NEUREM3 Interim Research Report

Lukáš Burget and Ondřej Bojar - editors

January 2022

Executive summary

Neural Representations in Multi-modal and Multi-lingual Modeling (NEUREM3) is 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. This report covers its first three 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 multi-lingual concepts. Its description is clustered according to 5 tasks defined in the project proposal and several technical topics. In several of these, we have reached beyond state-of-the-art results. 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 is 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 is competitive on the international level, the assessment is done via international technology evaluations (challenges), bibliographic metrics, and organization of top international scientific events. NEUREM3 team efficiently cooperates and is at the core of building both Czech and international speech/NLP/MT communities.

So far, the project led to a total of 96 publications of which 12 were in peer-reviewed journals, 49 at top conferences, and 35 at local workshops, challenge and evaluation workshops, etc. We are also regularly releasing research data and code in open repositories.

Contents

1 Introduction

Neural Representations in Multi-modal and Multi-lingual Modeling (NEUREM3) is 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 proposal is at the intersection of two important domains of artificial intelligence (AI) – natural language processing (NLP) and speech processing (SP). It proposes a systematic study of structures based on artificial neural networks (NN) for SP and NLP, addressing the current lack of fundamental understanding of neural representations and their interplays in machine learning. NEUREM3 concentrates on granularity/size and scope of neural representations, their interpretability, and interactions between long-span and short-span representations. It also investigates into embeddings and end-to-end systems both in single domains, and in systems combining NLP, SP and partly computer vison (CV).

The project is proposed by a joint team of two Czech institutions that can be considered among the world leaders in SP and NLP: Brno University of Technology Speech@FIT group (BUT), and Charles University Institute of Formal and Applied Linguistics (CUNI), with vast international cooperation network, and track in dozens of EU-, US- and locally funded research projects.

1.1 Purpose and structure of this document

GAČR requested an interim scientific report in the mid-term of the project, i.e. in June 2021, however, we were allowed to submit it only as a part of the 2021 year reporting. Thanks to this, we were able to include all the results of 2021. 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.* This section is divided into broad topics in which we are working and is based on publications that are the main output of our project - for each, a summary, the main results, and 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 previous outputs within the framework of international excellence;* concentrating especially on international technology evaluations (challenges), bibliographic metrics, 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.*

The official structure does not require plans for the next period, however, we find these quite important, therefore, we include a special Section 7 concluding the report. The last Section 8 lists all the project outputs - mainly the publications. These are divided into individual years of project performance, all 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 are 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. Our research in NEUREM3 can be categorized into five broad areas:

- 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 are 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 of the first period

The detailed research plan and methodology for the first part of the project consist of five tasks. The following paragraphs summarize the relevant work, with the numbers of respective detailed sections.

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

Multi-linguality was studied on several levels. On one hand, we attempted to suppress the information on language as a nuisance (such as our work towards robust speaker

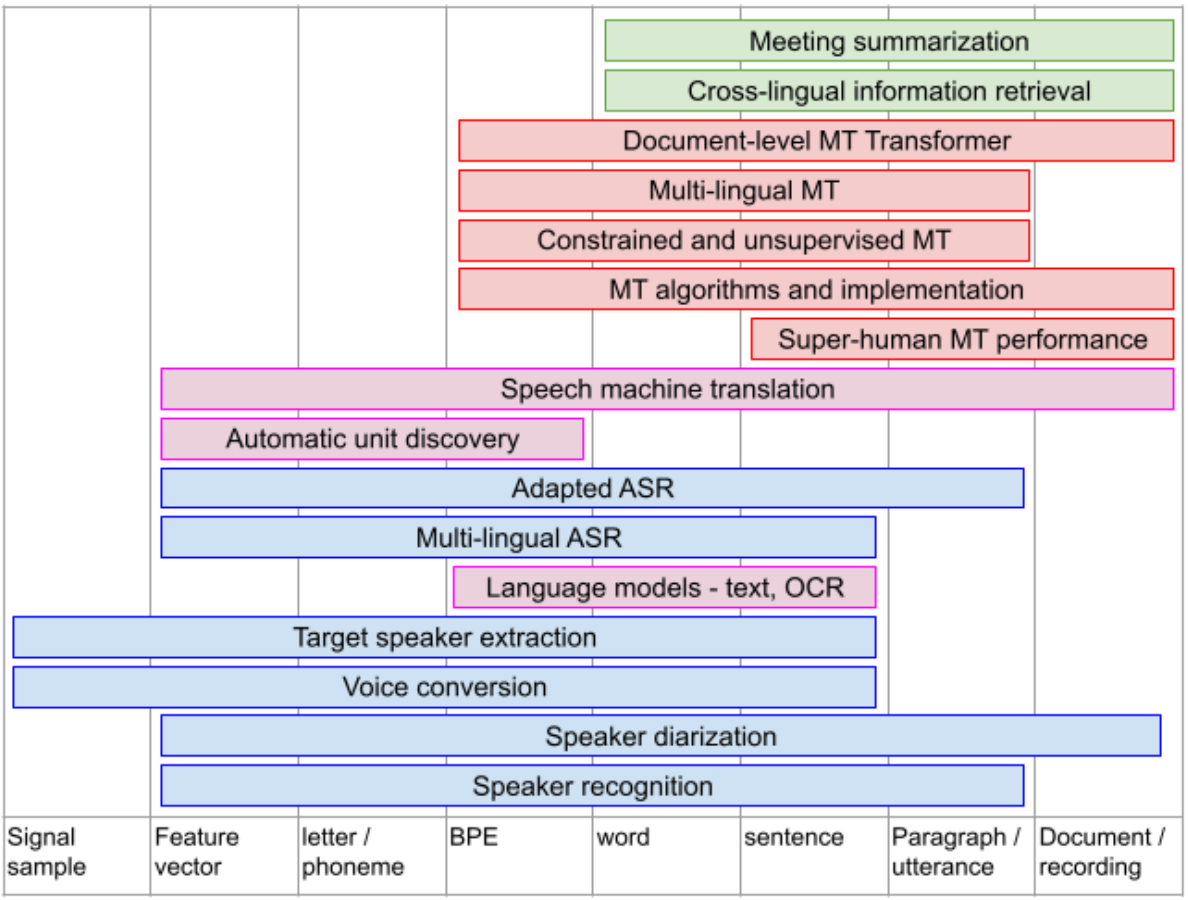


Figure 1: Overall scheme of investigated neural representations across unit granularity.

recognition embeddings 2.2.1 or in mitigating the phonetic information in SR embeddings 2.2.3). However, the multi-linguality is most often the target of our investigations, whether to create new data for ASR training (2.3.1), or in building ASR systems for several languages based on limited data 2.3.2. The whole MT Task (see below for Task 5) is inherently multi-lingual, but an MT approach making direct use of multi-lingual data for Better MT of Low-Resource Languages 2.5.10 is a clear representative of a successful Task 1 work.

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

Here, the results include "obvious" multi-modal investigations, primarily Eye-Tracking of Multi-Modal Human Translation 2.7.1 and Machine Translation Supported by Images and Image Captioning 2.7.2. However, we also extended the notion of multi-modality to possible re-use of investigated techniques on other than the original modality by Synthesizing Training Data for Handwritten Music Recognition 2.7.3 or using language model training and adaptation techniques not only for ASR and MT, but also for optical character recognition (OCR) of historical texts 2.4.2. Finally, for speech processing, "modality" can also be understood as making use of several kinds of information in the speech signal - an area that we largely explored for example in voice conversion 2.1.6 by taking into account speaker and content embeddings, and in target speaker separation 2.1.7 where

the information on the speaker is used to clean the signal for ASR either step-by-step or in an end-to-end system 2.1.7.

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 Interpretable Complex Filter 2.1.5) till the highest information levels such as whole scientific papers that were subject to information retrieval 2.8.8. The longitudinal analysis of SR systems on historical and current data 2.2.5 actually covers an extensive investigation of modeled units, from individual speech feature vectors to entire utterances. Unusual representation we explored include Phase Encoding 2.1.4 and Interpretable Complex Filters 2.1.5 in speaker recognition and probabilistic approaches for speaker extraction, speaker recognition and speaker diarization (variational auto-encoders and spatial clustering for multi-channel speech separation 2.1.8, Speaker diarization based on Bayesian HMM with eigenvoice priors (BHMM) 2.2.6 and Probabilistic embeddings for speaker diarization 2.2.8). The investigation of representations was however extensive also for text and translation, in studying Document-level Transformer 2.5.4, Linguistic Annotation for Transformers 2.5.2 or gradually growing input 2.6.1. Finally, human factors intervened in this task, especially in the research of Relation between MT System Quality and Time Savings in Translators' Workflow 2.5.3 and quest for Explainable MT Quality Estimation 2.5.12.

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 Data-driven approaches for ASR 2.3.4, we have attempted to automatically infer acoustic units (AUD, acoustic units discovery) that would perform equally or better than hand-crafted representations, especially for unknown languages. The second way was to explore cycle-consistent training making use of a differentiable chain of ASR and TTS (text to speech synthesis). We need to note our work on end-to-end SMT and OOV processing (2.6.3, 2.4.1), as the techniques used a hierarchy of units. The units however do not need to represent only speech sounds or text entities - we have also explored variational auto-encoders and spatial clustering to find probable positions of speakers for multi-channel speech separation 2.1.8. In our work on Multi-Source ASR 2.3.3, explored different unit levels for transfer learning for Czech ASR - we started with an intermediate "coarse" alphabet (i.e., the Czech alphabet without accents) and proceeded to the full alphabet. Finally, in our "semantic" work on Representations of Sentence Meaning 2.8.1, Translation into Many Paraphrases 2.8.2 and Compositionality in Sequence-to-Sequence Models 2.8.3 (as well as in our corpus COSTRA of Complex Sentence Transformations 2.8.4), the investigated unit is a sentence, and in our works toward Meeting Summarization 2.8.6 and Multimodal meeting summarization 2.8.7, the ultimate units are the whole meetings.

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

Significant efforts were devoted both to text-only MT and speech MT. Data was taken into account in studying Adequacy by Distinguishing Genuine and Synthetic Data 2.5.1. Work has also been done on the MT itself, such as considering the translation quality for Model Distillation 2.5.5. We have also experimented with lexically and grammatically Constrained Machine Translation 2.5.6 and dealt with the theoretical/implementation issues such as catastrophic forgetting 2.5.7. Robustness of Machine Translation to Noisy Inputs is covered in 2.5.9 and Length Generalization Issue in 2.5.8. Finally, the lack of training data was tackled with an investigation into Unsupervised Machine Translation and Sentence Embeddings 2.5.11. Speech machine translation (SMT) also took a significant portion of the team's energy. IWSLT 2021 shared task (5.1.1) was crucial both from the viewpoints of the organization by organizing the Shared Tasks in Machine Translation and Speech Translation and organizing another task in Machine Translation Evaluation. We have also provided pre-trained models for machine translation metric quality, pre-trained machine translation metric quality, and performed an analysis in Neural Machine Translation Quality estimation. Last but not least, also within IWSLT 2020, we worked in Non-Native Simultaneous ASR and SMT 2.6.1.

