to detect artificially generated voices. This is important since, due to the evolution of deep learning, it is now nearly trivial to generate high-quality speech with any lexical content of any desired target voice using only a small amount of example speech from the target voice. The ASVspoof5 challenge [Wan+24] involved two tracks. Track 1, as in previous years, was a conventional deepfake detection task, i.e., to develop a system that can discriminate between bona fide and spoofed speech. Track 2 was a more complex task in which deepfake detection must be done in combination with automatic speaker verification, i.e. the system should not just discriminate between bona fide and spoofed speech but also, in the case of bona fide speech, tell whether it is from a claimed identity or not. This task is commonly referred to as "spoofing-robust automatic speaker verification" (SASV) [Jun+22]. For each of the two tracks there were both an open condition (any training data can be used) and a closed condition (the training data are fixed and the same for all participants). We participated in all tracks/conditions. For Track 1, closed condition, we followed the top-ranked teams from previous years and utilized ResNet18 as our submitted system. In addition, we explore the influence of training with different labeling schemes. For the open condition, given the promising performance of self-supervised learning (SSL) models for spoof detection [WY22; Tak+22; Kaw+23] we compared different SSL models as front-end. We used our previously proposed Multihead Factorized Attentive Pooling (MHFA) [Pen+23a] to efficiently aggregate information from the SSL model. For Track 2, we combined our best deepfake detection systems from Track 1 with a state-of-the-art automatic speaker verification system based on ResNet221. The two systems were combined by computing the likelihood ratio for the joint task which allows for optimal decisions to be taken. Deteailed descriptions of our systems can be found in [Roh+24].

5.5.4 Speech recognition

Albayzin-RTVE 2022 Speech-to-Text Challenge We participated in the Albayzin-RTVE 2022 Text-to-Speech Challenge. This challenge focuses on developing robust Automatic Speech Recognition (ASR) systems capable of accurately transcribing challenging Spanish broadcast news and TV shows, encompassing diverse audio sources such as political debates, interviews, and documentaries. To address this demanding task, we developed a multi-system approach that takes advantage of hybrid and end-to-end architectures [Koc+22b]. This diversity aims to capture the strengths of different modeling paradigms and improve overall robustness. To effectively combine the diverse outputs, we implemented a novel system fusion strategy. A crucial component of this strategy involved calibrating word-level confidences from each system, addressing biases and inconsistencies arising from different model architectures and training data. This calibration step significantly improved the reliability of the fusion process, leading to a substantial improvement in the overall transcription accuracy [BKB24].

Our system fusion approach achieved outstanding results, securing first place in the challenge. This success underscores the importance of employing diverse models and implementing robust fusion techniques, particularly when dealing with the inherent variability and biases present in individual ASR systems. The calibrated confidence scores played a pivotal role in enabling the fusion algorithm to effectively leverage the strengths of each model, ultimately leading to significant improvements in transcription accuracy.

CHiME-7 Task 1 We participated in the CHIME-7 DASR Challenge in 2023, focusing on the far-field acoustic robustness sub-track of Task 1—Distant Automatic Speech Recognition. Our system integrated Guided Source Separation (GSS) to enhance single-channel input for ASR, leveraging self-supervised learning-based speech representations and multi-system fusion. We improved training data preparation by augmenting LibriSpeech with background speakers and refining segmentation strategies. Our ASR models were built using ESPnet and HuggingFace toolkits, incorporating fine-tuned WavLM and Conformer-Transformer architectures. We explored both neural and non-neural speech enhancement approaches. Our non-neural GSS-based enhancement incorporated mask-based post-filtering and MVDR beamforming, while our neural models utilized Target Speaker Extraction (TSE) with SpeakerBeam, DPCCN, and TF-GridNet architectures. Additionally, we experimented with K2's Zipformer-Transducer framework, demonstrating promising results for streaming ASR. Our final system employed the Hystoc tool [BKB24] for efficient system fusion, leveraging multiple ASR hypotheses with confidence-based selection and NIST ROVER voting. The fusion of our best models resulted in state-of-the-art performance across CHiME-6, Mixer6, and Dipco datasets.

CHiME-8 Task 2 - NOTSOFAR We participated in the CHIME 8 NOTSOFAR Challenge in 2024 in both single- and multi-channel tracks, receiving the Jury Prize for developing one of the most practical, efficient, and novel systems. Our approach integrates a modified Pyannote diarization pipeline with WavLM-based end-to-end neural diarization [Han+24] and introduces two diarization-aware ASR methods: Query-Key Biasing, which adapts Whisper's attention mechanism, and Frame-Level Diarization-Dependent Transformations, which apply trainable transformations to Whisper's inputs [Pol+24b; Pol+25]. These methods enable target-speaker ASR and significantly improve multi-speaker transcription, including an absolute 12.9 % ORC-WER reduction on NOTSOFAR-1 compared to baseline. Additionally, we extended both systems to multi-channel setups with cross-channel communication, achieving state-of-the-art results across datasets like AMI, and Libri2Mix. Detailed descriptions of our systems can be found in [Pol+24a].

5.5.5 AutoMin: Meeting summarization shared tasks

NEUREM3 allowed us to create and organize two instances of the shared task automatic meeting summarization: AutoMin 2021 [Gho+21][48] and AutoMin 2023 [Gho+23][49].

The tasks clearly document the technological change that came with pre-trained models and primarily large language models (LLMs). While the first instance in 2021 was demonstrating effectiveness of the pre-trained BART model and the limitations of small Transformer models, the second instance in 2022 was clearly dominated by LLMs, with GPT-4 delivering best summarization results and also serving best in automatic evaluation of summaries.

As of now, AutoMin 2025 is in preparation for SIGDIAL 2025, adding the cross-linguality and question answering as new challenges.

[48] https://elitr.github.io/automatic-minuting/
[49] https://ufal.github.io/automin-2023/

5.5.6 Evaluation of European speech and language technologies

From December 2024, part of the NEUREM3 team is "switching sides" and from participants in evaluations, it is becoming the evaluators. The EDF[50]-supported project "Advancing Robust and Creative Human language technologies through CHallenge Events and Research – ARCHER" coordinates by Airbus, aims to enhance language processing capabilities, particularly in the areas of speech recognition, machine translation, and natural language processing. NEUREM3 team members will be active in preparation of evaluation pipeline, and produce several competitive baselines in the areas of ASR and OCR.

5.6 International rankings

Although the metrics used by different rankings are often quite obscure, we were pleasantly surprised to find the BUT team among Arnetminer 2000 five "Speech Recognition Most Influential Organizations", and to have the PI of our project, Lukáš Burget, among the 100 "Speech Recognition Most Influential Scholars".[51]

According to Research.com, Lukáš Burget also ranks as the 5th most cited Czech researcher in the area of Computer Science while Ondřej Bojar has a solid 21st position.[52]

5.7 Best paper awards

- Junyi Peng's paper "Effective Phase Encoding for End-to-end Speaker Verification" [Pen+21a] was short-listed for the ISCA Interspeech 2021 Best Student Paper.[53]
- Anna Silnova's paper "Probabilistic Embeddings for Speaker Diarization" [Sil+20] obtained the inaugural Jack Godfrey's Best Student Paper Award at Odyssey 2020, The Speaker and Language Recognition Workshop (planned for Tokyo and held virtually dur to COVID).
- Our paper "SLTev: Comprehensive Evaluation of Spoken Language Translation" [Ans+21] received an outstanding demo paper award at EACL 2021 for its ambition to simply evaluate all the three key axes of speech translation: quality, delay and flicker.

5.8 Organization of international events

Events organized during 2019-21 (namely Interspeech 2021 conference, WMT Shared Tasks, WAT Shared Tasks, SummDial Session at SIGDIAL and AutoMin Shared Task, collocated with Interspeech 2021) were covered in the interim report. During 2022-24, we organized or co-organized the following events:

[50] European Defence Fund
[51] https://www.aminer.org/ai2000/sr
[52] https://research.com/scientists-rankings/computer-science/cz
[53] https://www.interspeech2021.org/best-student-paper-shortlist

5.8.1 MT Marathons 2022 and 2024

CUNI continueed in its tradition of approximately bi-annual organization of the Machine Translation Marathon, a week-long gathering of researchers, students as well as practitioners of machine translation field. MT Marathons feature introductory lectures (2022 and 2024 Intro given by Ondřej Bojar), keynote talks (Friday keynote in 2024 given by Věra Kloudová) and projects in the form of a hackathon.

Jindřich Helcl was the main organizer of both MT Marathons.

5.8.2 JSALT 2025

For the last 30 years,[54] each summer, Center for Language and Speech Processing of the Johns Hopkins University (CLSP JHU) organizes and hosts a few international teams for an intensive 6-week research workshop on speech and language engineering, nowadays tightly linked to machine learning, and artificial intelligence (ML/AI). These very successful workshops have had a widespread impact on the Human Language Technology community. Many workshop teams have had a lasting influence on subsequent research and industrial implementations through the publications, software, and data that they produce. For many workshop participants, the biggest benefit is the interaction with other researchers, seeding new and lasting collaborations. the PI, coPI and several NEUREM3 team members have an extensive track in JSALT workshops.

In 2025, for its 32nd edition, this workshop will be held in at BUT, from June 23rd to August 1st. It will be co-hosted by the Brno University of Technology and Phonexia.

[54] https://www.clsp.jhu.edu/workshops/

6 Team work

6.1 Cooperation of Brno and Prague teams

In the project proposal, we planned "three physical events every year: two project strategy meetings, including the senior members of the team and the leaders of BUT and CUNI groups (prof. Černocký and prof. Hajič) and one full-day NEUREM3 workshop" — this plan was followed in 2019 but was heavily disrupted by the COVID-19 epidemics raging (in Czechia) from the "black Friday" March 13th 2020. The physical meetings were replaced by zoom teleconferences.

BUT and CUNI teams intensively cooperated on speech translation — Hari Vydana launched this activity making use of his extensive know-how of end-to-end ASR systems [Vyd+21a] and both teams cooperated on organizing and participating in the IWSLT challenge [Ana+21].

Jonáš Kratochvíl, an MSc student of CUNI, visited BUT for 2 months in 2019 to get acquainted with speech recognition. He was mainly trained by NEUREM3 team members Karel Veselý and Ekaterina Egorova, and he subsequently made use of this know-how when building speech translations systems back at CUNI.

After the end of COVID, the teams resumed normal operations with regular meetings of team members, teleconferences and other common activities. An important activity was ensuring follow-up funding in years 2023 and 2024, we concentrated mainly on the GAČR EXPRO call for proposals in 2024 and Programme Johannes Amos Comenius (OP JAK) sponsored by the Ministry of Education of the Czech Republic (MoE) in 2022 and 2024. We were successful in securing the funding with our 2024 OP JAK proposal "Linguistics, Artificial Intelligence and Language and Speech Technologies: from Research to Applications", coordinated by Prof. Jan Hajič from Charles University, in which the PI and co-PI of NEUREM3, Lukáš Burget and Ondřej Bojar, have important roles of work-package leaders, see also Section 3.5.

6.2 Professional elevation of team members

During NEUREM3, 13 PhD students (either directly funded by the project, supervised by team senior members, or working on related research at CUNI and BUT) defended their PhD theses, see Section 5.4. We are happy that most of our PhD graduates received offers for positions in prestigious academic and industrial research laboratories, and that some of them stayed with our labs.

Helped by the research results obtained during NEUREM3, The co-PI, Ondřej Bojar, applied for a full Professor position at CUNI. The process (rather complicated in Czech Universities, and especially at CUNI) involves a defense at the Faculty level (done on January 8, 2025) followed by the University level (expected in spring or early summer 2025) and concludes with the nomination from the Czech president two times a year.