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 individual tasks specified in the project proposal for the first period (Section 1.3):

- Target-application aware neural signal processing
- Speaker Recognition and Diarization
- Automatic speech recognition
- Between speech and NLP
- Neural machine translation
- Speech machine translation
- Multimodal approaches
- Towards Semantics
- Human Performance and Human Interfaces

The details are in the following sections.

2.1 Target-application aware neural signal processing

2.1.1 DNN Speech enhancement for robust speaker recognition

In [Nov+19a], we have conducted an extensive study focused on building a speaker recognition system that is robust against adverse channel effects, noise and reverberation. Specifically, we were exploring the use of DNN-based autoencoders for speech enhancement, denoising and de-reverberation to achieve the robustness alone or in combination with standard methods, such as multi-condition training when we try to include various noisy conditions in the training of speaker recognition backend, such as Probabilistic Linear Discriminant Analysis (PLDA).

Our autoencoder consists of three hidden layers with 1500 neurons. The input of the autoencoder is a central frame of a log-magnitude spectrum with a context of +/- 15 frames (in total 3999-dimensional input). The output is a 129-dimensional enhanced central frame log-magnitude spectrum, see the topology in Figure 2.

As an objective function for training the autoencoder, we used the Mean Square Error (MSE) between the autoencoder outputs from training utterances and spectra of their clean variants. We were using both clean and augmented recordings (added noise and reverberation) as the DNN input during the training as we wanted the autoencoder to keep its robustness and produce good results also on relatively clean data.

We confirmed via numerous experiments, that DNN speech enhancement improves the robustness of speaker recognition system and can be combined with other techniques (such as multi-condition training) to further improve the performance.

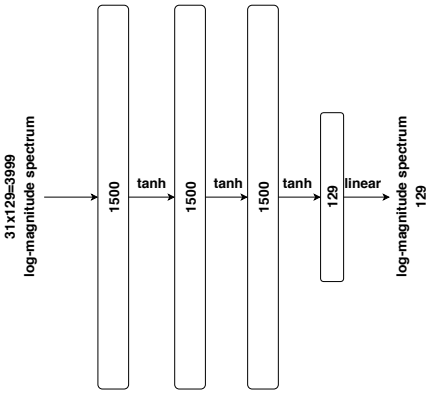


Figure 2: Topology of speech enhancement autoencoder.

2.1.2 Spectral augmentation for embedding learning

It has been experimentally confirmed that data multiplying the training data via augmentation (adding noise, reverberation, music, etc.) is important during speaker embedding training on publicly available datasets (e.g. Voxceleb or conversational telephone data available from NIST SREs). In [Wan+20], we were exploring a simple technique which is based on on-the-fly randomly masking bands in the log Mel spectogram during the embedding extractor training. Such method has been previously successfully applied for speech recognition. We have performed an extensive experimental study covering both narrowband telephone (NIST SRE16) data and wideband conversational data (voxceleb) with state-of-the art architectures of embedding extractors, namely 1-D convolutional TDNN and 2-D convolutional ResNet34, trained with either Softmax or AAMSoftmax loss. Our experiments confirmed, that spectral augmentation is a viable option to the classical offline augmentation via adding noise, music and reverberation. Systems trained with spectral augmentation had in some cases better performance or performed similarly to the baseline augmentation method. We see the main benefit in simplified training and smaller footprint of the training data.

2.1.3 Channel-wise correlation based pooling

In [SRB21], we explored an alternative pooling mechanism for NN based speaker embedding extractors. Inspired by works in computer vision, this pooling mechanism calculates the *channel-wise* correlations of the output of the last layer before pooling (recall that the standard pooling mechanism is to calculate the mean and/or the standard deviation of the channel-frequency pairs). An important difference between our work and the works in computer vision is that we calculate the correlation matrix by pooling only along the time-axis, i.e., not along the frequency axis, whereas in computer vision, pooling is done along both axes of the image. The reason for this difference is that the statistical properties of speech are not invariant along the frequency axis. Accordingly we have one correlation matrix for each frequency component. This increases the dimension of the pooling output and to mitigate this we merged the statistics of adjacent frequency components. In experiments we show that the the proposed pooling mechanism substantially outperforms a state-of-the art baseline on the Voxceleb benchmark.

2.1.4 Phase Encoding

The widely used magnitude spectrum based features have shown their superiority in the field of speech processing. In contrast, the importance of phase spectrum is always ignored. This is because the patterns hidden in phase cannot be intuitively modelled and interpreted, due to phase wrapping phenomenon. To address this problem, we explore novel phase spectrum based features, named Learnable Group Delay (LearnGD) [Pen+21a], to capture useful information in speech signals. Specifically, firstly, the negative of the spectral derivative of the phase spectrum, called group delay (GD), is used to unwrap the phase. Then, to suppress the spiky nature of GD, which is caused by its roots close to the unit circle in the Z domain, a carefully designed light convolutional smoothing layer is employed to reconstruct the GD. Finally, an exponential hyper-parameter is introduced to reconstruct GD features to restore the spectrum range and generate LearnGD features.

We observed that with the same feature extractor, the real and imaginary based system outperforms the magnitude spectrum based system by 0.2% EER. Compared to phase spectrum, group delay achieves a relative improvement of 16% in EER. This means that besides the widely used magnitude-based feature, the phase-based features can also encode some speaker-related information.

2.1.5 Interpretable Complex Filter

Recently, extracting speaker embedding directly from raw waveform has drawn increasing attention in the field of speaker verification. Parametric real-valued filters in the first convolutional layer are learned to transform the waveform into time-frequency representations. However, these methods only focus on the magnitude spectrum and the poor interpretability of the learned filters limits the performance. To solve this problem, we propose a complex speaker embedding extractor [Pen+21b], named ICSpk, with higher interpretability and fewer parameters. Specifically, at first, to quantify the speaker-related frequency response of waveform, we modify the original short-term Fourier transform filters into a family of complex exponential filters, named interpretable complex (IC) filters. Each IC filter is confined by a complex exponential filter parameterized by frequency. Then, a deep complex-valued speaker embedding extractor is designed to operate on the complex-valued output of IC filters.

We tested ICspk with ResNet34 using different types of input acoustic features, including magnitude, concatenation of real and imaginary parts of complex spectrogram and raw waveform. Firstly, using real-valued ResNet34 as front-end model, "Real+Imag" based system outperforms "Magnitude" based system (i.e. 2.34% v.s. 2.51%), implying that the phase part is also embedded with speaker related information that has been neglected. Learning band-pass filters (sinc-conv) outperforms the magnitude based system, while exhibiting slight worse performance than "Real+Imag" based system. Replacing the real-valued speaker embedding extractor (ResNet34) with the proposed $\mathbb{C}$ResNet34, 14% relative improvement (i.e. 2.02% v.s. 2.34%) is achieved with the same number of parameters.

2.1.6 Voice conversion

Voice conversion is a task of changing properties of utterance in such a way, that it sounds as another speaker. Voice conversion can be divided into several scenarios, based on its generality. One-to-one models achieve conversion only on one pair of speakers which they were trained on, many-to-many models change characteristics to and from any speaker in a training data-set. The hardest scenario in voice conversion field is any-to-any conversion, where the goal is to be able to modify voice of and to any speaker. Typically, two main approaches are used, one is direct conversion, where speech characteristics are changed continuously in several layers of neural network. The second approach is based on disentanglement, where speech is decomposed into mutually independent components. These components are typically speaker characteristics in form of speaker embedding (this part of the system is frequently trained separately) and some form of content characteristics, which is in form of sequence of content embeddings. Recent publications also try to use prosody information and fundamental frequency contour.

Our work focuses on improving level of disentanglement using adversarial speaker classifier in bottleneck layer. This modifications benefits are two-fold. First, it allows information bottleneck, which is usually used, to be bigger, therefore to increase amount of content information preserved. Second, it further reduces amount of information of original speaker in utterance.

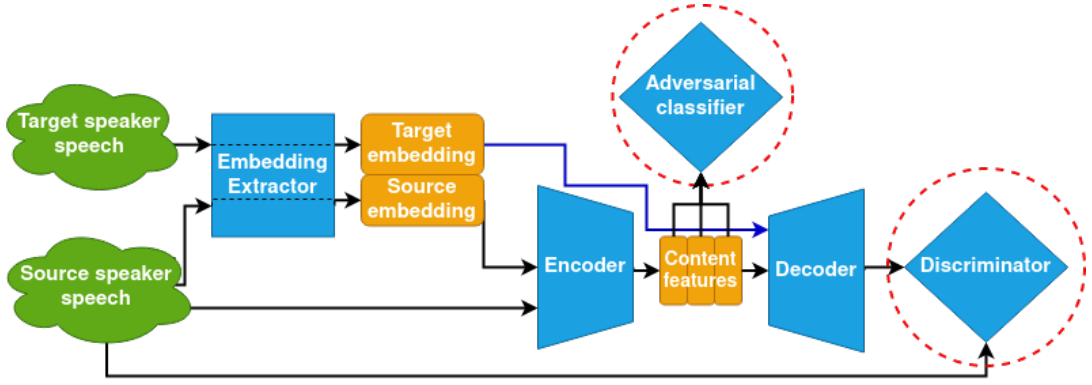


Figure 3: Scheme of voice conversion model with adversarial losses.

Our next modification comes from the fact, that model, which we use, is based on auto-encoder architecture. Therefore, converted speech's spectrogram is over-smoothed, which results in poor quality of voice. We address this issue by using additional discriminator network and training the whole system also with GAN objective. Whole system is shown in Figure 3. This work was submitted for ICASSP 2022, unfortunately, it was not accepted and it will be submitted to next conference after incorporating reviewers' remarks.

2.1.7 SpeakerBeam for target speech extraction

Speech technologies often suffer from problems when presented with an overlapping speech of multiple speakers. In our work [Žmo+19], we aim to alleviate this problem by pre-processing the mixture of multiple speakers and extracting only the signal of a target speaker. The target speaker is determined by an additional adaptation utterance,

i.e. short utterance spoken by the target speaker only. We propose a method called SpeakerBeam based on a neural network, which takes the mixture and the adaptation utterance at the input and outputs the extracted target speech. The overall scheme of the method can be seen in Figure 4. This method can be contrasted with more common blind speech separation methods that extract signals of all speakers in the mixture. Formulating the problem as target speech extraction instead avoids certain issues such as label permutation and the need to determine the number of speakers in the mixture.

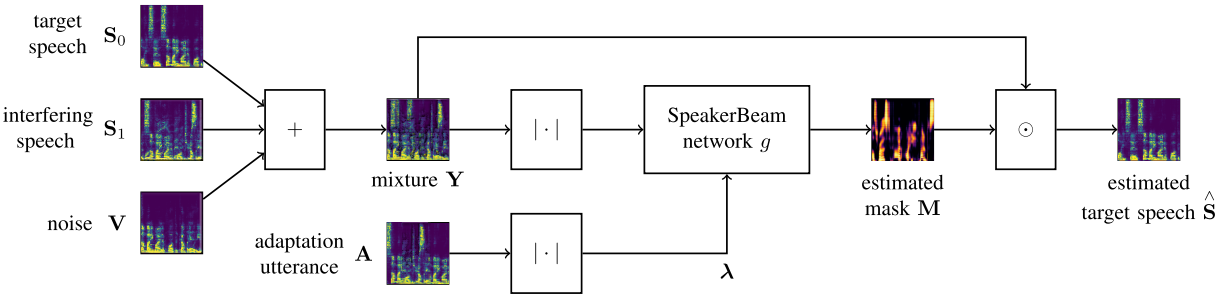


Figure 4: Scheme of SpeakerBeam for target speech extraction.

In [Žmo+19], we compare different ways to extract speaker embedding from the enrollment utterance and different ways of using this embedding to inform the neural network about the target speaker. For this, we make use of ideas from speaker verification and speaker adaptation fields. SpeakerBeam is compared with two speech separation methods, Permutation invariant training, and Deep clustering, and shows an advantage, especially on longer utterances. We further analyze the model and show that the speaker embeddings learned by the network properly encode the speaker and especially gender information.

2.1.8 Integration of variational auto-encoders and spatial clustering for multi-channel speech separation

In [Žmo+21], we look at the problem of speech separation from a different perspective. We focus on multi-channel speech separation, which makes use of spatial information obtained from multiple microphones. For this, we use spatial clustering methods and combine them with a spectral model represented by a variational auto-encoder. By doing so, we can exploit the modeling power of neural networks, but at the same time, keep

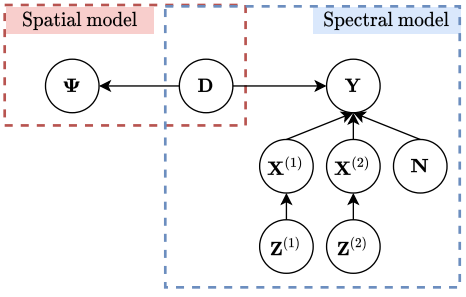


Figure 5: Overall scheme of the model combining variation auto-encoder as a spectral model with spatial clustering.

a structured model, as shown in Figure 5. The structured model is more interpretable than methods based on discriminative neural networks. Such a model can be advantageous when adapting to new noise conditions as only the noise part of the model needs to be modified. We show experimentally, that our model significantly outperforms the previous factorial model based on the Gaussian mixture model, performs comparably to the integration of permutation invariant training with spatial clustering, and enables us to easily adapt to new noise conditions.

2.2 Speaker recognition and diarization

2.2.1 Embeddings for speaker recognition

A *speaker embedding* is a fixed sized representation of an utterance that captures as much as possible of the information related to the speaker identity and, ideally, as little as possible of other information such as properties of the recording device and the environment as well as the physical and emotional state of the speaker. In the last five years, the state-of-the-art paradigm for speaker embeddings has transitioned from being based on Gaussian mixture models (GMM) to being based on neural networks (NN). The work on speaker embeddings within the Neurem3 project has been part of our efforts in this direction.

Initially, most of our efforts went into *end-to-end* NN based based speaker verification systems. This work is described in [Roh+19]. While end-to-end training is theoretically appealing, it eventually became apparent that multiclass discriminative training (MDT), i.e. the training objective is to identify who is speaking among all speakers in the training set, works better in practice. Much of our efforts focused on analysis and improvement of such training. In [Nov+19b] we analysed how MDT performs when applied to a classic i-vector model, i.e., the predecessor to NN based speaker embeddings. In [Wan+19] we combined MDT with training objectives that aims to amplify and/or attenuate phonetic information at various stages of the NN. In [Sta+19] we proposed a new training objective suitable for unlabeled data which also can be combined with MDT in the case of labeled data. Finally, in [Mat+20a] we analysed different speaker recognition systems that have been state of the art throughout history, including GMM and NN embedding based ones. The above works are described in the following paragraphs in the order they were introduced above.

Early work on NN based speaker embeddings suggested that the training of models for embedding extraction was prone to overfitting. In [Roh+19], we therefore proposed an NN based speaker verification system that is initialized to mimic the whole speaker verification pipeline from features to score (end-to-end) of a traditional system based on i-vectors and probabilistic linear discriminant analysis (PLDA). As such, the system has three distinct modules; a *feature to statistics* (**f2s**) module, a *statistics to i-vector (embedding)* (**s2i**) module, and finally discriminatively trained PLDA (DPLDA) as a *i-vector to score* module. The system was then further trained in an end-to-end manner, i.e., all modules of the system are trained jointly for the task at hand, namely speaker verification which means that the system should tell whether two speech segments are spoken by the same speaker. During training, the system was regularized so that it did not deviate too far from the initial system. In this way we mitigate overfitting which normally limits the performance of end-to-end systems.

The main conclusions are:

- The proposed systems outperforms the baselines for most metrics and data sets
- Replacing individual modules with NNs may deteriorate the performance
- Joint training was effect only for the first two modules, likely due to computational constraints in joint training of the three modules.

2.2.2 From i-vectors to x-vectors

In [Nov+19b], we focused more deeply on the embedding extractor part of our end-to-end approach. The aim of this work was to develop an embedding extractor with similar mathematical form as an i-vector extractor but with fewer parameters similar to state-of-the art NN embedding extractors at the time. This was achieved through a factorization of a standard i-vector model. In this way, we could effectively apply multiclass discriminative training that had proven effective for training NN based embedding extractors also to the i-vector extractor. This is important from an analysis point of view because it helps understand whether the excellent performance of the NN embeddings was due to the training objective or to the model. The developed model, was also a practical alternative the embedding extractor (**s2i**) in the end-to-end system. In conclusion, MDT of i-vector extractors improves over generative training (GT) for the full model. Further, the factorized model with much fewer parameters trained with MDT is as good as the full model trained with GT.

2.2.3 Phonetic information in SR embeddings

In [Wan+19] we explored how to best use phonetic information when training the embedding extractor. By this time, state of the art architecture for speaker recognition had evolved to consist of an NN block that processes framewise features, followed by a *pooling mechanism*, for example taking the mean, over the frame-wise outputs of one or more layers in the NN. The pooling layer is then processed further with another NN block. Finally a softmax layer is used for speaker classification. The speaker embedding is then the output from one of the intermediate layers of the last NN block. Earlier work had shown that it is effective to add phonetic information before the statistics pooling layer via frame-level multitask training. However, intuitively, for the text-independent speaker recognition task, phonetic information in the embeddings is just an extra source of noise. We therefore explored a training scheme were phonetic information is added before to the framewise layer by with multitask training, while being removed from the speaker embedding using adversarial training.

The main conclusions are:

- Multi-task training on its own helps no matter where it is applied
- Adversarial training helps if it is applied at the segment level but not if it is applied at the frame-level
- The combination of multi-task training at the frame-level and adversarial training at the segment level is overall the best training strategy

2.2.4 Self-supervised training of SR embeddings

In [Sta+19], we explored a *self-supervised* approach to train an embedding extractor. Here the training objective is to reconstruct the features of a speech segment using only phoneme labels and a speaker embedding from another speech segment of the same utterance. The model for extracting the speaker embedding is updated based on the reconstruction loss. Since both speaker and phoneme information is needed for accurate reconstruction, this forces the embedding extractor to create embeddings that contains speaker information. This training objective can be used unsupervised as well as in combination with the speaker classification loss.

Interestingly, the self-supervised training performs very well on its own although it should be noted that supervision was still used in the backend (PLDA) in this experiment. It is also apparent that the combination of the two objectives is clearly the best.

2.2.5 Longitudinal analysis of SR systems on historical and current data

Finally, in [Mat+20a] we analysed different speaker recognition systems that have been state of the art throughout history. This includes GMM embedding based system such as i-vector and the joint factor analysis (JFA) as well as x-vector which is an NN based embedding. This study focused on how speaker recognition technology had improved throughout time, and to what extent the improvements were due to better modelling technique or more data. The main conclusions were that NN based embeddings almost always outperformed the GMM based ones, even for the relatively small training data sizes that were available more than a decade ago.

2.2.6 Speaker diarization based on Bayesian HMM with eigenvoice priors (BHMM)

The method is a Bayesian approach to speaker diarization, where the sequence of speech features representing a conversation (frame-by-frame MFCCs) is assumed to be generated from a Bayesian hidden Markov model (HMM). In our model, HMM states represent speakers and the transitions between the states correspond to speaker turns. The speaker specific distributions are modeled by Gaussian mixture models (GMMs). In order to robustly learn the speaker-specific distributions, a strong informative prior is imposed on the GMM parameters, which makes use of eigenvoices just like i-vectors or joint factor analysis (JFA) – the standard techniques for speaker recognition. Such prior facilitates discrimination between speaker voices in an input recording. The proposed Bayesian model offers a very elegant approach to speaker diarization as a straightforward and efficient variational Bayes (VB) inference in a single probabilistic model addresses the complete speaker diarization problem.

In [Die+20], we provided a full description of the model, giving derivations of all update formulae. An extensive analysis for all model parameters was made, including also the newly proposed speaker regularization coefficient which provides extra control on the number of inferred speakers.

2.2.7 Bayesian HMM clustering of x-vector sequences (VBx)

The VBx diarization method [Die+19] was developed as a simplified version of the original Bayesian HMM with eigenvoice priors [Die+20]. Unlike the former BHMM, which operates on a frame-by-frame basis, VBx clusters sequences of x-vectors, which allows a simpler probabilistic linear discriminant analysis (PLDA) based model for modeling speaker distributions. The combination of the Bayesian inference with the discriminative power of x-vectors and the simplicity of the model makes the method very efficient and powerful.

In [Lan+22], we presented the derivation and update formulae for the 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.

In [Lan+22], we also pointed out the lack of a standardized evaluation protocol for AMI dataset and we proposed a new protocol for both Beamformed and Mix-Headset audios based on the official AMI partitions and transcriptions, which was nicely adopted by the community: several works and open source diarization toolkits considered and/or switched to this protocol[1][2].

Together with the publication, we released both the open source code to train the x-vector extractor network[3] and the diarization recipe[4] (see Section 8.1 for more details on the software).

We have successfully used VBx diarization in several diarization evaluations: we were the winners of the second DIHARD diarization challenge (on all four tracks), obtained the second position in the diarization track of VoxCeleb challenge (on the VoxConverse dataset) and ranked 5th (on a very close contest between 3rd, 4th and 5th positions) in the third DIHARD diarization challenge. More details about our participation in evaluations can be found in Section 5.1.2.

It is worth to highlight that from June 2019 to January 2021, Alicia Lozano-Diez joined the team with the Marie-Curie grant "Robust End-To-End SPEAKER recognition based on deep learning and attention models". Part of her work focused on exploring end-to-end neural diarization (EEND) approaches and significantly contributed to improving the state-of-the-art of the group.

2.2.8 Probabilistic embeddings for speaker diarization

In traditional approaches to speaker verification or diarization, every speech segment is assumed to be represented as a single vector – embedding. Then, the back-end model works with the embeddings as with the observed data. This approach has its benefits, it represents both training and test data in a compact way, allows for relatively simple back-end models to be used and, with some back-end models (i.e. heavy-tailed PLDA), there is a possibility for uncertainty propagation. However, the uncertainty, in this case,

[1] https://github.com/kaldi-asr/kaldi/tree/master/egs/ami/s5c
[2] https://github.com/pyannote/AMI-diarization-setup
[3] https://github.com/phonexiaresearch/VBx-training-recipe
[4] https://github.com/BUTSpeechFIT/VBx

is contained in the distribution of the embeddings, not in the individual embeddings themselves.