6.3 Community building

6.3.1 Czech speech / NLP days

Working on a flag-ship Czech speech/NLP project, we felt obliged (and actually also very happy) to be the leaders and organizers of the Czech speech and NLP community.

- On Monday 21st October 2019, BUT, CUNI and the University of West Bohemie (UWB) organized the "1st Czech speech/NLP day" at CIIRC CTU Prague. It attracted more than 100 participants from both academic and industrial speech and NLP labs, the event was financially supported by the Czech speech and NLP indutry. Kyunghyun Cho (New York University and Facebook AI Research)[55] was invited as Keynote speaker.
- In 2021, the event was replaced by Interspeech 2021[56] in Brno – the most important speech conference, organized by te BUT team, (with detailed description in the interim report) that attracted most of the Czech speech and NLP community.
- The 2nd edition "Czech speech/NLP day" took place in Semilasso Cultural Center close to FIT BUT on October 12 2023. It attracted more than 110 participants from both academic and industrial speech and NLP labs, and guests from Slovakia. The event was again financially supported by the Czech speech and NLP industry. An esteemed guest, Dr. Bhuvana Ramabhadran[57] from Google Research was invited as keynote speaker, with a talk "Multilingual Speech Recognition and Synthesis".

The team plans continue these events in 2025 and bi-annually beyond, with 2025 edition targeting Pilsen in West Bohemia.

Figure 2: 2023 Czech speech/NLP community gathering at Semilasso Center in Brno.

6.3.2 Organizations and efforts supporting AI

Team members are involved in community efforts supporting the R&D, business and education in artificial intelligence, of which speech and NLP are important components.

[55] https://kyunghyuncho.me/
[56] https://www.interspeech2021.org/proceedings
[57] https://research.google/people/bhuvanaramabhadran

In 2019, several team members were at the founding of AICzechia[58] — an open initiative supporting Czech labs and teams in the AI area. AICzechia includes representatives of mainly academia, and lobbies for recognition of AI by the political representation. Team members contributed to the write-up speech and NLP R&D survey that now forms part of the National AI strategy adopted by the Czech Government in 2019 and updated in 2024.[59]

The team members are also active in the efforts of regional organizations supporting AI in Prague[60] and Brno.[61]

6.4 Impact on teaching

We are aware of our responsibility to transmit the know-how to the young generation. In addition to the "education by research" practised with the PhD students present on the NEUREM3 team, the NEUREM3 staff is involved in under-graduate course:

At BUT, the PI of the project, Lukáš Burget, is responsible for specialization "Machine learning" and is involved in specialization "Sound, Speech and Natural Language Processing" of Master program "Information Technology and Artificial Intelligence"[62]. We are training our students in courses SUI (Artificial Intelligence and Machine Learning), SUR (Machine Learning and Recognition), BAYa (Bayesian Models for Machine Learning), ZRE (Speech signal processing), and others.

CUNI finished its LangTech project (2017-2022)[63] where both Ondřej Bojar and Pavel Pecina were involved. The project has modernized most of the classes on NLP taught by CUNI, including the class on machine translation by Ondřej Bojar[64] and it was also very beneficial for attracting foreign PhD students and for supporting CUNI PhD students in internships.

Under the header of prg.ai, Ondřej Bojar has initiated the idea of "prg.ai Minor".[65] While not an accredited degree or programme, prg.ai Minor fosters collaboration and permeability of Prague's AI schools for the best talents. To pass prg.ai Minor, students have to attend AI-related subjects at three of four involved faculties (two at Charles University, two at Czech Technical University) across a set of topic groups. One of the topic groups used to be Natural Language Processing, recently renamed to Perception because it includes also computer vision. Based on the advances in multimodal processing (incl. within NEUREM3), we expect some courses merging these modalities to appear in the future years. One recent addition to prg.ai Minor is the CUNI subject "NPFL140 Large Language Models"[66] taught since 2023 collectively by 9 teachers who also recruited from NEUREM3 team: Jindřich Helcl is one of the guarantors and Dominik Macháček and Peter Polák are on the team.

[58] https://www.aiczechia.cz/

[59] https://mpo.gov.cz/assets/en/business/2024/9/Narodni-strategie-umele-intelience-CR-2030.pdf

[60] https://prg.ai/; with Ondřej Bojar being one of the initiators of the whole idea

[61] https://www.brno.ai/

[62] https://www.fit.vut.cz/applicants/degree-programme/.en#mgr

[63] https://ufal.mff.cuni.cz/grants/langtech

[64] https://ufal.mff.cuni.cz/courses/npfl087

[65] https://prg.ai/minor/

[66] https://ufal.mff.cuni.cz/courses/npfl140

7 ERC Proposals

A proposal to the European Research Council (ERC) was a required output of the project. The team submitted two such proposals. While none of them was accepted in this very competitive funding scheme competition, the preparation of the proposals on its own was very instrumental for both proposers, forcing them to formulate their long-term research visions.

7.1 Pinpointing Language Understanding – PILAU

- ERC scheme and year: Synergy 2024
- Corresponding Principal Investigator (cPI): Ondřej Bojar
- Corresponding Host Institution: Charles University
- Other Principal investigators (PIs): Anders Søgaard
- Host institutions: University of Copenhagen
- Proposal duration: 72 months
- Result of evaluation: B

Proposal summary: Large language models (LLMs) exhibit impressive performance, but do they really understand? This question has important consequences for explainability, AI alignment, and AI risks, but surprisingly, there is widespread disagreement about whether large language models exhibit understanding or not. Some of the most widely cited scholars in computer science, linguistics, cognitive science, and philosophy, have written about this, but in favour of radically different views. Some (Bender, Marcus, Chiang) have argued that language models do not exhibit *any* form of understanding. Others (Tenenbaum, Hill, Linzen, Butlin) have argued that language models exhibit *some* forms of understanding, but disagreed about which forms that would be. Finally, some (Cappelen, Dever, King) have argued that it is perfectly possible that language models exhibit *near-complete* understanding in a way that is very similar to humans. Evaluation against benchmarks such as BigBench, designed to evaluate the performance of language models across a wide range of tasks, has been deemed insufficient by most scholars, since it can be difficult to rule out trivial solutions to such tasks (e.g. Ned Block's Giant Lookup Table). PILAU will:

(i) help the community arrive at a common, operational **definition** of understanding;

(ii) redesign **evaluation** of current language models to ensure its scientific soundness and maintain discerning power in black box settings when applied to Internet-scale models, and to humans;

(iii) **develop** ways to explicitly optimize for human-like understanding, targeting alignment with human values, trustworthiness, true reflection of model's abilities and considering natural human limits on processing.

The project is a synergy between Charles University and the University of Copenhagen, but it will include a wide portfolio of community efforts – through surveys, interviews, academic conferences, and anthologies – to bring people together around this extremely important problem.

7.2 Aligning pre-trained models via an interpretable latent space for robust Artificial Intelligence – ALPINE -AI

- ERC scheme and year: Starting 2024
- Corresponding Principal Investigator (cPI): Santosh Kesiraju
- Corresponding Host Institution: BUT
- Proposal duration: 60 months
- Result of evaluation: C

Proposal summary: The usage of large pre-trained models has become ubiquitous in several fields of Artificial Intelligence (AI). The recent developments and capabilities of large language models are a prime example. Similar trends are seen in areas such as speech technology, computer vision, and across disciplines related to medicine and healthcare. In speech and language processing, current state-of-the-art models are trained independent of each other, and a majority of them are uni-modal at their input. Whereas, a number of applications such as spoken language translation, task-oriented dialogue systems and atypical speech assessment either require or benefit from a careful combination of two or more models. A naive way of building a cascade pipeline results in error propagation and compounding, while joint-training causes catastrophic forgetting, where the benefits of pre-training diminish. Combined with these limitations, the black-box nature of the models make them hard to interpret; moreover, they propagate harmful biases acquired from the massive web-crawled training data. To overcome these limitations of the current state-of-the-art, ALPINE -AI aims to develop theoretically-motivated methods for aligning any arbitrary pre-trained models via an interpretable latent space. The alignment will enable to join the models without requiring to fine-tune them. The interpretable latent space will ease the study and identification of the linguistic, para-linguistic, and fairness attributes that are encoded in the pre-trained models. This will also allow the explainability of the models' output in human-centred applications related to medicine and healthcare such as atypical speech and language assessment. The shared latent space enables to use efficient data augmentation and bias mitigation methods that will enhance the robustness of speech and language applications. The project has a clear potential to lead ground-breaking results that will positively impact several areas in AI.

8 Publications

2019 Publications, reported in interim report

[Ala+19] Jahangir Alam et al. "ABC System Description for NIST Multimedia Speaker Recognition Evaluation 2019". In: *Proceedings of NIST 2019 SRE Workshop*. Sentosa, Singapore, SG, 2019, pp. 1–7. URL: https://www.fit.vut.cz/research/publication/12164.

[Bar+19] Loïc Barrault et al. "Findings of the 2019 Conference on Machine Translation (WMT19)". In: *Fourth Conference on Machine Translation - Proceedings of the Conference*. Ed. by Ondřej Bojar. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 1–61. ISBN: 978-1-950737-27-7. URL: https://aclanthology.org/W19-5301/.

[BBW19] Ondřej Bojar, Raffaella Bernardi, and Bonnie L. Webber. "Representation of sentence meaning (A JNLE Special Issue)". In: *Natural Language Engineering* 25.4 (2019), pp. 427–432. ISSN: 1351-3249. URL: http://ufal.mff.cuni.cz/biblio/attachments/2019-bojar-m1285039693733276521.pdf.

[ÇB19a] Erion Çano and Ondřej Bojar. "Efficiency Metrics for Data-Driven Models: A Text Summarization Case Study". In: *Proceedings of the 12th International Conference on Natural Language Generation (INLG 2019)*. Association for Computational Linguistics. Stroudsubrgh, PA, USA: Association for Computational Linguistics, 2019, pp. 229–239. ISBN: 978-1-950737-94-9. URL: https://www.aclweb.org/anthology/W19-8630.

[ÇB19b] Erion Çano and Ondřej Bojar. "Keyphrase Generation: A Text Summarization Struggle". In: *The 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Hyatt Regency Hotel). NAACL-HLT 2019. Minneapolis, USA: NAACL-HLT 2019, 2019, pp. 666–672. ISBN: 978-1-950737-13-0. URL: https://www.aclweb.org/anthology/N19-1070.

[Die+19] Mireia Diez et al. "Bayesian HMM based x-vector clustering for Speaker Diarization". In: *Proceedings of Interspeech*. Vol. 2019. 9. Graz, AT, 2019, pp. 346–350. DOI: 10.21437/Interspeech.2019-2813. URL: https://www.fit.vut.cz/research/publication/12085.

[HLP19] Jindřich Helcl, Jindřich Libovický, and Martin Popel. "CUNI System for the WMT19 Robustness Task". In: *Fourth Conference on Machine Translation - Proceedings of the Conference*. Ed. by Ondřej Bojar. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 738–742. ISBN: 978-1-950737-27-7. URL: https://www.aclweb.org/anthology/W19-5364/.

[KCB19] Daniel Kondratyuk, Ronald Cardenas, and Ondřej Bojar. "Replacing Linguists with Dummies: A Serious Need for Trivial Baselinesin Multi-Task Neural Machine Translation". In: *The Prague Bulletin of Mathematical Linguistics* 113 (2019), pp. 31–40. ISSN: 0032-6585. URL: https://ufal.mff.cuni.cz/pbml/113/art-kondratyuk-cardenas-bojar.pdf.

[Ma+19] Qingsong Ma et al. "Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges". In: *Fourth Conference on Machine Translation - Proceedings of the Conference*. Ed. by Ondřej Bojar. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 62–90. ISBN: 978-1-950737-27-7. URL: https://aclanthology.org/W19-5302/.