In [Sil+20], we propose an alternative approach to use uncertainty information in speaker diarization (similar approach can be used for verification): we consider the embeddings as hidden variables in the model. The desired property for these hidden variables is that for high-quality audio segment the posterior distribution for the embedding is sharp, while for short or noisy segments it should be flat. In other words, we assume the existence of some ideal true embedding; one should be certain about it for high-quality audio and be uncertain where the embedding is for low-quality recordings. In our work, we assume that the embedding distribution is a normal distribution with parameters depending on the observed speech segment (sequence of MFCC features). To estimate these parameters we utilize existing pre-trained x-vector extractor. We augment x-vector neural network with an additional block responsible for extracting the precision of the probabilistic embedding, while the original x-vector serves as the mean of probabilistic embedding. We also propose to modify widely-used PLDA model to be able to incorporate it into probabilistic embedding framework. We train the model discriminatively to cluster sets of several speech segments into speaker clusters. When training it, we jointly learn the parameters of PLDA model and of the neural network extracting parameters of the probabilistic embedding distribution. We tested this model on DIHARD 2019 diarization task. Diarization was performed by using Agglomerative Hierarchical Clustering (AHC) where we used LLR scores computed with the baseline PLDA or our discriminatively trained PLDA with or without embedding uncertainty as pair-wise similarities between speech segments. The results were generated with stopping threshold for AHC set analytically and optimally.

We conclude that if one is allowed to set the stopping threshold optimally, the performance of the baseline and both discriminatively trained models is similar. However, when the threshold is set analytically, discriminative PLDA with embedding uncertainty provides the best performance. Moreover, for this model, the gap in performance between oracle and blind thresholds is the smallest, indicating that this model provides better calibrated LLR scores then the two other models.

2.3 Automatic speech recognition

2.3.1 Data for ASR

Large Speech Corpus for Czech Neural speech recognition systems are heavily dependent on the volume of training data compared to the more traditional hybrid systems. For some languages, e.g. English or Spanish, there are enough corpora available while for the majority of the languages there is a very limited number. To bridge this gap for Czech, we processed, published and released ParCzech 3.0 — a Large Corpus of Czech Parliament Plenary Hearings [KPB20]. The corpus consists of approximately 400 hours of speech data and corresponding text transcriptions. The whole corpus is segmented into short audio segments making it suitable for both training and evaluation of ASR systems. We also demonstrate that the dataset is useful for training both, the neural end-to-end and hybrid ASR systems. The release is focused both on detailed metadata annotation (sessions, speakers, etc.) as well as on an easy exploitation for training of ASR systems. Specifically, we designed and applied a sensitive filtering technique to

allow corpus users select only speech segments of a sufficient quality, incl. the match between the words uttered and the official transcript.

Speech test set with additional relevant texts In [Mac+19], we created and published Antrecorp, a transcribed speech test set of students' practice business presentations that are supplemented by additional relevant texts to the presentation topic, such as slides and webpages. The speeches are in English, the speakers are second language learners of English, often having a strong non-native accent. The recordings contain strong background noise. The speaker's metadata about their country of origin are preserved. The test set is therefore suitable for analyzing robustness of ASR and speech translation against noise and non-native accents.

2.3.2 Low-resource speech recognition

Low-resource speech recognition for air traffic domain: Although there is a track of applied research in the area of ASR applied to Air Traffic Communication, the speech technology is not yet used much in practice. One of the reason is the relatively limited amount of data. In [Zul+20] we prepared several existing in-domain databases, and merged them into three sets of training data: all databases and two subgroups. As part of it we did some data standardization. We trained several types of ASR systems, we aimed at also cross-generalization between the source databases. In the scenarios we achieve WER between 5% to 15% trained with the merged database. The main issues of ATC speech recognition are noise in the radio transmission, very fast speech and many multi-lingual accents of English.

Later, we experimented with using contextual information (call-sign lists from air-traffic monitoring) to improve the accuracy of ASR transcripts. The mechanism is that we give score discounts to some rare 'words' in the lattice-generation step, or we give score discounts to some 'word sequences' by rescoring lattices of alternative hypotheses. The call-sign lists come from the air-traffic monitoring databases of OpenSky Network. We further improve the ASR system by using untranscribed data in semi-supervised training of the acoustic model.

Lightly-supervised ASR training for Spanish television data [Koc+21] describes the ASR effort for the Albayzin 2020 Challenge. We tested a neural-based music separator to filter-out background music from the spoken parts of TV shows. We also describe our efforts in retrieving high-quality transcripts for training, from a set of TV captions that were available. A fusion of our best systems achieved 13.3% and 23.24% WER on RTVE2018 and RTVE2020 test sets respectively.

Our end-to-end fully-neural ASR model shows relatively competitive performance on the RTVE2018 test set in comparison with its hybrid HMM-based counterpart. However, its performance on the RTVE2020 evaluation set exposes that the model was not able to generalize very well since this database turns out to contain slightly different acoustic conditions. Despite of this fact, the model still managed to improve the results in the final fusion with hybrid systems. An exploration on background music removal shows that it yields the best results for lower SNR ranges, thus having a different impact depending on the acoustic conditions of each TV show.

Speaker adaptation for ASR using x-vectors In [Kar+21], we studied usability of x-vectors for speaker adaptation of automatic speech recognition systems. X-vectors are Neural Network based speaker embeddings recently proposed in speaker recognition. The x-vectors get concatenated to the input of the neural network. All experiments were done on ASR for the latest IARPA MATERIAL evaluation running on Pashto language.

Extensive analysis shows suitability of this technique for low resource ASR even if the target language is not part of x-vector training data. The x-vectors trained on sufficient amount of well balanced telephone data show robustness to channel and language mismatch. They overcome baseline i-vectors by impressive 2% absolute gain. The obtained improvements are persistent when a significantly more complex ASR system is used. Over 1% absolute improvement was observed with x-vectors over traditional i-vectors, even when the x-vector extractor was not trained on target Pashto data.

2.3.3 Multi-Source ASR

Another way how to tackle the scarcity of training data for a particular language is transfer learning [Tan+18]. In particular, in [PB21] we explored different strategies for transfer learning for Czech ASR. We proposed a coarse-to-fine transfer where the training starts with an intermediate "coarse" alphabet (i.e., the Czech alphabet without accents). During the training, we switch to the full alphabet. We compared this proposed approach with a vanilla cross-lingual transfer from English (i.e., the Czech ASR model is initialized with weights from an English model). The newly proposed transfer outperformed the vanilla by almost 4 % WER absolute on the test set (16.57 vs. 20.19). We also tested the proposed method with a combination of cross-lingual and coarse-to-fine transfer. Both approaches performed on par. Though, an important observation is that the combination of the methods converged very quickly, which is interesting for example in restricted computational resources scenario.

2.3.4 Data-driven approaches for ASR

Self-supervised training for E2E ASR Self-supervised ASR-TTS model is a recently proposed cycle-consistency based training regime to handle low-resource condition in E2E ASR. The ASR-TTS model jointly trains ASR and TTS model as a whole differentiable pipeline. Although the model provides improvements on in-domain low-resource condition, it suffers under out-of-domain data conditions. To overcome this constraint, we propose an enhanced ASR-TTS (EAT) [Bas+21] model that incorporates the following main features:

- We incorporate a pre-trained RNNLM regularization term in the ASR REINFORCE loss for speech only (SO) training, increasing its robustness to bad latent ASR hypotheses. This is named as Language Model Penalty (LMP).

- We introduce an α hyper-parameter for text-only (TO) training, to attenuate the influence of the ASR encoder by scaling the attention-encoded context. This allows us to reduce the focus on acoustic information when the latent speech quality is poor, effectively alternating between ASR and a more language-model-like behaviour.

- We incorporate latest training strategies and architectures such as data augmentation and data annealing. The TTS module is also built using the Transformer architecture, which shows higher robustness and memory efficiency. Multi-head attention is used as ASR encoder layers to attain additional gains and for reduced model complexities.

- We show that these techniques greatly improve performance and particularly attain the target goal achieving good performance in limited data and out-of-domain scenarios with cycle consistency techniques.

The experimental analysis shows the impact of SO, TO and speech+text only (ST) training on BABEL-Pashto dataset. The ST pipeline is trained using a multi-task objective with both cross-entropy loss from TO and REINFORCE loss from SO. The results highlight the importance of α scaling during TO training as it mitigates the effect of badly synthesized speech and learns context from unsupervised text data. Training with ST provides additional gains compared to SO and TO pipeline. In addition the EAT model is shown to be complementary to specaugmentation technique.

Data-driven units for speech processing

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

In [Ond+19], we proposed the Subspace Hidden Markov Model (SHMM) for AUD. The SHMM is a generative model of speech which assumes the speech in a language is generated from a (potentially unbounded) number of discrete units each of which can be modeled by a Hidden Markov Model (HMM). The crux of the SHMM is that the parameters are forced to dwell in a low-dimensional subspace of the parameter space, i.e., we ensure that the parameters of each HMM all dwell in the column space of some low-rank matrix. Each unit's parameter vector is generated by multiplying the low-rank matrix with a corresponding low-dimensional embedding vector. The training procedure for the SHMM comprises two phases: in the first phase, transcribed data from some source languages is used to estimate the subspace matrix along with the embeddings of the phones of the source languages. The learned subspace is fixed and transferred for any new target (unannonated) language where new unit embeddings are inferred for the target language in an unsupervised manner. Thus the units of the target language are constrained to resemble phones from the source languages.

The SHMM assumes that the parameters of the units can be constrained to a single subspace regardless of the units. Hypothesizing that this is too stringent a constraint, we proposed the Hierarchical Subspace Hidden Markov Model (H-SHMM) in [Yus+21]. The H-SHMM is a hierarchical model which allows us to adapt the subspace to each language without completely removing the constraint on the parameter space. We achieve this by introducing a "hyper-space" of subspaces. Just as the SHMM constrains HMM parameters to dwell on a subspace, the H-SHMM constrains subspaces to dwell on another, hyper, subspace. The low-rank matrix with which unit embeddings are to be

multiplied is itself generated by multiplying another low-rank matrix with a *language* embedding and reshaping the resulting vector into the desired matrix dimensions. H-SHMM training follows a similar procedure to SHMM training with the caveat that it is now the subspace of subspaces which is transferred from the source languages. In the target language, both a language embedding and unit embeddings are jointly inferred with an unsupervised procedure. The former, in combination with the hyper-subspace, defines a parameter subspace for that language. The latter define the actual units which dwell on that subspace.

We report the AUD performance of various systems in terms of normalized mutual information (NMI) and F-score for three target languages: English, Mboshi and Yoruba. The baseline is an HMM-based AUD system without the subspace constraining its parameters.

We observe that the SHMM significantly outperforms the unconstrained HMM in all languages and on all metrics showing that constraining the parameters with a subspace is allows us to learn more meaningful units. Furthermore, we observe that the H-SHMM provides consistent improvements over the SHMM giving credence to the hypothesis that it is beneficial to adapt the subspace to each target language.

2.4 Between speech and NLP

2.4.1 Neural handling OOVs in speech recognition

Out-of-vocabulary words (OOVs) pose one of the persistent problems in ASR. In [Ego+21], we explore the effectiveness of detecting OOVs in a word-predicting end-to-end neural network system trained to predict words. We experiment with attention and CTC architectures and use two approaches to finding the position of OOVs: one is using attention matrix and another is using CTC alignment.

We show that CTC is more precise in the task of producing the start and end times of OOVs than attention information. When detecting OOVs with CTC training, we're able to achieve 81.5% recall and 35.3% precision, whereas attention requires time shift and can only reach 34.4% recall and 14.9% precision for the detection task.

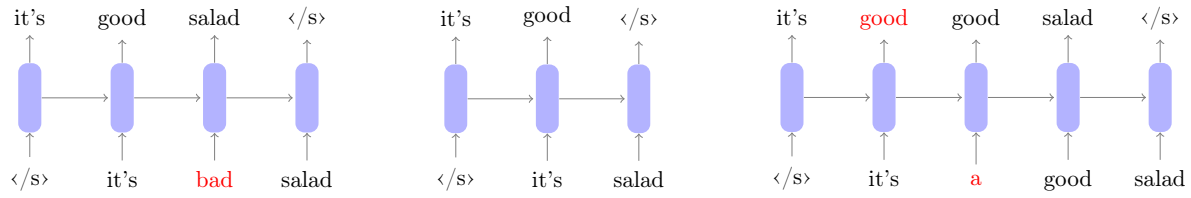
We also process the output of OOV detection step with an OOV recovery pipeline and assess the effect detection has on recovery task. For the recovery, phoneme strings from the detected OOV time frames are extracted from a phoneme recognition system and clustered in an attempt to recover the pronunciation of repeating OOVs. Attention output has poor performance in recovery task (5.5% recall), whereas CTC detection provides a much better input for recovery task (14% recall).

2.4.2 Language modeling for speech and OCR

Previous work on semi-supervised adaptation of language models suggests, that a careful selection of sentences is needed to obtain good quality of the final model. We reexamine this assumption in context of Optical Character Recognition (OCR) in [KBH21]. We use a seed system trained on data from related domain and a small amount of target domain data. Then, we use the seed system to obtain machine annotations of the rest of the target domain. Evaluating on challenging printed and hand-written datasets, we see

that, with a single exception, the best quality of LMs is achieved by simply considering all of the machine-annotated data.

The improvements in language modeling are relevant in the big picture, e.g. for the more difficult handwritten data system without LM achieves 6.43 % character error rate (CER), introducing a generic LM reduces the CER to 5.43 % and finetuning the LM on the machine-annotated data brings the figure down to 4.17 %. Practically, this allows the user of an OCR system to benefit greatly even from modest human effort on transcribing their specific data.



(a) Substitution (b) Deletion (c) Insertion

Figure 6: How we introduce errors for training of neural LMs. The original sentence is "*it's good salad*".

In automatic speech recognition, we have explored data augmentation for language modeling motivated by high error rates in some application scenarios, where the correct transcription may not even be present in 3000-best outputs from the system. To mitigate the impact on LM, we proposed to train it on similarly corrupted data [BB21]. We have compared several different ways of introducing the errors into the training data and to our great surprise, the best results were achieved with a simple data-agnostic scheme, where we only control the overall amount of substitutions, deletions and insertions. Specifically, it increased the absolute improvement in Word Error Rate brought by neural LM from 1.1 % when trained with clean data to 1.9 % when trained with our augmentation scheme.

Through additional experiments, we have established that it is the random nature of the input 0-gram augmentation what makes it so successful. From this and the previous experiments on OCR, we conclude that LMs are rather robust to noise in training data and can actually benefit from it. For ASR, this means that there is a very simple and computationally cheap, yet efficient data augmentation. We have confirmed its effect on other data too, though the improvements do decrease as the task becomes easier.

2.5 Neural machine translation

2.5.1 Achieving Human-Level Adequacy by Distinguishing Genuine and Synthetic Data

Our paper [Pop+20] is the result of several years of previous research on English-to-Czech machine translation. Gradually, over a decade, we have built and refined our large training corpus, studied properties and difficulties when translating from a morphologically poor language to a morphologically very rich one. Shortly before the start of NEUREM3, we carried out and extensive hyperparameter search for the Transformer architecture [PB18].

In NEUREM3, we continued and focused on another parameter of training: what is the best use of genuinely parallel texts and of monolingual texts. In NMT, monolingual texts are utilized not to train a separate language model, but rather to extend the parallel data by creating a synthetic version of the missing side [BT11; SHB16]. In [Pop+20], a new training regime was proposed that switches training on genuine and synthetic parallel data after longer periods of training, thereby moving between two distinct optima. By averaging model checkpoints with a certain balance of checkpoints saved during the genuine phase and checkpoints saved during the synthetic phase, the model obtains outstanding properties.

In a very careful manual evaluation that included larger context of the document, the system was found to surpass human reference translation (created by a translation agency) in terms of adequacy. Put simply, both humans and MT make adequacy errors but in this case, MT was more careful in details.

2.5.2 Analysis of the Utility of Linguistic Annotation for Transformers

Transformers are exceptionally effective in machine translation, in part because they have the capability of direct modelling of relations between words, i.e. some form of empirical syntax. Based on existing treebanks and parsers, we wanted to find out if aligning this unsupervised form of syntax with the linguistically adequate one is beneficial for translation quality. The results were surpsing.

Promoting source syntax in Transformer NMT is not needed In [PMB19b], we proposed two ways of enriching source syntax in Transformer NMT: simple alternating multi-tasking with POS and dependency parsing, and retrieving a dependency parse tree in a form of a matrix from one of the Transformer's attention head.

In the first way, it turns out that alternating multi-tasking with a linguistic task is beneficial for the MT task more than a simple linguistically unrelated contrastive task, but the overall multi-tasking cost outweighs the benefits, so that MT with multi-tasking is worse than single-tasking baseline.

In the second examined way, we found out that learning jointly to parse and translate improves a baseline MT quality. However, contrasting parsing task with a dummy linear tree that does not bear any linguistic information gains the same MT quality enhancement. We therefore conclude that the improvements were not from promoting syntax, but from other, unintended and non-investigated reasons.

Multi-tasking to Predict Target Syntax along with Translation In [KCB19] we focused on target sentence syntax. If the model is explicitly considering target sentence structure, it may be in better position to preserve e.g. various long-distance dependencies. This intuition has been already explored by [Nad+17].

As expected, we observe gains in translation quality from predicting target syntax in the form of interleaved output: a single model produces one sequence of target tokens, which are alternating between the words in the sentence and their CCG tags, i.e. a compact representation of their syntactic role in the target sentence.

The surprising observation is that training on *random* CCG tags is equally beneficial. In other words, it is not the explicit modelling of target syntax, but rather some other

difference in the model that improves translation quality. Our intuition is that the decoder, when producing twice as long sequences, effectively has a higher number of layers during which it can refine the choice of the next output word.

2.5.3 Relation between MT System Quality and Time Savings in Translators' Workflow

It has been shown for phrase-based machine translation that the better the MT, the more time is saved during professional post-editing. In [Zou+21b] we focused on state-of-the-art neural MT and performed an experimental study involving over 30 professional translators for English → Czech translation. We examined the relationship between NMT performance and post-editing time and quality. Across all models, we found that better MT systems indeed lead to fewer changes in the sentences in this industry setting. The relation between system quality and post-editing time is however not straightforward and, contrary to the results on phrase-based MT, BLEU is definitely not a stable predictor of the time or final output quality.

2.5.4 Document-level Transformer

In [Mac+20], we describe CUNI submission for WMT 2019 news translation task. We propose a way to exploit hierarchies of text units in NMT, namely by using a larger context than one sentence during translation. In our system, we translate overlapping windows of N sentences at once, and consider only some part of the window for the finalized translation.

2.5.5 Model Distillation Considering Translation Quality

Model distillation is a way to train smaller and more efficient models (possibly with privacy guarantees) from a larger model. A standard way to approach model distillation is to mix original data and the same data translated by the large model, resulting in two translations per one source sentence. In [Zou21] we showed that it is possible to improve this method by supersampling sentences in the new training set which are estimated to be of higher quality. Overall, based on the same input data, this results in models which have higher translation quality (measured automatically).

2.5.6 Constrained Machine Translation

Lexically constrained translation allows modification of generated translation by providing target words and phrases that are known beforehand, e.g. specific terminology for given domain. [Jon+21c] explores possibilities of implementing constrained translation in morphologically rich languages, where constraints need to be inflected into a correct surface form to match the rest of the translation. Previous approaches were able to enforce terms to appear in the translation, however they often struggled to make the constraint word form agree with the rest of the generated output. We investigate mechanisms to allow neural machine translation to infer the correct word inflection given lemmatized constraints. In particular, we focus on methods based on training the model with constraints provided as part of the input sequence. Our experiments on the English-Czech language pair show that this approach improves the translation of constrained

terms in both automatic and manual evaluation by reducing errors in agreement. Our approach thus eliminates inflection errors, without introducing new errors or decreasing the overall quality of the translation.

[Jon+21b] implements the approach described above for different language pair (English to French) in Covid-19 related domain. We submitted the final system to WMT21 Machine Translation using Terminologies Shared Task.

2.5.7 Exploration of Pretrained LM Reuse in MT via Elastic Weight Consolidation

[VB19] explores the effect of regularization using Elastic Weight Consolidation (EWC) on the unsupervised NMT pretraining using monolingual corpora. Deep neural networks often suffer from "catastrophic forgetting", a phenomenon occurring during incremental learning where the network weights optimized for a prior task A are being overwritten during optimization for the following task B, disregarding the weight values optimized for the prior task.

The main idea behind EWC is to identify weights that are crucial for the prior task and put a heavy regularization penalty on modifying these particular weights during optimization on the subsequent tasks. This should result in the model avoiding modification of the important weights and using its unused capacity to learn to solve subsequent tasks.

In our experiments, we focus on separate pretraining of the encoder and decoder using larger monolingual corpora and making use of this prior knowledge in the encoder/decoder during fine-tuning on the low-resource parallel data for MT. The results show that the monolingual pretraining with bilingual fine-tuning in combination with EWC is on par with the previous pretraining method only when decoder is pretrained (and encoder is randomly initialized) prior to fine-tuning. However, the pretraining methods fall short when compared to widely popular backtranslation approaches.

We hypothesize that the difference between the monolingual and bilingual tasks is too large to benefit from the EWC regularization approach. In the future, we plan to explore the EWC regularization (or similar techniques) in the multi-lingual NMT scenarios (using bilingual data only).

2.5.8 Length Generalization Issue in Machine Translation

We show in [VB21] that decoders within the Transformer architecture have a strong tendency to overfit to the length of the output sequences seen during training. We demonstrate on both a simple string editing task and the machine translation (MT) task that Transformer decoder has a clear tendency of producing sentences of the length similar to the lengths of the referential sequences in the training data regardless of the input sequence length.