[Mac+19] Dominik Macháček et al. "A Speech Test Set of Practice Business Presentations with Additional Relevant Texts". In: *Lecture Notes in Artificial Inteligence, Statistical Language and Speech Processing* (Jožef Stefan Institut, Ljubljana). Lecture Notes in Computer Science 11816. IRDTA. Cham, Switzerland: Springer Nature Switzerland AG, 2019, pp. 151–161. ISBN: 978-3-030-31371-5. URL: https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3023.

[Mat+19] Pavel Matějka et al. "Analysis of BUT Submission in Far-Field Scenarios of VOiCES 2019 Challenge". In: *Proceedings of Interspeech*. Vol. 2019. 9. Graz, AT, 2019, pp. 2448–2452. DOI: 10.21437/Interspeech.2019-2471. URL: https://www.fit.vut.cz/research/publication/12090.

[Nak+19] Toshiaki Nakazawa et al. "Overview of the 6th Workshop on Asian Translation". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 1–35. ISBN: 978-1-950737-90-1. URL: https://www.aclweb.org/anthology/D19-5201.pdf.

[NB19] Anna Nedoluzhko and Ondřej Bojar. "Towards Automatic Minuting of Meetings". In: *Proceedings of the 19th Conference ITAT 2019: Slovenskočeský NLP workshop (SloNLP 2019)*. Ed. by Petra Barančíková et al. Vol. 2473. CEUR Workshop Proceedings. P.J.Šafárik University in Košice. Košice, Slovakia: CreateSpace Independent Publishing Platform, 2019, pp. 112–119. URL: http://ceur-ws.org/Vol-2473/.

[Nov+19a] Ondřej Novotný et al. "Analysis of DNN Speech Signal Enhancement for Robust Speaker Recognition". In: *Computer Speech and Language* 2019.58 (2019), pp. 403–421. ISSN: 0885-2308. DOI: 10.1016/j.csl.2019.06.004. URL: https://www.fit.vut.cz/research/publication/12039.

[Nov+19b] Ondřej Novotný et al. "Factorization of Discriminatively Trained i-Vector Extractor for Speaker Recognition". In: *Proceedings of Interspeech*. Vol. 2019. 9. Graz, AT, 2019, pp. 4330–4334. DOI: 10.21437/Interspeech.2019-1757. URL: https://www.fit.vut.cz/research/publication/12091.

[Ond+19] Francois Antoine Lucas Ondel et al. “Bayesian Subspace Hidden Markov Model for Acoustic Unit Discovery”. In: *Proceedings of Interspeech 2019*. Vol. 2019. 9. Graz, AT, 2019, pp. 261–265. DOI: 10.21437/Interspeech.2019-2224. URL: https://www.fit.vut.cz/research/publication/12084.

[Pal+19] Shruti Palaskar et al. “Multimodal Abstractive Summarization for How2 Videos”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 6587–6596. ISBN: 978-1-950737-48-2. URL: https://aclanthology.org/P19-1659/.

[PBD19] Shantipriya Parida, Ondřej Bojar, and Satya Dash. “Hindi Visual Genome: A Dataset for Multimodal English-to-Hindi Machine Translation”. In: *Computación y Sistemas* 23.4 (2019), pp. 1499–1505. ISSN: 1405-5546. URL: https://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/3294.

[PMB19a] Shantipriya Parida, Petr Motlíček, and Ondřej Bojar. “Idiap NMT System for WAT 2019 Multi-Modal Translation Task”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 175–180. ISBN: 978-1-950737-90-1. URL: https://www.aclweb.org/anthology/D19-5223.pdf.

[PMB19b] Thuong-Hai Pham, Dominik Macháček, and Ondřej Bojar. “Promoting the Knowledge of Source Syntax in Transformer NMT Is Not Needed”. In: *Computación y Sistemas* 23.3 (2019), pp. 923–934. ISSN: 1405-5546. URL: https://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/3265/2708.

[Pop+19] Martin Popel et al. “English-Czech Systems in WMT19: Document-Level Transformer”. In: *Fourth Conference on Machine Translation - Proceedings of the Conference*. Ed. by Ondřej Bojar. 2. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 342–348. ISBN: 978-1-950737-27-7. URL: http://www.statmt.org/wmt19/pdf/53/WMT37.pdf.

[Roh+19] A. Johan Rohdin et al. “End-to-end DNN based text-independent speaker recognition for long and short utterances”. In: *Computer Speech and Language* 2020.59 (2019), pp. 22–35. ISSN: 0885-2308. DOI: 10.1016/j.csl.2019.06.002. URL: https://www.fit.vut.cz/research/publication/12038.

[SP19] Shadi Saleh and Pavel Pecina. “Term Selection for Query Expansion in Medical Cross-Lingual Information Retrieval”. In: *Advances in Information Retrieval; 41st European Conference on IR Research, ECIR 2019*. Ed. by Leif Azzopardi et al. 11438 1. Springer. Berlin, Germany: Springer Interna-

tional Publishing, 2019, pp. 507–522. ISBN: 978-3-030-15719-7. URL: https://link.springer.com/chapter/10.1007/978-3-030-15712-8_33.

[Sta+19] Themos Stafylakis et al. "Self-supervised speaker embeddings". In: *Proceedings of Interspeech*. Vol. 2019. 9. Graz, AT, 2019, pp. 2863–2867. DOI: 10.21437/Interspeech.2019-2842. URL: https://www.fit.vut.cz/research/publication/12092.

[VB19a] Dušan Variš and Ondřej Bojar. "Unsupervised Pretraining for Neural Machine Translation Using Elastic Weight Consolidation". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 130–135. ISBN: 978-1-950737-47-5. URL: https://www.aclweb.org/anthology/P19-2017.pdf.

[VB19b] Dušan Variš and Ondřej Bojar. "Unsupervised Pretraining for Neural Machine Translation Using Elastic Weight Consolidation". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Ed. by Fernando Alva-Manchego, Eunsol Choi, and Daniel Khashabi. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 130–135. DOI: 10.18653/v1/P19-2017. URL: https://aclanthology.org/P19-2017/.

[Wan+19] Shuai Wang et al. "On the Usage of Phonetic Information for Text-independent Speaker Embedding Extraction". In: *Proceedings of Interspeech*. Vol. 2019. 9. Graz, AT, 2019, pp. 1148–1152. DOI: 10.21437/Interspeech.2019-3036. URL: https://www.fit.vut.cz/research/publication/12087.

[ZČB19] Hossein Zeinali, Jan Černocký, and Lukáš Burget. "A multi purpose and large scale speech corpus in Persian and English for speaker and speech Recognition: the DeepMine database". In: *IEEE Automatic Speech Recognition and Understanding Workshop - Proceedings (ASRU)*. Sentosa, Singapore, SG, 2019, pp. 397–402. ISBN: 978-1-7281-0306-8. DOI: 10.1109/ASRU46091.2019.9003882. URL: https://www.fit.vut.cz/research/publication/12153.

[Zei+19] Hossein Zeinali et al. "BUT System Description to VoxCeleb Speaker Recognition Challenge 2019". In: *Proceedings of The VoxCeleb Challange Workshop 2019*. Graz, AT, 2019, pp. 1–4. URL: https://www.fit.vut.cz/research/publication/12224.

[Žmo+19] Kateřina Žmolíková et al. "SpeakerBeam: Speaker Aware Neural Network for Target Speaker Extraction in Speech Mixtures". In: *IEEE Journal of Selected Topics in Signal Processing* 13.4 (2019), pp. 800–814. ISSN: 1932-4553. DOI: 10.1109/JSTSP.2019.2922820. URL: https://www.fit.vut.cz/research/publication/12078.

2020 Publications, reported in interim report

[Ala+20] Jahangir Alam et al. "Analysis of ABC Submission to NIST SRE 2019 CMN and VAST Challenge". In: *Proceedings of Odyssey 2020 The Speaker and Language Recognition Workshop*. Vol. 2020. 11. Tokyo, JP, 2020, pp. 289–295. DOI: 10.21437/Odyssey.2020-41. URL: https://www.fit.vut.cz/research/publication/12292.

[Ans+20] Ebrahim Ansari et al. "FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN". In: *Proceedings of the 17th International Conference on Spoken Language Translation*. Ed. by Marcello Federico et al. ACL. Online: Association for Computational Linguistics, 2020, pp. 1–34. ISBN: 978-1-952148-07-1. URL: https://www.aclweb.org/anthology/2020.iwslt-1.1.pdf.

[BB20a] Petra Barančíková and Ondřej Bojar. "COSTRA 1.0: A Dataset of Complex Sentence Transformations". In: *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)* (Le Palais du Pharo). Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, 2020, pp. 3535–3541. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.434.

[BB20b] Petra Barančíková and Ondřej Bojar. "Costra 1.1: An Inquiry into Geometric Properties of Sentence Spaces". In: *Lecture Notes in Artificial Intelligence, 23rd International Conference on Text, Speech and Dialogue*. Ed. by Petr Sojka et al. Lecture Notes in Computer Science. Faculty of Informatics, Masaryk University Brno. Cham, Switzerland: Springer, 2020, pp. 135–143. ISBN: 978-3-030-58322-4. DOI: 10.1007/978-3-030-58323-1_14.

[Bar+20] Loïc Barrault et al. "Findings of the 2020 Conference on Machine Translation (WMT20)". In: *Fifth Conference on Machine Translation - Proceedings of the Conference*. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 1–55. ISBN: 978-1-948087-81-0. URL: http://www.statmt.org/wmt20/pdf/2020.wmt-1.1.pdf.

[Bur+20] Lukáš Burget et al. "BUT System Description to SdSV Challenge 2020". In: *Proceedings of Short-duration Speaker Verification Challenge 2020 Workshop*. Shanghai, on-line event of Interspeech 2020 Conference, CN, 2020, pp. 1–5. URL: https://www.fit.vut.cz/research/publication/12481.

[ÇB20a] Erion Çano and Ondřej Bojar. "How Many Pages? Paper Length Prediction from the Metadata". In: *4th International Conference on Natural Language Processing and Information Retrieval*. ACM. New York, USA: ACM, 2020, pp. 91–95. ISBN: 978-1-4503-7760-7. URL: https://dl.acm.org/doi/10.1145/3443279.3443305.

[ÇB20b] Erion Çano and Ondřej Bojar. "Two Huge Title and Keyword Generation Corpora of Research Articles". In: *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)* (Le Palais du Pharo). Ed. by Nicoletta Calzolari et al. Marseille, France: European

Language Resources Association, 2020, pp. 6663–6671. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.823.

[Die+20] Mireia Diez et al. "Analysis of Speaker Diarization based on Bayesian HMM with Eigenvoice Priors". In: *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING* 28.1 (2020), pp. 355–368. ISSN: 2329-9290. DOI: 10.1109/TASLP.2019.2955293. URL: https://www.fit.vut.cz/research/publication/12139.

[KAB20a] Hadi Abdi Khojasteh, Ebrahim Ansari, and Mahdi Bohlouli. *Large-Scale Colloquial Persian 0.5*. 2020. URL: https://iasbs.ac.ir/~ansari/lscp/.

[KAB20b] Hadi Abdi Khojasteh, Ebrahim Ansari, and Mahdi Bohlouli. "LSCP: Enhanced Large Scale Colloquial Persian Language Understanding". In: *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)* (Le Palais du Pharo). Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, 2020, pp. 6323–6327. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.776.

[KPB20] Jonáš Kratochvíl, Peter Polák, and Ondřej Bojar. "Large Corpus of Czech Parliament Plenary Hearings". In: *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)* (Le Palais du Pharo). Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, 2020, pp. 6363–6367. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.781/.

[KKB20] Ivana Kvapilíková, Tom Kocmi, and Ondřej Bojar. "CUNI Systems for the Unsupervised and Very Low Resource Translation Task in WMT20". In: *Fifth Conference on Machine Translation - Proceedings of the Conference*. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 1123–1128. ISBN: 978-1-948087-81-0. URL: https://aclanthology.org/2020.wmt-1.133.pdf.

[Kva+20] Ivana Kvapilíková et al. "Unsupervised Multilingual Sentence Embeddings for Parallel Corpus Mining". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Ed. by Shruti Rijhwani et al. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 255–262. ISBN: 978-1-952148-03-3. URL: https://www.aclweb.org/anthology/2020.acl-srw.34/.

[Lib+20] Jindřich Libovický et al. "Expand and Filter: CUNI and LMU Systems for the WNGT 2020 Duolingo Shared Task". In: *Proceedings of the Fourth Workshop on Neural Generation and Translation*. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 153–160. ISBN: 978-1-952148-17-0. URL: https://www.aclweb.org/anthology/2020.ngt-1.18/.

[Loz+20] Alicia Díez Lozano et al. “BUT Text-Dependent Speaker Verification System for SdSV Challenge 2020”. In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 2020. 10. Shanghai, CN, 2020, pp. 761–765. DOI: 10.21437/Interspeech.2020-2882. URL: https://www.fit.vut.cz/research/publication/12378.

[Mac+20] Dominik Macháček et al. “ELITR Non-Native Speech Translation at IWSLT 2020”. In: *Proceedings of the 17th International Conference on Spoken Language Translation*. Ed. by Marcello Federico et al. ACL. Online: Association for Computational Linguistics, 2020, pp. 200–208. ISBN: 978-1-952148-07-1. URL: https://www.aclweb.org/anthology/2020.iwslt-1.25.pdf.

[Mat+20a] Pavel Matějka et al. “13 years of speaker recognition research at BUT, with longitudinal analysis of NIST SRE”. In: *Computer Speech and Language* 2020.63 (2020), pp. 1–15. ISSN: 0885-2308. DOI: 10.1016/j.csl.2019.101035. URL: https://www.fit.vut.cz/research/publication/12211.

[Mat+20b] Nitika Mathur et al. “Results of the WMT20 Metrics Shared Task”. In: *Fifth Conference on Machine Translation - Proceedings of the Conference*. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 688–725. ISBN: 978-1-948087-81-0. URL: http://www.statmt.org/wmt20/pdf/2020.wmt-1.77.pdf.

[Nak+20] Toshiaki Nakazawa et al. “Overview of the 7th Workshop on Asian Translation”. In: *Proceedings of the 7th Workshop on Asian Translation (WAT2020)*. AACL-IJCNLP. Stroudsburg, USA: Association for Computational Linguistics, 2020, pp. 1–44. URL: https://www.aclweb.org/anthology/2020.wat-1.1/.

[Par+20] Shantipriya Parida et al. “ODIANLP’s Participation in WAT2020”. In: *Proceedings of the 7th Workshop on Asian Translation (WAT2020)*. AACL-IJCNLP. Stroudsburg, USA: Association for Computational Linguistics, 2020, pp. 103–108. URL: https://www.aclweb.org/anthology/2020.wat-1.10/.

[Pol+20] Peter Polák et al. “CUNI Neural ASR with Phoneme-Level Intermediate Step for Non-Native SLT at IWSLT 2020”. In: *Proceedings of the 17th International Conference on Spoken Language Translation*. Ed. by Marcello Federico et al. ACL. Online: Association for Computational Linguistics, 2020, pp. 191–199. ISBN: 978-1-952148-07-1. URL: https://www.aclweb.org/anthology/2020.iwslt-1.24.

[Pop+20] Martin Popel et al. “Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals”. In: *Nature Communications* 11.4381 (2020), pp. 1–15. ISSN: 2041-1723. URL: https://doi.org/10.1038/s41467-020-18073-9.

[SP20] Shadi Saleh and Pavel Pecina. "Document Translation vs. Query Translation for Cross-Lingual Information Retrieval in the Medical Domain". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 6849–6860. ISBN: 978-1-952148-25-5. URL: https://aclanthology.org/2020.acl-main.613.pdf.

[Sil+20] Anna Silnova et al. "Probabilistic embeddings for speaker diarization". In: *Proceedings of Odyssey 2020 The Speaker and Language Recognition Workshop*. Vol. 2020. 11. Tokyo, JP, 2020, pp. 24–31. DOI: 10.21437/Odyssey.2020-4. URL: https://www.fit.vut.cz/research/publication/12288.

[Wan+20] Shuai Wang et al. "Investigation of Specaugment for Deep Speaker Embedding Learning". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Barcelona, ES, 2020, pp. 7139–7143. ISBN: 978-1-5090-6631-5. DOI: 10.1109/ICASSP40776.2020.9053481. URL: https://www.fit.vut.cz/research/publication/12278.

[ZB20] Vilém Zouhar and Ondřej Bojar. "Outbound Translation User Interface Ptakopet: A Pilot Study". In: *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)* (Le Palais du Pharo). Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, 2020, pp. 6967–6975. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.860.

[ZVB20] Vilém Zouhar, Tereza Vojtěchová, and Ondřej Bojar. "WMT20 Document-Level Markable Error Exploration". In: *Fifth Conference on Machine Translation - Proceedings of the Conference*. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 371–380. ISBN: 978-1-948087-81-0. URL: https://aclanthology.org/2020.wmt-1.41/.

[Zul+20] Juan Zuluaga-Gomez et al. "Automatic Speech Recognition Benchmark for Air-Traffic Communications". In: *Proceedings of Interspeech 2020*. Vol. 2020. 10. Shanghai, CN, 2020, pp. 2297–2301. DOI: 10.21437/Interspeech.2020-2173. URL: https://www.fit.vut.cz/research/publication/12404.

2021 Publications, reported in interim report

[Akh+21a] Farhad Akhbardeh et al. "Findings of the 2021 Conference on Machine Translation (WMT21)". In: *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics. Online: Association for Computational Linguistics, 2021, pp. 1–88. ISBN: 978-1-954085-94-7. URL: https://aclanthology.org/2021.wmt-1.1.pdf.

[Ana+21] Antonios Anastasopoulos et al. “FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN”. In: *Proceedings of the 18th International Conference on Spoken Language Translation*. ACL. Stroudsburg, USA: Association for Computational Linguistics, 2021, pp. 1–29. ISBN: 978-1-954085-74-9. URL: https://aclanthology.org/2021.iwslt-1.1/.

[Ans+21] Ebrahim Ansari et al. “SLTev: Comprehensive Evaluation of Spoken Language Translation”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Ed. by Dimitra Gkatzia and Djamé Seddah. EACL. Stroudsburg, PA, USA: Association for Computational Linguistics (ACL), 2021, pp. 71–79. ISBN: 978-1-954085-05-3. URL: https://aclanthology.org/2021.eacl-demos.9.

[AP21] Michal Auersperger and Pavel Pecina. “Solving SCAN Tasks with Data Augmentation and Input Embeddings”. In: *Proceedings of the Recent Advances in Natural Language Processing*. INCOMA Ltd. Shoumen, Bulgaria: INCOMA Ltd., 2021, pp. 86–91. ISBN: 978-954-452-072-4. URL: https://aclanthology.org/2021.ranlp-main.11.pdf.

[BVB21] Niyati Bafna, Martin Vastl, and Ondřej Bojar. “Constrained Decoding for Technical Term Retention in English-Hindi MT”. Silchar, India, 2021.

[Bas+21] K. Murali Baskar et al. “Eat: Enhanced ASR-TTS for Self-Supervised Speech Recognition”. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, Ontario, CA, 2021, pp. 6753–6757. ISBN: 978-1-7281-7605-5. DOI: 10.1109/ICASSP39728.2021.9413375. URL: https://www.fit.vut.cz/research/publication/12524.

[BB21] Karel Beneš and Lukáš Burget. “Text Augmentation for Language Models in High Error Recognition Scenario”. In: *Proceedings Interspeech 2021*. Vol. 2021. 8. Brno, CZ, 2021, pp. 1872–1876. DOI: 10.21437/Interspeech.2021-627. URL: https://www.fit.vut.cz/research/publication/12606.

[Ego+21] Ekaterina Egorova et al. “Out-of-Vocabulary Words Detection with Attention and CTC Alignments in an End-to-End ASR System”. In: *Proceedings Interspeech 2021*. Vol. 2021. 8. Brno, CZ, 2021, pp. 2901–2905. DOI: 10.21437/Interspeech.2021-1756. URL: https://www.fit.vut.cz/research/publication/12608.

[Fre+21] Markus Freitag et al. “Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain”. In: *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics. Online: Association for Computational Linguistics, 2021, pp. 733–774. ISBN: 978-1-954085-94-7. URL: https://aclanthology.org/2021.wmt-1.73/.

[Geb+21] Petr Gebauer et al. “CUNI Systems in WMT21: Revisiting Backtranslation Techniques for English-Czech NMT”. In: *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics. Online: Association for Computational Linguistics, 2021, pp. 123–129. ISBN: 978-1-954085-94-7. URL: https://aclanthology.org/2021.wmt-1.7/.

[HB21] Michael Hanna and Ondřej Bojar. “A Fine-Grained Analysis of BERTScore”. In: *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics. Online: Association for Computational Linguistics, 2021, pp. 507–517. ISBN: 978-1-954085-94-7. URL: https://aclanthology.org/2021.wmt-1.59.pdf.

[HM21] Michael Hanna and David Mareček. “Analyzing BERT’s Knowledge of Hypernymy via Prompting”. In: *Proceedings of the 4th Workshop on Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 275–282. ISBN: 978-1-955917-06-3. URL: https://aclanthology.org/2021.blackboxnlp-1.20.pdf.

[Jon+21a] Josef Jon et al. “CUNI systems for WMT21: Multilingual Low-Resource Translation for Indo-European Languages Shared Task”. In: *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics. Online: Association for Computational Linguistics, 2021, pp. 354–361. ISBN: 978-1-954085-94-7. URL: https://aclanthology.org/2021.wmt-1.42/.

[Jon+21b] Josef Jon et al. “CUNI systems for WMT21: Terminology translation Shared Task”. In: *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics. Online: Association for Computational Linguistics, 2021, pp. 828–834. ISBN: 978-1-954085-94-7. URL: https://aclanthology.org/2021.wmt-1.82/.

[Jon+21c] Josef Jon et al. “End-to-End Lexically Constrained Machine Translation for Morphologically Rich Languages”. In: *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 4019–4033. ISBN: 978-1-954085-52-7. URL: https://aclanthology.org/2021.acl-long.311.

[Kar+21] Martin Karafiát et al. “Analysis of X-Vectors for Low-Resource Speech Recognition”. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, Ontario, CA, 2021, pp. 6998–7002. ISBN: 978-1-7281-7605-5. DOI: 10.1109/ICASSP39728.2021.9414725. URL: https://www.fit.vut.cz/research/publication/12525.

[KBH21] Martin Kišš, Karel Beneš, and Michal Hradiš. “AT-ST: Self-Training Adaptation Strategy for OCR in Domains with Limited Transcriptions”. In: *Lladós J., Lopresti D., Uchida S. (eds) Document Analysis and Recognition - ICDAR 2021*. Lecture Notes in Computer Science. Lausanne, CH, 2021,

pp. 463–477. ISBN: 978-3-030-86336-4. DOI: 10.1007/978-3-030-86337-1_31. URL: https://www.fit.vut.cz/research/publication/12464.