Additionally, we show that this tendency to overfit to the lengths of the training sequences can be exploited when creating artificial training examples of longer length of reference by concatenating available shorter training examples without the necessity of the concatenated examples being related to each other. This can in turn help with training system where there is lack of sufficient amount of longer training examples. Furthermore, the creation of additional training data can be beneficial in the low-resource translation settings.

In our future work, we plan to investigate in more detail, how much is this phenomenon influenced by the similarity between the training and validation data and how big of a role the intrinsic positional encoding influences this behaviour.

2.5.9 Robustness of Machine Translation to Noisy Inputs

In [HLP19], we presented a submission to the WMT 2019 Machine Translation Robustness Shared Task. The goal of the task was to develop MT systems that can handle noisy data on the input. In our submission, we used the CUNI Transformer system [Pop18] as our main model which was already quite robust against the input noise compared to the official shared task baseline. We show that using a strong baseline is the key for creating a robust MT model.

In [KAB20b], we present our dataset [KAB20a] of 27k tweets in Persian equipped with linguistic annotation (POS tagging, dependency parsing, named entity recognition, sentiment analysis and machine-translated versions of each tweet in five target languages. This dataset will serve as the basis for future experiments on robustness, multilinguality and – via the sentiment – possibly also semantics.

2.5.10 Multilinguality for Better MT of Low-Resource Languages

[Jon+21a] is focused on tackling the issue of low-resource MT using multilingual models. The paper describes our submission to WMT21 Multilingual Low-Resource Translation for Indo-European Languages Shared Task, where we competed in translation from Catalan into Romanian, Italian and Occitan. We show that using joint model for multiple similar language pairs improves upon translation quality in each pair. We also demonstrate that chararacter-level bilingual models are competitive for very similar language pairs (Catalan-Occitan) but less so for more distant pairs. We also describe our experiments with multi-task learning, where aside from a textual translation, the models are also trained to perform grapheme-to-phoneme conversion.

2.5.11 Unsupervised Machine Translation and Sentence Embeddings

When training a machine translation system in low-resource conditions, it is vital to be able take advantage of monolingual data, which are significantly easier to collect. We investigated various methods of training unsupervised machine translation systems without using any parallel resources for training. We competed in the unsupervised task of WMT20 [KKB20] with our neural MT system relying on a pretrained bilingual encoder and online back-translation and ranked second among four teams.

A lack of parallel text resources is a problem not only when translating between low-resource languages pairs but also when translating in narrow domains. In [KB21] we showed how transfer learning and multilingual training can improve translation quality in the specialized domain of Covid-19.

We also studied how pretrained representations can aid machine translation in low-resource scenarios. We devised a method to find parallel sentences in comparable corpora using unsupervised sentence embeddings (see [Kva+20]) and create new parallel data sets.

2.5.12 Explainable MT Quality Estimation

We participated in the Eval4NLP shared task with Explainable Quality Estimation: CUNI Eval4NLP Submission [PSB21]. The goal of the shared task was to explore solutions for reference-free evaluation of the machine translation (MT), i.e., the inputs of the evaluation are source sentence and MT hypothesis. We used a pre-trained Encoder-Decoder Transformer model XLM-R [Con+19] and we fine-tuned this model for reference-free MT evaluation. The advantage of using a pre-trained multilingual model is that it is able to perform an evaluation on unseen language pairs too. In this case, the training language pairs are Estonian-English and Romanian-English. The test set includes unseen language pairs German-Chines and Russian-German. We demonstrate that our system is capable to perform the reference-free MT evaluation on these pairs, too. In the supervised track, our system achieves almost state-of-the-art results and places 2nd and 3rd on the supervised Ro-En and Et-En test sets.

2.6 Speech machine translation

2.6.1 Non-Native Simultaneous ASR and ST in IWSLT 2020

In this section, we describe our investigations on how to handle gradually growing input for MT, a setting critical for simultaneous speech translation. We took part in the non-native speech translation task at IWSLT 2020 (see also Section 5.1.1) with two sets of systems.

In [Mac+20], we describe the first system submission. We describe systems for offline ASR, real-time ASR, and our cascaded approach to offline SLT and real-time SLT. We select our primary candidates from a pool of pre-existing systems, develop a new end-to-end general ASR system, and a hybrid ASR trained on non-native speech. The provided small validation set prevents us from carrying out a complex validation, but we submit all the unselected candidates for contrastive evaluation on the test set.

[Pol+20] is our second submission for the IWSLT 2020 shared task. In this submission, we use a cascaded speech-to-text translation system. We propose to use phonemes as an intermediate representation of the source language instead of usual graphemes. We demonstrate that the proposed solution outperforms commercially available ASR and translation systems (Google and Microsoft).

2.6.2 Multi-Source Simultaneous Speech Translation

The neural machine translation (NMT) has the capability to handle more source or target languages at once [Joh+17; DCK20].

There are events with multi-lingual audience that use simultaneous interpreting by human experts. However, the set of interpreting target languages is often smaller than what would the audience need. An existing technology of automatic simultaneous speech translation [Mül+16; Nie+18] may increase the set of target languages. Unfortunately, the simultaneous speech translation from one source is often imperfect, due to speaker variance, background noise, non-native accent, etc. A machine could be processing multiple parallel speech signals as source for translation, e.g. the main speaker, and the interpreters. We hypothesize that multi-sourcing may bring benefit to translation quality.

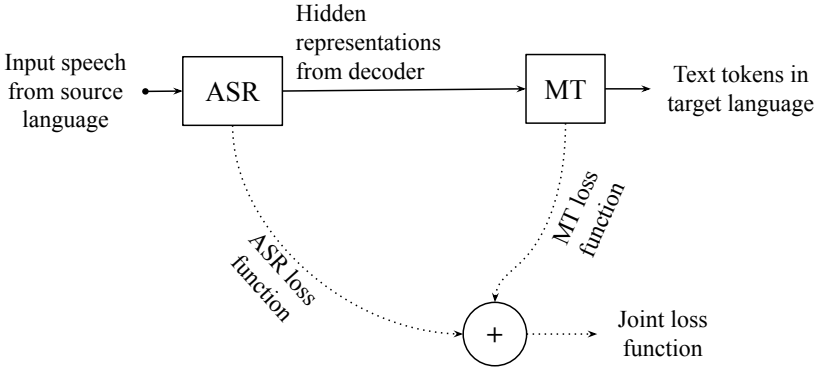


Figure 7: Block diagram of ASR-MT model with joint (multi-task) loss.

As data preparation, we created an evaluation corpus of interpreting from European Parliament ESIC [MŽB21].Our initial experiment with this dataset, described in the same paper concludes that automatic evaluation cannot be used to find out if it is better to translate from the original speaker or from the interpreter. We show that translating from the interpreter introduces a delay comparable to relay interpreting but has the advantage of producing shorter outputs with significantly less complex vocabulary. A limited human assessment shows that more information is preserved in direct translation than in interpreting-based translations, and that far more content survives in translation from gold transcripts than from online ASR.

2.6.3 End-to-end Spoken language translation

A typical spoken language translation (or speech machine translation) is a cascade of ASR and MT systems. The input speech from source language is automatically transcribed into n-best hypotheses (in source language), which are then passed to an MT that produces the text in the target language. The errors from the ASR system are propagated into the MT system and cannot be avoided in such a cascaded setup, thus producing sub-optimal results. We proposed a methodology [Vyd+21b] that aims to overcome the aforementioned problem, where we jointly optimize both the ASR and MT systems. This so-called end-to-end approach is achieved by creating a differentiable path between the final objective function and the model parameters of both the ASR and MT systems.

Both the ASR and MT models are based on Transformer encoder-decoder neural network architecture. The differentiable path is established by passing the final hidden representations from the decoder of ASR to the encoder of MT module. This scheme is illustrated in Fig 7. The experiments on HOW2 speech translation dataset (English speech $\rightarrow$ Portuguese text) have shown improvements on BLEU scores from 43.6 to 46.9. The achieved results are on par with the best systems from IWSLT 2019, while having $5\times$ less model parameters.

2.7 Multimodal approaches

2.7.1 Eye-Tracking of Multi-Modal Human Translation

While traditional translation is thought of as being about printed words, the rise of the Internet and new media have made multimodal texts ubiquitous. This implies that there is a need for a systematic understanding of translation processes for multimodal texts. This understanding, we posit, should arise from learning from how both human and machine "brains" function. A comparative analysis of human and machine intelligence should lead to a better understanding of intelligence. It is thus fair to assume that cognitive signal data from a participant performing a set of tasks involving language use would be a great resource to understand how the brain works with language. And if the tasks are designed carefully such that they can also be presented to trained deep neural network algorithms, they can be used to compare the human and machine perspective of handling the same problem. A similar approach in computer vision has led to simpler and yet equally effective models for object detection.

In our eye-tracking study (conducted in early 2021), we collected cognitive data in the form of gaze, audio and EEG signals from participants while they translated English sentences in multimodal conditions into Czech. The source sentences were comprised of ambiguous and unambiguous sentences, accompanied by an image. The images were either related the sentences or unrelated and were used in the anticipation that it would demonstrate how the participants utilized the information from the images during translation i.e. if the images were helpful at all. We anticipate that this Eyetracked Multi-Modal Translation (EMMT) dataset [Bha+22] would help us study how behavioral patterns (in terms of cognitive signals) change with stimuli of differing modalities. The tasks that the participants were subjected to have been designed so that similar experiments can be performed on the state-of-the-art algorithms for natural language processing. A paper containing a description of the data collected was submitted to Nature Scientific Data and is currently under review.

Additionally, we did a preliminary analysis of the collected gaze data. We investigated the specific aspects of translating ambiguous and unambiguous sentences, and simultaneously, we focused on the possible impact of visual information on the translation process. We recorded how the Stroop effect was visible in the experimental setup. We also found that in all the scenarios, the participants scanned through the entire screen trying to extract as much information before selectively attending to specific parts of the visual area containing the relevant stimulus. We also found that ambiguity in sentences is detected naturally, and the increased processing effort is reflected in the eye-tracking data. Finally, we saw that the image stimuli helped the participants but only in ambiguous conditions. A manuscript detailing the experiments has been almost prepared and will be soon submitted to the Journal of Neurolinguistics.

2.7.2 Machine Translation Supported by Images and Image Captioning

Early at the beginning of NEUREM3, we finished our dataset for multimodal experiments: Hindi Visual Genome [PBD19] (HVG).[5]

[5] https://ufal.mff.cuni.cz/hindi-visual-genome

The dataset originally comes from the Visual Genome[6] (VG) collection of images and English captions of selected areas in the image. VG provides additional types of annotation, such as a graph of objects in the image and their relations, which is beneficial for future research.

Our HVG consists of 29k training items and three separate test sets: a development test and a test set sample randomly in the same way as the test, and a challenge set which was selected based on the presence of one of 19 ambiguous words where the image may help to choose the correct translation from English into Hindi (e.g. "penalty" for a fine vs. the football action). Later analyses have documented that in many cases, a human reader can correctly pick the meaning already from the text. For machine translation, even these cases may remain a challenge.