[KBP21] Věra Kloudová, Ondřej Bojar, and Martin Popel. "Detecting Post-edited References and Their Effect on Human Evaluation". In: *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*. EACL 2021. Stroudsburg, USA: Association for Computational Linguistics, 2021, pp. 114–119. ISBN: 978-1-954085-10-7. URL: https://aclanthology.org/2021.humeval-1.13.pdf.

[Koc+21] Martin Kocour et al. "BCN2BRNO: ASR System Fusion for Albayzin 2020 Speech to Text Challenge". In: *Proceedings of IberSPEECH 2021*. Vallaloid, ES, 2021, pp. 113–117. DOI: 10.21437/IberSPEECH.2021-24. URL: https://www.fit.vut.cz/research/publication/12577.

[Kop+21a] Matyáš Kopp et al. *ParCzech 3.0*. Prague, Czechia, 2021. URL: http://hdl.handle.net/11234/1-3631.

[Kop+21b] Matyáš Kopp et al. "ParCzech 3.0: A Large Czech Speech Corpus with Rich Metadata". In: *Lecture Notes in Artificial Intelligence, 24th International Conference on Text, Speech and Dialogue*. Ed. by Kamil Ekštein, František Pártl, and Miroslav Konopík. Vol. 12848. Lecture Notes in Computer Science. University of West Bohemia. Cham, Switzerland: Springer, 2021, pp. 293–304. ISBN: 978-3-030-83526-2. URL: https://link.springer.com/content/pdf/10.1007%5C%2F978-3-030-83527-9_25.pdf.

[KB21] Ivana Kvapilíková and Ondřej Bojar. "Machine Translation of Covid-19 Information Resources via Multilingual Transfer". In: *ITAT 2021 2nd Workshop on Automata, Formal and Natural Languages – WAFNL 2021* (Hotel Hel'pa). Ed. by František Mráz, Dana Pardubská, and Martin Plátek. MFF UK. Praha, Czechia: Faculty of Mathematics and Physics, 2021, pp. 176–181. URL: https://ics.upjs.sk/~antoni/ceur-ws.org/Vol-0000/paper26.pdf.

[Lan+21a] Federico Landini et al. "Analysis of the BUT Diarization System for Voxconverse Challenge". In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, Ontario, CA, 2021, pp. 5819–5823. ISBN: 978-1-7281-7605-5. DOI: 10.1109/ICASSP39728.2021.9414315. URL: https://www.fit.vut.cz/research/publication/12520.

[Lan+21b] Federico Landini et al. "BUT System Description for The Third DIHARD Speech Diarization Challenge". In: *Proceedings available at Dihard Challenge Github*. on-line by LDC and University of Pennsylvania, US, 2021, pp. 1–5. URL: https://www.fit.vut.cz/research/publication/12478.

[MŽB21] Dominik Macháček, Matúš Žilinec, and Ondřej Bojar. "Lost in Interpreting: Speech Translation from Source or Interpreter?" In: *Proceedings of INTERSPEECH 2021*. ISCA. Baxas, France: ISCA, 2021, pp. 2376–2380. DOI: 10.21437/Interspeech.2021-2232.

[MP21] Jiří Mayer and Pavel Pecina. "Synthesizing Training Data for Handwritten Music Recognition". In: *Document Analysis and Recognition – ICDAR 2021*. Ed. by Josep Lladós, Daniel Lopresti, and Uchida Seiichi. Vol. 12823. Lecture Notes in Computer Science. University of Applied Sciences and Arts Western Switzerland. Cham, Switzerland: Springer International Publishing, 2021, pp. 626–641. ISBN: 978-3-030-86333-3. URL: https://link.springer.com/content/pdf/10.1007%5C%2F978-3-030-86334-0.pdf.

[Nak+21] Toshiaki Nakazawa et al. "Overview of the 8th Workshop on Asian Translation". In: *Proceedings of the 8th Workshop on Asian Translation*. AACL-IJCNLP. Stroudsburg, USA: Association for Computational Linguistics, 2021, pp. 1–45. URL: https://aclanthology.org/2021.wat-1.1/.

[Pen+21a] Junyi Peng et al. "Effective Phase Encoding for End-To-End Speaker Verification". In: *Proceedings Interspeech 2021*. Vol. 2021. 8. Brno, CZ, 2021, pp. 2366–2370. DOI: 10.21437/Interspeech.2021-2025. URL: https://www.fit.vut.cz/research/publication/12607.

[Pen+21b] Junyi Peng et al. "ICSpk: Interpretable Complex Speaker Embedding Extractor from Raw Waveform". In: *Proceedings Interspeech 2021*. Vol. 2021. 8. Brno, CZ, 2021, pp. 511–515. DOI: 10.21437/Interspeech.2021-2016. URL: https://www.fit.vut.cz/research/publication/12597.

[PB21] Peter Polák and Ondřej Bojar. "Coarse-To-Fine And Cross-Lingual ASR Transfer". In: *ITAT 2021 2nd Workshop on Automata, Formal and Natural Languages – WAFNL 2021* (Hotel Hel'pa). Ed. by František Mráz, Dana Pardubská, and Martin Plátek. MFF UK. Praha, Czechia: Faculty of Mathematics and Physics, 2021, pp. 154–160. URL: https://ics.upjs.sk/~antoni/ceur-ws.org/Vol-0000/paper09.pdf.

[PSB21] Peter Polák, Muskaan Singh, and Ondřej Bojar. "Explainable Quality Estimation: CUNI Eval4NLP Submission". In: *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 250–255. URL: https://aclanthology.org/2021.eval4nlp-1.24.pdf.

[SGB21] Muskaan Singh, Tirthankar Ghosal, and Ondřej Bojar. "An Empirical Performance Analysis of State-of-the-Art Summarization Models for Automatic Minuting" (Shanghai International Studies University). 209 N. Eighth Street, Stroudsburg PA 18360, USA, 2021.

[SRB21] Themos Stafylakis, A. Johan Rohdin, and Lukáš Burget. "Speaker embeddings by modeling channel-wise correlations". In: *Proceedings Interspeech 2021*. Vol. 2021. 8. Brno, CZ, 2021, pp. 501–505. DOI: 10.21437/Interspeech.2021-1442. URL: https://www.fit.vut.cz/research/publication/12596.

[VB21] Dušan Variš and Ondřej Bojar. "Sequence Length is a Domain: Length-based Overfitting in Transformer Models". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 8246–8257. ISBN: 978-1-955917-09-4. URL: https://aclanthology.org/2021.emnlp-main.650.pdf.

[Vyd+21c] K. Hari Vydana et al. "Jointly Trained Transformers Models for Spoken Language Translation". In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, Ontario, CA, 2021, pp. 7513–7517. ISBN: 978-1-7281-7605-5. DOI: 10.1109/ICASSP39728.2021.9414159. URL: https://www.fit.vut.cz/research/publication/12522.

[Yus+21] Bolaji Yusuf et al. "A Hierarchical Subspace Model for Language-Attuned Acoustic Unit Discovery". In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, Ontario, CA, 2021, pp. 3710–3714. ISBN: 978-1-7281-7605-5. DOI: 10.1109/ICASSP39728.2021.9414899. URL: https://www.fit.vut.cz/research/publication/12523.

[Žmo+21] Kateřina Žmolíková et al. "Integration of Variational Autoencoder and Spatial Clustering for Adaptive Multi-Channel Neural Speech Separation". In: *Proceedings of SLT 2021*. Shenzhen - virtual , CN, 2021, pp. 889–896. ISBN: 978-1-7281-7066-4. DOI: 10.1109/SLT48900.2021.9383612. URL: https://www.fit.vut.cz/research/publication/12553.

[Zou21] Vilém Zouhar. "Sampling and Filtering of Neural Machine Translation Distillation Data". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2021, pp. 1–8. ISBN: 978-1-954085-50-3. URL: https://aclanthology.org/2021.naacl-srw.1.pdf.

[Zou+21a] Vilém Zouhar et al. "Backtranslation Feedback Improves User Confidence in MT, Not Quality". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 151–161. ISBN: 978-1-954085-46-6. URL: https://aclanthology.org/2021.naacl-main.14/.

[Zou+21b] Vilém Zouhar et al. "Neural Machine Translation Quality and Post-Editing Performance". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 10204–10214. ISBN: 978-1-955917-09-4. URL: https://aclanthology.org/2021.emnlp-main.801.pdf.

2021 Publications, not yet reported

[Gho+21] Tirthankar Ghosal et al. "Overview of the First Shared Task on Automatic Minuting (AutoMin) at Interspeech 2021". In: *First Shared Task on Automatic Minuting at Interspeech 2021*. 2021, pp. 1–25. DOI: 10.21437/AutoMin.2021-1.

[Sen+21] Arghyadeep Sen et al. "Bengali Visual Genome: A Multimodal Dataset for Machine Translation and Image Captioning". In: *9th International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA 2021)*. Vol. 266. Smart Innovation, Systems and Technologies (SIST). Singapore: Springer Nature Singapore, 2021, pp. 63–70. ISBN: 978-981-16-6624-7.

[Vyd+21b] Hari Krishna Vydana et al. "The IWSLT 2021 BUT Speech Translation Systems". In: *Proceedings of 18th International Conference on Spoken Language Translation (IWSLT)*. Bangkok, on-line, TH, 2021, pp. 75–83. ISBN: 978-1-7138-3378-9. DOI: 10.18653/v1/2021.iwslt-1.7. URL: https://www.fit.vut.cz/research/publication/12702.

2022 Publications

[Abd+22] Idris Abdulmumin et al. "Hausa Visual Genome: A Dataset for Multi-Modal English to Hausa Machine Translation". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, June 2022, pp. 6471–6479. URL: https://aclanthology.org/2022.lrec-1.694/.

[Ala+22] Jahangir Alam et al. "Development of ABC systems for the 2021 edition of NIST Speaker Recognition evaluation". In: *Proceedings of The Speaker and Language Recognition Workshop (Odyssey 2022)*. Beijing, CN, 2022, pp. 346–353. DOI: 10.21437/Odyssey.2022-48. URL: https://www.fit.vut.cz/research/publication/12843.

[Ana+22] Antonios Anastasopoulos et al. "FINDINGS OF THE IWSLT 2022 EVALUATION CAMPAIGN". In: *Proceedings of the 19th International Conference on Spoken Language Translation*. ACL. Stroudsburg, USA: Association for Computational Linguistics, 2022, pp. 98–157. ISBN: 978-1-955917-41-4.

[AP22] Michal Auersperger and Pavel Pecina. "Defending Compositionality in Emergent Languages". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*. Ed. by Daphne Ippolito et al. Hybrid: Seattle, Washington + Online: Association for Computational Linguistics, July 2022, pp. 285–291. DOI: 10.18653/v1/2022.naacl-srw.35. URL: https://aclanthology.org/2022.naacl-srw.35/.

[Bas+22] K. Murali Baskar et al. “Speaker adaptation for Wav2vec2 based dysarthric ASR”. In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 9. 9. Incheon, KR, 2022, pp. 3403–3407. DOI: 10.21437/Interspeech.2022-10896. URL: https://www.fit.vut.cz/research/publication/12854.

[BKB22] Sunit Bhattacharya, Rishu Kumar, and Ondrej Bojar. “Team ÚFAL at CMCL 2022 Shared Task: Figuring out the correct recipe for predicting Eye-Tracking features using Pretrained Language Models”. In: *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. Ed. by Emmanuele Chersoni et al. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 130–135. DOI: 10.18653/v1/2022.cmcl-1.15. URL: https://aclanthology.org/2022.cmcl-1.15/.

[BZB22] Sunit Bhattacharya, Vilém Zouhar, and Ondrej Bojar. “Sentence Ambiguity, Grammaticality and Complexity Probes”. In: *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Ed. by Jasmijn Bastings et al. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 40–50. DOI: 10.18653/v1/2022.blackboxnlp-1.4. URL: https://aclanthology.org/2022.blackboxnlp-1.4/.

[Bha+22a] Sunit Bhattacharya et al. *EMMT Release (version 1.0)*. 2022. DOI: 10.17605/OSF.IO/HXYMJ. URL: https://osf.io/hxymj/.

[Bru+22] Niko Brummer et al. “Probabilistic Spherical Discriminant Analysis: An Alternative to PLDA for length-normalized embeddings”. In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 2022. 9. Incheon, KR, 2022, pp. 1446–1450. DOI: 10.21437/Interspeech.2022-731. URL: https://www.fit.vut.cz/research/publication/12845.

[Ego+22] Ekaterina Egorova et al. “Spelling-Aware Word-Based End-to-End ASR”. In: *IEEE Signal Processing Letters* 29.29 (2022), pp. 1729–1733. ISSN: 1558-2361. DOI: 10.1109/LSP.2022.3192199. URL: https://www.fit.vut.cz/research/publication/12803.

[Gar+22] Muskan Garg et al. “Multimodality for NLP-Centered Applications: Resources, Advances and Frontiers”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, June 2022, pp. 6837–6847. URL: https://aclanthology.org/2022.lrec-1.738/.

[Han+22] Jiangyu Han et al. “DPCCN: Densely-Connected Pyramid Complex Convolutional Network for Robust Speech Separation and Extraction”. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Singapore, SG, 2022, pp. 7292–7296. ISBN: 978-1-6654-0540-9. DOI: 10.1109/ICASSP43922.2022.9747340. URL: https://www.fit.vut.cz/research/publication/12787.

[Hel22] Jindřich Helcl. "Non-Autoregressive Neural Machine Translation". PhD thesis. Praha, Czech Republic: Charles University in Prague, Faculty of Mathematics and Physics, 2022.

[HHB22] Jindřich Helcl, Barry Haddow, and Alexandra Birch. "Non-Autoregressive Machine Translation: It's Not as Fast as it Seems". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 1780–1790. DOI: 10.18653/v1/2022.naacl-main.129. URL: https://aclanthology.org/2022.naacl-main.129/.

[JMB22] Dávid Javorský, Dominik Macháček, and Ondřej Bojar. "Continuous Rating as Reliable Human Evaluation of Simultaneous Speech Translation". In: *Proceedings of the Seventh Conference on Machine Translation*. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 154–164.

[Kiš+22] Martin Kišš et al. "Importance of Textlines in Historical Document Classification". In: *Uchida, S., Barney, E., Eglin, V. (eds) Document Analysis Systems*. Vol. 13237. Lecture Notes in Computer Science. La Rochelle, FR, 2022, pp. 158–170. ISBN: 978-3-031-06554-5. DOI: 10.1007/978-3-031-06555-2_11. URL: https://www.fit.vut.cz/research/publication/12706.

[Koc+22b] Martin Kocour et al. "BCN2BRNO: ASR System Fusion for Albayzin 2022 Speech to Text Challenge". In: *Proceedings of IberSpeech 2022*. Granada, ES, 2022, pp. 276–280. DOI: 10.21437/IberSPEECH.2022-56. URL: https://www.fit.vut.cz/research/publication/12859.

[KP22] Mateusz Krubiński and Pavel Pecina. "From COMET to COMES – Can Summary Evaluation Benefit from Translation Evaluation?" In: *Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems*. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 21–31.

[KB22a] Nalin Kumar and Ondrej Bojar. "Genre Transfer in NMT:Creating Synthetic Spoken Parallel Sentences using Written Parallel Data". In: *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*. Ed. by Md. Shad Akhtar and Tanmoy Chakraborty. New Delhi, India: Association for Computational Linguistics, Dec. 2022, pp. 224–233. URL: https://aclanthology.org/2022.icon-main.28/.

[KB22b] Ivana Kvapilíková and Ondrej Bojar. "CUNI Submission to MT4All Shared Task". In: *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*. Ed. by Maite Melero, Sakriani Sakti, and Claudia Soria. Marseille, France: European Language Resources Association, June 2022, pp. 78–82. URL: https://aclanthology.org/2022.sigul-1.10/.

[Lan+22a] Federico Landini et al. “Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: Theory, implementation and analysis on standard tasks”. In: *Computer Speech and Language* 71.101254 (2022), pp. 1–16. ISSN: 0885-2308. DOI: 10.1016/j.csl.2021.101254. URL: https://www.fit.vut.cz/research/publication/12619.

[Lan+22b] Federico Landini et al. “From Simulated Mixtures to Simulated Conversations as Training Data for End-to-End Neural Diarization”. In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 2022. 9. Incheon, KR, 2022, pp. 5095–5099. DOI: 10.21437/Interspeech.2022-10451. URL: https://www.fit.vut.cz/research/publication/12846.

[MP22] Jiří Mayer and Pavel Pecina. “Obstacles with Synthesizing Training Data for OMR”. In: *Proceedings of the 4th International Workshop on Reading Music Systems*. Ed. by Jorge Calvo-Zaragoza, Alexander Pacha, and Elona Shatri. University of Alicante. Alicante, Spain: University of Alicante, 2022, pp. 15–19.

[Moš+22a] Ladislav Mošner et al. “Multi-Channel Speaker Verification with Conv-Tasnet Based Beamformer”. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Singapore, SG, 2022, pp. 7982–7986. ISBN: 978-1-6654-0540-9. DOI: 10.1109/ICASSP43922.2022.9747771. URL: https://www.fit.vut.cz/research/publication/12786.

[Moš+22b] Ladislav Mošner et al. “MultiSV: Dataset for Far-Field Multi-Channel Speaker Verification”. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Singapore, SG, 2022, pp. 7977–7981. ISBN: 978-1-6654-0540-9. DOI: 10.1109/ICASSP43922.2022.9746833. URL: https://www.fit.vut.cz/research/publication/12785.

[Nad+22] Vijaya Nadimpalli et al. “Resources and Benchmarks for Keyword Search in Spoken Audio From Low-Resource Indian Languages”. In: *IEEE Access* 10.2022 (2022), pp. 34789–34799. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3162854. URL: https://www.fit.vut.cz/research/publication/12952.

[Nak+22] Toshiaki Nakazawa et al. “Overview of the 9th Workshop on Asian Translation”. In: *Proceedings of the 9th Workshop on Asian Translation*. Gyeongju, Republic of Korea: International Conference on Computational Linguistics, Oct. 2022, pp. 1–36. URL: https://aclanthology.org/2022.wat-1.1/.

[Ned+22] Anna Nedoluzhko et al. “ELITR Minuting Corpus: A Novel Dataset for Automatic Minuting from Multi-Party Meetings in English and Czech”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, June 2022, pp. 3174–3182. URL: https://aclanthology.org/2022.lrec-1.340/.

[Ond+22] Lucas Ondel et al. “Non-Parametric Bayesian Subspace Models for Acoustic Unit Discovery”. In: *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING* 30.5 (2022), pp. 1902–1917. ISSN: 2329-9290. DOI: 10.1109/TASLP.2022.3171975. URL: https://www.fit.vut.cz/research/publication/12791.

[Pen+22a] Junyi Peng et al. “Learnable Sparse Filterbank for Speaker Verification”. In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. 9. Incheon, KR, 2022, pp. 5110–5114. DOI: 10.21437/Interspeech.2022-11309. URL: https://www.fit.vut.cz/research/publication/12851.

[Pen+22b] Junyi Peng et al. “Progressive contrastive learning for self-supervised text-independent speaker verification”. In: *Proceedings of The Speaker and Language Recognition Workshop (Odyssey 2022)*. Beijing, CN, 2022, pp. 17–24. DOI: 10.21437/Odyssey.2022-3. URL: https://www.fit.vut.cz/research/publication/12835.

[Pol+22a] Peter Polák et al. “ALIGNMEET: A Comprehensive Tool for Meeting Annotation, Alignment, and Evaluation”. In: *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)* (Le Palais du Pharo). Ed. by Nicoletta Calzolari et al. 549. Marseille, France: European Language Resources Association, 2022, pp. 1771–1779. ISBN: 979-10-95546-72-6.

[Pož+22] Borek Požár et al. “CUNI Submission to the BUCC 2022 Shared Task on Bilingual Term Alignment”. In: *Proceedings of the BUCC Workshop within LREC 2022*. Ed. by Reinhard Rapp, Pierre Zweigenbaum, and Serge Sharoff. Marseille, France: European Language Resources Association, June 2022, pp. 43–49. URL: https://aclanthology.org/2022.bucc-1.6/.

[RL22] Philipp J. Rösch and Jindřich Libovický. “Probing the Role of Positional Information in Vision-Language Models”. In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 1031–1041. DOI: 10.18653/v1/2022.findings-naacl.77. URL: https://aclanthology.org/2022.findings-naacl.77/.

[Sil+22] Anna Silnova et al. “Analyzing speaker verification embedding extractors and back-ends under language and channel mismatch”. In: *Proceedings of The Speaker and Language Recognition Workshop (Odyssey 2022)*. Beijing, CN, 2022, pp. 9–16. DOI: 10.21437/Odyssey.2022-2. URL: https://www.fit.vut.cz/research/publication/12834.

[Sta+22] Themos Stafylakis et al. “Training Speaker Embedding Extractors Using Multi-Speaker Audio with Unknown Speaker Boundaries”. In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 2022. 9. Incheon, KR, 2022, pp. 605–609. DOI: 10.21437/Interspeech.2022-10165. URL: https://www.fit.vut.cz/research/publication/12847.

2023 Publications

[Aga+23] Milind Agarwal et al. "FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN". In: *Proceedings of the 20th International Conference on Spoken Language Translation*. ACL. Stroudsburg, USA: Association for Computational Linguistics, 2023, pp. 1–61. ISBN: 978-1-959429-84-5.

[Del+23] Marc Delcroix et al. "Multi-Stream Extension of Variational Bayesian HMM Clustering (MS-VBx) for Combined End-to-End and Vector Clustering-based Diarization". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 2023. 08. Dublin, IE, 2023, pp. 3477–3481. DOI: 10.21437/Interspeech.2023-628. URL: https://www.fit.vut.cz/research/publication/13110.

[Gho+23] Tirthankar Ghosal et al. "Overview of the Second Shared Task on Automatic Minuting (AutoMin) at INLG 2023". In: *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*. Ed. by Simon Mille. Prague, Czechia: Association for Computational Linguistics, Sept. 2023, pp. 138–167. URL: https://aclanthology.org/2023.inlg-genchal.19/.

[JBY23] Dávid Javorský, Ondřej Bojar, and François Yvon. "Assessing Word Importance Using Models Trained for Semantic Tasks". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 8846–8856. DOI: 10.18653/v1/2023.findings-acl.563. URL: https://aclanthology.org/2023.findings-acl.563/.

[JB23a] Josef Jon and Ondřej Bojar. "Breeding Machine Translations: Evolutionary approach to survive and thrive in the world of automated evaluation". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 2191–2212. DOI: 10.18653/v1/2023.acl-long.122. URL: https://aclanthology.org/2023.acl-long.122/.

[JB23b] Josef Jon and Ondřej Bojar. "Character-level NMT and language similarity". In: *Proceedings of Machine Translation Summit XIX vol. 1: Research Track*. Asia-Pacific Association for Machine Translation (AAMT). Kyoto, Japan: Asia-Pacific Association for Machine Translation (AAMT), 2023, pp. 360–371. ISBN: 978-4-9913461-0-1.