The dataset was then used in three instances of a shared task in multimodal translation which we organized [Nak+19; Nak+20; Nak+21]. We also contributed our systems to these tasks [PMB19a; Par+20; Par+21]

2.7.3 Synthesizing Training Data for Handwritten Music Recognition

In [MP21] we focused on the visual modality of interpreted data, namely in the area of Handwritten Music Recognition (HMR). This task is analogous to Optical Character Recogniton (OCR) which aims at automatic recognition of images (scans) of printed, typed or handwritten text into machine-encoded text. In HMR, the goal is to recognize handwritten music scores. This task which is much more challenging not only due to the complexity of music scores but also due to very limited availability of training data. This work is based on the series of our previous publications in the are of HMR and focuses on exploitation of synthetic data which is a known technique to tackle the unavailability of sufficient amounts of training data in machine learning. in the area of HMR, this approach is very innovative and has not been explored before. In [MP21], we proposed, implemented and thoroughly evaluated a method to generate images of handwritten music scores that can be directly used for training a HMR model. Our HMR model based on a Convolutional Recurrent Neural Network (CRNN) with the Connectionist Temporal Classification (CTC) loss function achieved state-of-the-art results. It was evaluated on unseen real-world sheet music samples and achieved Symbol Error Rate (SER) of 25%. In manual error analysis, it outperformed previously published State of the Art. This work was presented at the ICDAR conference (CORE RANK A) and was conducted with a master student at MFF UK Jiří Mayer who is going to join the MFF UK team as a PhD student next year.

2.8 Towards Semantics

2.8.1 Representations of Sentence Meaning

Deep learning has brought about new possibilities for representing sentences. While continuous representations of words (word embeddings) have been extensively studied, this was not the case for continuous representations of whole sentences in 2019. We thus organized a special issue of JNLE and in its overview, we briefly summarized the

[6] http://visualgenome.org/

desirable properties of sentence meaning representation and compared if these are met by traditional symbolic methods and, separately, by continuous representations [BBW19].

2.8.2 Translation into Many Paraphrases

The meaning of a sentence in a natural language is an abstract entity, not strictly tied to the particular wording of a sentence. One practical approach towards capturing the abstract meaning is to enumerate paraphrases of a sentence, each of which and all of them together represent the meaning. Being able to generate paraphrases and to identify if a candidate string is or is not a good paraphrase of a sentence is thus a stepping stone towards modelling the meaning of the sentence.

To better understand modeling paraphrases, we took part in the Simultaneous Translation And Paraphrase for Language Education challenge. In our submission [Lib+20], we developed a system that is supposed to produce not only a single best translation hypotheses, but also generate as many correct translations as possible. Our best solution is based on filtering beam search output using a classifier based on multilingual representation. Further experiments with automatic output paraphrasing did not lead to further improvements.

2.8.3 Compositionality in Sequence-to-Sequence Models

Compositionality is one of the key aspects of units of meaning in natural languages.

We investigated to what degree sequence-to-sequence neural network models (simpler variants but very closely related to those used in MT) can combine learned pieces of information in novel contexts. This line of research is usually motivated by the differences between human and NN learners. Human knowledge is compositional, which makes it possible to build complex hierarchical representations by combining simpler units. This in turn leads to ability to generalize well even if the underlying data generating distribution changes.

We used the SCAN benchmark designed to test for compositionality in the domain of NLP. In [AP21] we reported a method of solving the tasks by combining previously published architectural changes and data augmentation.

Recently, there have been doubts whether the attempts of making NNs exhibit some compositional behavior are worthwhile. After all, NNs have been shown to generalize well. We argue (and are in the process of showing experimentally) that the key factor in deciding whether compositionality might be beneficial is the change (or lack thereof) in the data generating distribution.

2.8.4 COSTRA: Corpus of Complex Sentence Transformations

In our study of continuous sentence representations, i.e. sentence embeddings, we also approach the topic from the other side, from a set of concrete sentence examples that exhibit certain properties or that are in certain mutual relations.

In [BB20a], we presented COSTRA 1.0, a dataset of complex sentence transformations. The dataset is intended for the study of sentence-level embeddings beyond simple word alternations or standard paraphrasing. This first version of the dataset was limited to sentences in Czech but the construction method is universal. The dataset consists of

4,262 unique sentences with an average length of 10 words, illustrating 15 types of modifications, such as simplification, generalization, or formal and informal language variation. The hope is that with this dataset, we should be able to test semantic properties of sentence embeddings and perhaps even to find some topologically interesting "skeleton" in the sentence embedding space. A preliminary analysis using LASER, multi-purpose multi-lingual sentence embeddings suggests that the LASER space does not exhibit the desired properties. Furthermore, in [BB20b] we presented a new dataset for testing geometric properties of sentence embeddings spaces. In particular, we concentrate on examining how well sentence embeddings capture complex phenomena such paraphrases, tense or generalization. The dataset is a direct expansion of Costra 1.0 which we extended with more sentences and sentence comparisons. We show that available off-the-shelf embeddings do not possess essential attributes such as having synonymous sentences embedded closer to each other than sentences with a significantly different meaning. On the other hand, some embeddings appear to capture the linear order of sentence aspects such as style (formality and simplicity of the language) or time (past to future).

2.8.5 Cross-lingual Information Retrieval

One of the variety of multilingual NLP tasks is cross-lingual information retrieval (CLIR), a special case of information retrieval (IR). The task of IR is to retrieve documents relevant to an information need from a collection of those documents. The information need is usually specified as a textual query. In CLIR, the language of the query and the language of documents are different.

In [SP20], we published a thorough comparison of two principal approaches to CLIR: document translation (DT) and query translation (QT). The experiments were conducted using the CLEF eHealth cross-lingual test collection containing documents in English and queries in several European languages. Our results with statistical (SMT) and neural (NMT) models showed that the quality of QT by SMT is sufficient enough to outperform the retrieval results of the DT approach for all the languages with comparable results to monolingual retrieval. NMT then further improved results of both the QT and DT methods for most languages. Generally, QT provided better retrieval results than DT.

In [SP19], we presented a method for automatic query expansion improving CLIR performance. The method employed MT of source-language queries into a document language and prediction of the retrieval performance for each translated query when expanded with a candidate term. Candidate terms (in the document language) were taken from multiple various sources. The experiments were conducted using the CLEF eHealth 2013–2015 test collection and showed significant improvements in both cross-lingual and monolingual settings.

2.8.6 Towards Meeting Summarization

In [NB19], we defined the task of "minuting", i.e. summarizing meetings into the form of "minutes". This task is different from standard text summarization in that it crosses genres: the source is spoken domain while the target is written domain. Also, the input will very likely be noisy in terms of the structure of sentences (the concept of sentences generally does not apply well to spontaneous speech) as well as longer blocks. Our minuting

is closest to the task called "dialogue summarization", but we explicitly expect a multi-party meetings and we also tend to prefer structured minutes, i.e. bulleted lists rather than prose, although we acknowledge that there are many different styles of taking minutes.

To refine the generally imprecise definition of minuting, we built a corpus of meetings and equipped it with manual minutes. For multiple meetings, we have at least two different and independently created minutes.

We ran an initial shared task for this challenging goal, see Section 5.5.5, and we plan to carefully analyze the dataset and start building our summarization systems soon as discussed below of page 61.

2.8.7 Multimodal Summarization

We study abstractive summarization for open-domain videos [Pal+19]. Unlike the traditional text news summarization, the goal is less to "compress" text information but rather to provide a fluent textual summary of information that has been collected and fused from different source modalities, in our case video and audio transcripts (or text).

We experiment with multi-source sequence-to-sequence models with hierarchical multimodal attention (previously developed at CUNI) and that it can integrate information from different modalities into a coherent output.

We also propose a new evaluation metric (Content F1) for abstractive summarization task that measures semantic adequacy rather than fluency of the summaries, which is covered by metrics like ROUGE and BLEU.

2.8.8 Information Retrieval from Scientific Papers

As a separate stream of research, we have been involved in methods for condensing information from scientific papers.

We created large datasets for the task [ÇB20b] and experimented with very basic properties such as predicting the paper length from the paper's metadata [ÇB20a] or with generating keywords and key phrases from the title and abstract of the paper [ÇB19b].

Interestingly, we find out that modern deep learning methods are not any better than standard techniques on this task.

2.9 Human performance and human interfaces

This research area explicitly considers humans in text and speech processing tasks.

On one hand, we focus on humans as users of translation systems: (1) we design and empirically evaluate a user interface that supports users when producing text in a language they do not speak, and (2) we build upon humans as the ultimate judges of translation quality.

On the other hand, we analyze human performance when carrying out translation, motivated by the long-term vision of being able to simulate human processing of language closer and thus understanding it better and ideally also obtaining more natural outputs.

2.9.1 Machine Translation User Interface

Translating text into a language unknown to the text's author, dubbed outbound translation, is a modern need for which the user experience has significant room for improvement, beyond the basic machine translation facility. An initial experiment on Czech-German [ZB20] and a follow-up larger experiment on English-Estonian and English-Czech [Zou+21a] in collaboration with other institutions was conducted to find the effect of various outbound translation modules (backward translation, quality estimation with alignment and source paraphrasing, shown in Figure 8).

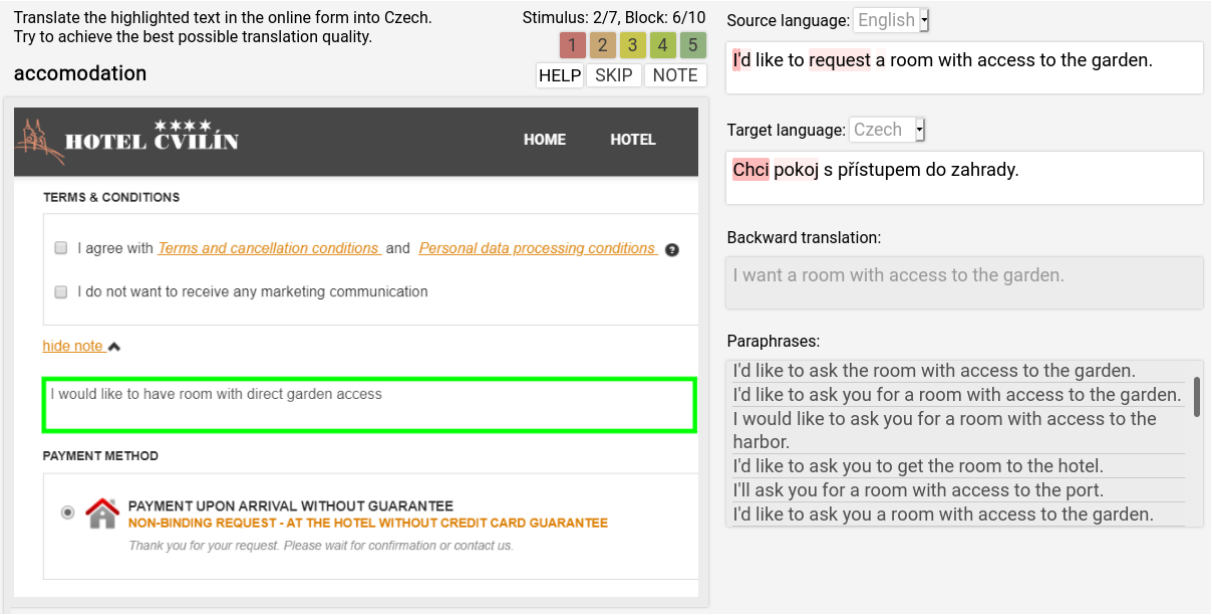


Figure 8: Example of an experimental user interface [Zou+21a] with all three modules enabled.