[JPB23] Josef Jon, Martin Popel, and Ondřej Bojar. "CUNI at WMT23 General Translation Task: MT and a Genetic Algorithm". In: *Proceedings of the Eighth Conference on Machine Translation*. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 119–127. ISBN: 979-8-89176-041-7.

[Jon+23] Josef Jon et al. “Negative Lexical Constraints in Neural Machine Translation”. In: *Proceedings of Machine Translation Summit XIX vol. 1: Research Track*. Asia-Pacific Association for Machine Translation (AAMT). Kyoto, Japan: Asia-Pacific Association for Machine Translation (AAMT), 2023, pp. 372–384. ISBN: 978-4-9913461-0-1.

[Kak+23] Sofoklis Kakouros et al. “Speech-Based Emotion Recognition with Self-Supervised Models Using Attentive Channel-Wise Correlations and Label Smoothing”. In: *Proceedings of ICASSP 2023*. Rhodes Island, GR, 2023, pp. 1–5. ISBN: 978-1-7281-6327-7. DOI: 10.1109/ICASSP49357.2023.10094673. URL: https://www.fit.vut.cz/research/publication/13054.

[Kes+23a] Santosh Kesiraju et al. “BUT Systems for IWSLT 2023 Marathi - Hindi Low Resource Speech Translation Task”. In: *20th International Conference on Spoken Language Translation, IWSLT 2023 - Proceedings of the Conference*. Toronto (in-person and online), CA, 2023, pp. 227–234. ISBN: 978-1-959429-84-5. DOI: 10.18653/v1/2023.iwslt-1.19. URL: https://www.fit.vut.cz/research/publication/13055.

[Kes+23b] Santosh Kesiraju et al. “Strategies for Improving Low Resource Speech to Text Translation Relying on Pre-trained ASR Models”. In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 2023. 08. Dublin, IE, 2023, pp. 2148–2152. DOI: 10.21437/Interspeech.2023-2506. URL: https://www.fit.vut.cz/research/publication/13109.

[KE23] Kristýna Klesnilová and Michelle Elizabeth. “Team Synapse @ AutoMin 2023: Leveraging BART-Based Models for Automatic Meeting Minuting”. In: *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*. Ed. by Simon Mille. Prague, Czechia: Association for Computational Linguistics, Sept. 2023, pp. 108–113. URL: https://aclanthology.org/2023.inlg-genchal.15/.

[Klo+23] Věra Kloudová et al. “Možnosti a meze tvorby tzv. optimálních referenčních překladů: po stopách „překladatelštiny“ v profesionálních překladech zpravodajských textů”. In: *Slovo a slovesnost* 84.2 (2023), pp. 122–156. DOI: 10.58756/s2228425.

[KP23a] Mateusz Krubiński and Pavel Pecina. *MLASK: Multimodal Summarization of Video-based News Articles*. Institute of Formal and Applied Linguistics, 2023.

[KP23b] Mateusz Krubiński and Pavel Pecina. “MLASK: Multimodal Summarization of Video-based News Articles”. In: *Findings of the Association for Computational Linguistics: EACL 2023*. Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 910–924. ISBN: 978-1-959429-47-0.

[KB23a] Ivana Kvapilíková and Ondřej Bojar. "Boosting Unsupervised Machine Translation with Pseudo-Parallel Data". In: *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*. Ed. by Masao Utiyama and Rui Wang. Macau SAR, China: Asia-Pacific Association for Machine Translation, Sept. 2023, pp. 135–147. URL: https://aclanthology.org/2023.mtsummit-research.12/.

[KB23b] Ivana Kvapilíková and Ondřej Bojar. "Low-Resource Machine Translation Systems for Indic Languages". In: *Proceedings of the Eighth Conference on Machine Translation*. Ed. by Philipp Koehn et al. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 954–958. DOI: 10.18653/v1/2023.wmt-1.90. URL: https://aclanthology.org/2023.wmt-1.90/.

[Lan+23] Federico Landini et al. "Multi-Speaker and Wide-Band Simulated Conversations as Training Data for End-to-End Neural Diarization". In: *Proceedings of ICASSP 2023*. Rhodes Island, GR, 2023, pp. 1–5. ISBN: 978-1-7281-6327-7. DOI: 10.1109/ICASSP49357.2023.10097049. URL: https://www.fit.vut.cz/research/publication/13051.

[MBD23] Dominik Macháček, Ondřej Bojar, and Raj Dabre. "MT Metrics Correlate with Human Ratings of Simultaneous Speech Translation". In: *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*. Ed. by Elizabeth Salesky, Marcello Federico, and Marine Carpuat. Toronto, Canada (in-person and online): Association for Computational Linguistics, July 2023, pp. 169–179. DOI: 10.18653/v1/2023.iwslt-1.12. URL: https://aclanthology.org/2023.iwslt-1.12/.

[MDB23] Dominik Macháček, Raj Dabre, and Ondřej Bojar. "Turning Whisper into Real-Time Transcription System". In: *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*. Ed. by Sriparna Saha and Herry Sujaini. Bali, Indonesia: Association for Computational Linguistics, Nov. 2023, pp. 17–24. DOI: 10.18653/v1/2023.ijcnlp-demo.3. URL: https://aclanthology.org/2023.ijcnlp-demo.3/.

[Mac+23] Dominik Macháček et al. "Robustness of Multi-Source MT to Transcription Errors". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 3707–3723. DOI: 10.18653/v1/2023.findings-acl.228. URL: https://aclanthology.org/2023.findings-acl.228/.

[Mat+23] Pavel Matějka et al. "Description and Analysis of ABC Submission to NIST LRE 2022". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 2023. 08. Dublin, IE, 2023, pp. 511–515. DOI: 10.21437/Interspeech.2023-1529. URL: https://www.fit.vut.cz/research/publication/13111.

[Moš+23] Ladislav Mošner et al. “Multi-Channel Speech Separation with Cross-Attention and Beamforming”. In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 2023. 08. Dublin, IE, 2023, pp. 1693–1697. DOI: 10.21437/Interspeech.2023-2537. URL: https://www.fit.vut.cz/research/publication/13108.

[Nak+23] Toshiaki Nakazawa et al. “Overview of the 10th Workshop on Asian Translation”. In: *Proceedings of the 10th Workshop on Asian Translation*. Ed. by Toshiaki Nakazawa et al. Macau SAR, China: Asia-Pacific Association for Machine Translation, Sept. 2023, pp. 1–28. URL: https://aclanthology.org/2023.wat-1.1/.

[NB23] Kristyna Neumannova and Ondřej Bojar. “The Role of Compounds in Human vs. Machine Translation Quality”. In: *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*. Ed. by Masao Utiyama and Rui Wang. Macau SAR, China: Asia-Pacific Association for Machine Translation, Sept. 2023, pp. 248–260. URL: https://aclanthology.org/2023.mtsummit-research.21/.

[Par+23] Shantipriya Parida et al. “HaVQA: A Dataset for Visual Question Answering and Multimodal Research in Hausa Language”. In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 10162–10183. DOI: 10.18653/v1/2023.findings-acl.646. URL: https://aclanthology.org/2023.findings-acl.646/.

[Pen+23a] Junyi Peng et al. “An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification”. In: *2022 IEEE Spoken Language Technology Workshop, SLT 2022 - Proceedings*. Doha, QA, 2023, pp. 555–562. ISBN: 978-1-6654-7189-3. DOI: 10.1109/SLT54892.2023.10022775. URL: https://www.fit.vut.cz/research/publication/12984.

[Pen+23b] Junyi Peng et al. “Improving Speaker Verification with Self-Pretrained Transformer Models”. In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 2023. 08. Dublin, IE, 2023, pp. 5361–5365. DOI: 10.21437/Interspeech.2023-453. URL: https://www.fit.vut.cz/research/publication/13112.

[Pen+23c] Junyi Peng et al. “Parameter-Efficient Transfer Learning of Pre-Trained Transformer Models for Speaker Verification Using Adapters”. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Rhodes Island, GR, 2023, pp. 1–5. ISBN: 978-1-7281-6327-7. DOI: 10.1109/ICASSP49357.2023.10094795. URL: https://www.fit.vut.cz/research/publication/13053.

[Per+23] Andrej Perković et al. “Shortening of the results of machine translation using paraphrasing dataset”. In: *Proceedings of the 23rd Conference Information Technologies – Applications and Theory (ITAT 2023)*. Ed. by Broňa

Brejová et al. CEUR-WS.org. Košice, Slovakia: 23rd Conference on Information Technologies – Applications and Theory, 2023, pp. 121–130.

[Pol23] Peter Polák. "Long-form Simultaneous Speech Translation: Thesis Proposal". In: *Proceedings of the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 13th International Joint Conference on Natural Language Processing: Student Research Workshop*. ACL. Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 64–74.

[Pol+23a] Peter Polák et al. "Incremental Blockwise Beam Search for Simultaneous Speech Translation with Controllable Quality-Latency Tradeoff". In: *Proceedings of the 24st Annual Conference of the International Speech Communication Association*. International Speech Communication Association. Baixas, France: International Speech Communication Association, 2023, pp. 3979–3983.

[Pol+23b] Peter Polák et al. "Towards Efficient Simultaneous Speech Translation: CUNI-KIT System for Simultaneous Track at IWSLT 2023". In: *Proceedings of the 20th International Conference on Spoken Language Translation*. ACL. Stroudsburg, USA: Association for Computational Linguistics, 2023, pp. 389–396. ISBN: 978-1-959429-84-5.

[Sil+23a] Anna Silnova et al. "ABC System Description for NIST LRE 2022". In: *Proceedings of NIST LRE 2022 Workshop*. Washington DC, US, 2023, pp. 1–5. URL: https://www.fit.vut.cz/research/publication/12986.

[Sil+23b] Anna Silnova et al. "Toroidal Probabilistic Spherical Discriminant Analysis". In: *Proceedings of ICASSP 2023*. Rhodes Island, GR, 2023, pp. 1–5. ISBN: 978-1-7281-6327-7. DOI: 10.1109/ICASSP49357.2023.10095580. URL: https://www.fit.vut.cz/research/publication/13052.

[Sta+23] Themos Stafylakis et al. "Extracting speaker and emotion information from self-supervised speech models via channel-wise correlations". In: *2022 IEEE Spoken Language Technology Workshop, SLT 2022 - Proceedings*. Doha, QA, 2023, pp. 1136–1143. ISBN: 978-1-6654-7189-3. DOI: 10.1109/SLT54892.2023.10023345. URL: https://www.fit.vut.cz/research/publication/12985.

[TSB23] Frantisek Trebuna, Kristína Szabová, and Ondrej Bojar. "Searching for Reasons of Transformers' Success: Memorization vs Generalization". In: *Text, Speech, and Dialogue - 26th International Conference, TSD 2023, Pilsen, Czech Republic, September 4-6, 2023, Proceedings*. Ed. by Kamil Ekstein, Frantisek Pártl, and Miloslav Konopík. Vol. 14102. Lecture Notes in Computer Science. Springer, 2023, pp. 25–32. DOI: 10.1007/978-3-031-40498-6_3. URL: https://doi.org/10.1007/978-3-031-40498-6%5C_3.

[TTB23] Iryna Tryhubyshyn, Aleš Tamchyna, and Ondřej Bojar. "Bad MT Systems are Good for Quality Estimation". In: *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*. Ed. by Masao Utiyama and Rui Wang. Macau SAR, China: Asia-Pacific Association for Machine Translation, Sept. 2023, pp. 200–208. URL: https://aclanthology.org/2023.mtsummit-research.17/.

[Var23] Dušan Variš. "Learning capabilities in Transformer Neural Networks". PhD thesis. Praha, Czech Republic: Charles University in Prague, Faculty of Mathematics and Physics, 2023.

[YBH23] Uladzislau Yorsh, Ondřej Bojar, and Martin Holeňa. "On Difficulties of Attention Factorization through Shared Memory". In: *Proc. of 10th European Starting AI Researchers' Symposium (STAIRS) co-located with the 26th European Conference on Artificial Intelligence (ECAI-23)*. Oct. 2023. URL: https://ecai2023.eu/STAIRS.

[Yu+23] Dong Yu et al. "Twenty-Five Years of Evolution in Speech and Language Processing". In: *IEEE Signal Processing Magazine* 40.5 (2023), pp. 27–39. ISSN: 1558-0792. DOI: 10.1109/MSP.2023.3266155. URL: https://www.fit.vut.cz/research/publication/13058.

[YČS23] Bolaji Yusuf, Jan Černocký, and Murat Saraçlar. "End-to-End Open Vocabulary Keyword Search With Multilingual Neural Representations". In: *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING* 31.08 (2023), pp. 3070–3080. ISSN: 2329-9290. DOI: 10.1109/TASLP.2023.3301239. URL: https://www.fit.vut.cz/research/publication/13057.

2024 Publications

[Ahm+24] Ibrahim Ahmad et al. "FINDINGS OF THE IWSLT 2024 EVALUATION CAMPAIGN". In: *Proceedings of the 21st International Conference on Spoken Language Translation*. ACL. Stroudsburg, USA: Association for Computational Linguistics, 2024, pp. 1–11. ISBN: 979-8-89176-141-4.

[BKB24] Karel Beneš, Martin Kocour, and Lukáš Burget. "Hystoc: Obtaining Word Confidences for Fusion of End-To-End ASR Systems". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Seoul, KR, 2024, pp. 11276–11280. ISBN: 979-8-3503-4485-1. DOI: 10.1109/ICASSP48485.2024.10446739. URL: https://www.fit.vut.cz/research/publication/13267.

[Ďur+24] Dominika Ďurišková et al. "Khan Academy Corpus: A Multilingual Corpus of Khan Academy Lectures". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari et al. Torino, Italia: ELRA and ICCL, May 2024, pp. 9743–9752. URL: https://aclanthology.org/2024.lrec-main.851/.

[Han+24] Jiangyu Han et al. "Diacorrect: Error Correction Back-End for Speaker Diarization". In: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Seoul, KR, 2024, pp. 11181–11185. ISBN: 979-8-3503-4485-1. DOI: 10.1109/ICASSP48485.2024.10446968. URL: https://www.fit.vut.cz/research/publication/13268.

[JB24] Josef Jon and Ondřej Bojar. "GAATME: A Genetic Algorithm for Adversarial Translation Metrics Evaluation". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari et al. Torino, Italia: ELRA and ICCL, May 2024, pp. 7562–7569. URL: https://aclanthology.org/2024.lrec-main.668/.

[Kle+24] Dominik Klement et al. "Discriminative Training of VBx Diarization". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Seoul, KR, 2024, pp. 11871–11875. ISBN: 979-8-3503-4485-1. DOI: 10.1109/ICASSP48485.2024.10446119. URL: https://www.fit.vut.cz/research/publication/13277.

[Klo+24] Věra Kloudová et al. "A Comparison of Human and Machine Performance in Simultaneous Interpreting as a Basis for Further Research in Automatic Simultaneous Speech Translation". In: *Advancing interdisciplinarity in empirical translation research*. Ed. by Tomáš Svoboda, Olga Nádvorníková, and Věra Kloudová. Berlin: Language Science Press, 2024, in press.

[KP24] Mateusz Krubiński and Pavel Pecina. "Towards Unified Uni- and Multimodal News Headline Generation". In: *Findings of the Association for Computational Linguistics: EACL 2024*. Ed. by Yvette Graham and Matthew Purver. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 437–450. URL: https://aclanthology.org/2024.findings-eacl.30/.

[Kva24] Ivana Kvapilíková. "Towards Machine Translation Based on Monolingual Texts". PhD thesis. Praha, Czech Republic: Charles University in Prague, Faculty of Mathematics and Physics, 2024.

[Lan+24] Federico Landini et al. "DiaPer: End-to-End Neural Diarization With Perceiver-Based Attractors". In: *IEEE Transactions on Audio, Speech, and Language Processing* 32.7 (2024), pp. 3450–3465. ISSN: 1558-7916. DOI: 10.1109/TASLP.2024.3422818. URL: https://www.fit.vut.cz/research/publication/13279.

[Mac24] Dominik Macháček. "Multi-Source Simultaneous Speech Translation". PhD thesis. Praha, Czech Republic: Charles University, Faculty of Mathematics and Physics, 2024.

[May+24] Jiří Mayer et al. "Practical End-to-End Optical Music Recognition for Pianoform Music". In: *Document Analysis and Recognition – ICDAR 2024*. Springer International Publishing. Cham, Switzerland: Springer International Publishing, 2024, pp. 55–73. ISBN: 978-3-030-86333-3.

[OJB24] Adam Osuský, Dávid Javorský, and Ondřej Bojar. "InsBERT: Word importance from artificial insertions". In: *Proceedings of the 24th Conference Information Technologies – Applications and Theory (ITAT 2024)* (Hotel Javorna). Ed. by Lucie Ciencialová et al. Univerzita Pavla Jozefa Šafárika v Košiciach, Košice, Slovakia. Košice, Slovakia: CEUR-WS.org, 2024, pp. 96–106.

[Pen+24a] Junyi Peng et al. "Probing Self-Supervised Learning Models With Target Speech Extraction". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Seoul, KR, 2024, pp. 535–539. ISBN: 979-8-3503-7451-3. DOI: 10.1109/ICASSPW62465.2024.10627502. URL: https://www.fit.vut.cz/research/publication/13276.

[Pen+24b] Junyi Peng et al. "Target Speech Extraction with Pre-Trained Self-Supervised Learning Models". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Seoul, KR, 2024, pp. 10421–10425. ISBN: 979-8-3503-4485-1. DOI: 10.1109/ICASSP48485.2024.10448315. URL: https://www.fit.vut.cz/research/publication/13275.

[WZB24] Hening Wang, Leixin Zhang, and Ondrej Bojar. "Human and Machine: Language Processing in Translation Tasks". In: *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*. Ed. by Mourad Abbas and Abed Alhakim Freihat. Trento: Association for Computational Linguistics, Oct. 2024, pp. 243–250. URL: https://aclanthology.org/2024.icnlsp-1.27/.

[Yor+24] Uladzislau Yorsh et al. "On Difficulties of Attention Factorization through Shared Memory". In: *The Twelveth International Conference on Learning Representations, ICLR 2024*. 2024. URL: https://openreview.net/pdf?id=pexcddsXGY.

[Zha+24] Leixin Zhang et al. "Unveiling Semantic Information in Sentence Embeddings". In: *Proceedings of the Fifth International Workshop on Designing Meaning Representations @ LREC-COLING 2024*. Ed. by Claire Bonial, Julia Bonn, and Jena D. Hwang. Torino, Italia: ELRA and ICCL, May 2024, pp. 39–47. URL: https://aclanthology.org/2024.dmr-1.5/.

[ZBB24] Vilém Zouhar, Sunit Bhattacharya, and Ondřej Bojar. "Multimodal Shannon Game with Images". In: *Proceedings of the 10th International Conference on Learning, Optimization and Data – LOD/ACAIN*. Castiglione della Pescaia, Italy: Springer Cham, Sept. 2024.

[ZB24] Vilém Zouhar and Ondřej Bojar. "Quality and Quantity of Machine Translation References for Automatic Metrics". In: *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*. Ed. by Simone Balloccu et al. Torino, Italia: ELRA and ICCL, May 2024, pp. 1–11. URL: https://aclanthology.org/2024.humeval-1.1/.

[Zou+24] Vilém Zouhar et al. "Evaluating optimal reference translations". In: *Natural Language Processing* (2024), pp. 1–24. DOI: 10.1017/nlp.2024.3.

2025 Publications

[Boj25] Ondřej Bojar. "Evaluation Techniques, Experiment Design and Common Misconceptions in Machine Translation Research". In: *Machine Translation: Best Practices using Deep Learning and Generative AI*. Ed. by Elizabeth Sherly et al. In print. CRC Press, Taylor and Francis Group, 2025.

[Kva25] Ivana Kvapilíková. *Unsupervised Machine Translation: How Machines Learn to Understand across Languages*. Praha, Czechia: Karolinum Press, 2025. ISBN: 978-80-246-6084-4.

References (not project's outputs)

[TKW21] Emiru Tsunoo, Yosuke Kashiwagi, and Shinji Watanabe. "Streaming transformer asr with blockwise synchronous beam search". In: *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 2021, pp. 22–29.

[Jun+22] Jee-weon Jung et al. "SASV 2022: The First Spoofing-Aware Speaker Verification Challenge". In: *Proc. Interspeech*. 2022, pp. 2893–2897. DOI: 10.21437/Interspeech.2022-11270.

[Pol+22b] Peter Polák et al. "CUNI-KIT System for Simultaneous Speech Translation Task at IWSLT 2022". In: *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*. Ed. by Elizabeth Salesky, Marcello Federico, and Marta Costa-jussà. Dublin, Ireland (in-person and online): Association for Computational Linguistics, May 2022, pp. 277–285. DOI: 10.18653/v1/2022.iwslt-1.24. URL: https://aclanthology.org/2022.iwslt-1.24/.

[Tak+22] Hemlata Tak et al. "Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation". In: *Proc. Odyssey*. 2022, pp. 112–119. DOI: 10.21437/Odyssey.2022-16.

[WY22] Xin Wang and Junichi Yamagishi. "Investigating Self-Supervised Front Ends for Speech Spoofing Countermeasures". In: *Proc. Odyssey*. 2022, pp. 100–106. DOI: 10.21437/Odyssey.2022-14.

[Kaw+23] Piotr Kawa et al. "Improved DeepFake Detection Using Whisper Features". In: *Proc. Interspeech*. 2023, pp. 4009–4013. DOI: 10.21437/Interspeech.2023-1537.

[Rad+23] Alec Radford et al. "Robust speech recognition via large-scale weak supervision". In: *International conference on machine learning*. PMLR. 2023, pp. 28492–28518.

[Moš+24] Ladislav Mošner et al. "Multi-Channel Extension of Pre-trained Models for Speaker Verification". In: *Proceedings of Interspeech 2024*. Kos, GR, 2024, pp. 2135–2139. URL: https://www.fit.vut.cz/research/publication/13322.

[Pol+24a] Alexander Polok et al. "BUT/JHU System Description for CHiME-8 NOTSOFAR-1 Challenge". In: *8th International Workshop on Speech Processing in Everyday Environments (CHiME 2024)*. 2024, pp. 18–22. DOI: 10.21437/CHiME.2024-4.

[Pol+24b] Alexander Polok et al. *DiCoW: Diarization-Conditioned Whisper for Target Speaker Automatic Speech Recognition*. Submitted to Computer Speech & Language (CSL) journal special issue on Multi-Speaker, Multi-Microphone, and Multi-Modal Distant Speech Recognition. 2024. arXiv: 2501.00114 [eess.AS]. URL: https://arxiv.org/abs/2501.00114.

[Roh+24] Johan Rohdin et al. "BUT systems and analyses for the ASVspoof 5 Challenge". In: *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*. 2024, pp. 24–31. DOI: 10.21437/ASVspoof.2024-4.

[Wan+24] Xin Wang et al. "ASVspoof 5: Crowdsourced Speech Data, Deepfakes, and Adversarial Attacks at Scale". In: *Proc. ASVspoof Workshop 2024 (accepted)*. 2024.

[Pol+25] Alexander Polok et al. "Target Speaker ASR with Whisper". In: *accepted to ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2025.