Through an analysis of collected user inputs and produced translations we conclude, that while word-level quality estimation appears promising at first, it is not yet production-ready to have any significant effect on the objective translation quality. In fact, because the deployed QE system was very critical, it only worsened the self-reported user confidence. Paraphrases had no positive or negative effects. Backward translation, a very common strategy among Internet users, turns out to strongly positively affect the self-reported user confidence, but not the objective translation quality.

2.9.2 Human Evaluation of Simultaneous Speech Translation

The goal was to examine methods for presenting the outputs of simultaneous speech translation systems to users. We examined possibilities of relaxing constraints such as reading speed and space limitations, achieving an improvement in overall quality of translated speech presentation. We provided a thorough analysis of selecting the best layout from a number of presentation options, recommending its usage given available resources. Additionally, the results brought new insights to the relation between the evaluation methods, suggesting less time-consuming evaluation for specific use cases in

future experiments. The paper has not been accepted at the ACL conference yet, but the work is captured in David Javorsky's master thesis.[7]

2.9.3 Machine, Human, and Superhuman Translations

Both human and machine translation quality assessment is one of the core research fields in translation studies and MT research, investigating the correlation of neural representations with human linguistic categories. Recently, various post-editing procedures and other human-machine interactions in this area have been thematized. In [KBP21] we provided a quick overview of possible methods how to detect that reference translations were actually created by post-editing an MT system. Two methods based on automatic metrics are presented: BLEU difference between the suspected MT and some other good MT and BLEU difference using additional references. These two methods revealed a suspicion that the WMT 2020 Czech reference is based on MT. The suspicion was confirmed in a manual analysis by finding concrete proofs of the post-editing procedure in particular sentences. Finally, a typology of post-editing changes is presented where typical errors or changes made by the post-editor or errors adopted from the MT are classified.

Furthermore, there is a large area of superhuman translations to be explored. In our article (submitted 7/2021, Slovo a slovesnost, in review process) *Exploring translationese: professional vs. superhuman translation of newspaper articles* we explored translation quality, particularly style including the presence of source-language interference, by analysing the quality of Czech translations of two English newspaper articles, done by three translators hired by a translation company; the other area explored by the study is the creation of superhuman translations, i.e. translations thought to be the best possible under the given conditions. The primary focus of the quality assessment is on the style of the existing translations, and the problems identified fall into a wide range of linguistic categories such as spelling, morphosyntax, grammar, lexicon and word formation. Special emphasis is placed on interference, and the draft typology compiled by the authors can be expanded to include several other types of interference recurrent in the translations analysed. The superhuman translations, resulting from the teamwork of two translators-cum-theoreticians who synthesised the best of the three existing translations while employing a considerable amount of their own creativity, may be used in future assessments of excellent machine translations.

[7] https://dspace.cuni.cz/bitstream/handle/20.500.11956/147964/120397331.pdf

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. Please see Section 4 to see the details on their track record, activity on NEUREM3, etc.

- Hari Krishna Vydana, senior, India
- Santosh Kesiraju, senior, India
- Anna Silnova, PhD student, 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 the first three years of NEUREM3:

- Dr. Alicia Lozano[8] (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].
- 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].
- 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

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

published within NEUREM3 [PBD19], our first dataset for multi-modal translation. During NEUREM3, our collaboration continues, although Shantipriya is now funded from his Swiss sources. 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].

3.3 Organization and participation in international workshops, challenges and evaluations

This important international activity is covered in Section 5.

3.4 Summer internships

Summer 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.

- 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.
- From June till October 2011, 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. ATM employs an external ASR model or scorer to weight unsupervised input samples The results showcase the importance of data selection under mismatched pretraining and finetuning data conditions.
- 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 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]).

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 moire application-oriented scenarios:

- IARPA MATERIAL[9] (coordinator Raytheon BBN Technologies Corp., main researcher at BUT Dr. Martin Karafiát) was completed in 09/2021. The project

[9] https://www.iarpa.gov/research-programs/material

aimed to develop a system for obtaining information of the type "English-in, English-out". Based on a domain-dependent English query, the MATERIAL system searches for relevant data in a large multilingual repository and presents it as a summary. BUT focused on the low-resource ASR systems in the project.

- Horizon 2020 ROXANNE[10] (Real time network, text, and speaker analytics for combating organized crime, IDIAP coordinator) is a large European project dealing with speech processing and network analysis techniques for detecting criminal networks. NN representations of speakers are key component in the project.

- Horizon 2020 ATCO^2;[11] (Automatic collection and processing of voice data from air-traffic communications, coordinated by IDIAP) and HAWAAII[12] (Highly automated air traffic controller workstations with artificial intelligence integration, coordinated by DLR) are aiming at the data and systems for transcribing the communication of pilot and air-traffic controllers (ATC). Although seemingly application projects, they bring a number of research issues ranging from low-resource ASR training, through robust voice activity detection to recognition of language from very short speech segments.

- H2020 ELITR[13] (European Live Translator) project is coordinated by Ondřej Bojar at CUNI. It aims to overcome the language barrier in communication within and among European citizens, companies, institutes and organizations at large assemblies like conferences, in smaller live discussions like workshops and in discussions held over long distance like formal or informal online meetings.

- H2020 WELCOME[14] (Multiple Intelligent Conversation Agent Sevices for Reception, Management and Integration of Third Country Nationals, coordinated by UPF Barcelona) aims to research and develop intelligent technologies for support of the reception and integration of migrants and refugees in Europe. Both BUT and CUNI participate in this project concentrating on speech and language technologies respectively.

- CELSA CELL[15] (ContExtual machine Learning of Language translations) A joint project with the Charles University and KU Leuven focusing on the research of neural representations in the multimodal context.

- ROZKAZ[16] (Robust processing of recordings for operations and security) is supported by Czech Ministry of Interior (MoI) and targets information mining from real recordings in the field of security and close cooperation with police and national security organizations. Martin Karafiát (previously on NEUREM3 team) is the PI of ROZKAZ since October 2020.

[10] https://roxanne-euproject.org/
[11] https://www.atco2.org/
[12] https://www.haawaii.de/wp/
[13] https://elitr.eu/the-project/
[14] https://welcome-h2020.eu/
[15] https://researchportal.be/en/project/cell-contextual-machine-learning-language-translations
[16] https://www.fit.vut.cz/research/project/1428/.en

3.6 Networking and International research infrastructure projects

- H2020 ICT-48 Humane AI Net[17] (European network of human-centered artificial intelligence, coordinated by DFKI) is a large networking project built around ethics values and trust (Responsible AI). BUT and CUNI are both involved in this project, including participation in its μ-projects.
- H2020-MSCA-RISE ESPERANTO[18] (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.
- i-AIDA[19] (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[20] (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.

[17] https://www.humane-ai.eu/
[18] http://esperanto.univ-lemans.fr/en/index.html
[19] https://www.i-aida.org/
[20] https://www.clarin.eu/

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 ot 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. 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 (see Section 3.5) [Zul+20].
- 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 still intensively 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+21b].

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 unfortunately left the team in July 2021 for Huawei, Finland.

- 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 relatively recently (August 2021), therefore none of his recent publications bears NEUREM3 acknowledgement. He has taken the speech translation work from Hari K. Vydana and supervises Marek Sarvas (an MSc student, see below).
- Pavel Pecina, Ph.D. is an associate professor working in the area of machine translation , cross-lingual information retrieval, and multimodal-data processing. He's been supervising Michal Auersperger, Shadi Saleh, and Petra Galuščáková.
- Petra Galuščáková, Ph.D. was planned to join the project from year 2 and work primarily on multimodal information retrieval but due to an extension of her postdoc stay in the U.S. and later due to Covid-19 pandemic, she has not relocated back to Prague yet.
- Shadi Saleh, Ph.D. participated in two strong publications in the area of cross-lingual information retrieval [SP19; SP20]. He defended his dissertation in 2020 and left the university for a research position in industry.
- Jindřich Libovický, Ph.D. was a postdoc at Center for Speech and Language Processing at the Ludwig Maximilian University of Munich and is rejoining NEUREM3 in 2022. He worked mostly on character-level methods for machine translation and NLP in general and analysing and improving language neutrality of multilingual sentence representations.
- Jindřich Helcl is a research associate at the University of Edinburgh until June 2022, working on non-autoregressive neural machine translation, which is the topic of his doctoral thesis. During his stay in Edinburgh, he was working on the Gourmet project, which aims at improving machine translation of news articles under very low-resource conditions.
- 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. She contributed to the research of superhuman references and translation evaluation [KBP21].

4.3 PhD and MSc students

BUT:

- Karel Beneš is MSc. graduate of BUT. He is working in the area of language model adaptation [BB21]. Karel is expected to graduate in 2023-24.
- Murali Karthick Baskar is MSc. graduate of IIT Madras, India. He is working in end-to-end ASR and its training on disjoint speech and text resources using an ASR-TTS loop [Bas+21]. He is also involved in (so far unpublished) ASR of dysarthria patients. Karthick is expected to graduate by end 2022.

- Anna Silnova. She 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 evaluation efforts [Ala+19; Mat+19; Zei+19; Ala+20; Bur+20; Loz+20] as well as summary publication [Mat+20a]. She also participated in the research of embeddings for speaker recognition [Roh+19] and diarization [Lan+21a; Lan+21b]. Her main topic are probabilistic embeddings in speaker recognition and diarization [Sil+20]. She is expected to graduate in 2022.
- Jan Brukner is MSc. graduate of BUT. He started his PhD in 2020 and works on synthesis and voice modification with neural architectures.
- 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. Although only in the 1st year, he is already the authors of several publications, of which [Pen+21a; Pen+21b] are dedicated to NEUREM3. He works in the area of neural signal processing aiming at speaker recognition. He started on the project in September 2021.
- 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]
- Marek Sarvaš is an MSc. student under the supervision of Santosh Kesiraju. He worked on Bc. thesis "Interpretability of Neural Networks in Speech Processing" and currently concentrates on speech machine translation. He started on the project in August 2021.

CUNI:

- Michal Auersperger is a PhD student at CUNI. His research interests are representation learning, specifically at the concepts of disentanglement and compositionality. His recent work focused on solving the SCAN tasks [AP21].
- Sunit Bhattacharya is a M.Sc graduate of Central University of Rajasthan, India. He started his Ph.D in 2019 and works on representations for multimodal learning.
- Dominik Macháček is PhD candidate at CUNI. He specializes in multi-lingual machine and speech translation.
- Ivana Kvapilíková is a PhD student at CUNI. She specializes in low-resource machine translation and unsupervised training of machine translation systems. She is currently on her maternity leave.
- Peter Polák is a PhD researcher at CUNI. He specializes in speech recognition and translation. He is supported by START Programme[21] of Charles University.

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

4.4 Support staff

At BUT, Renata Kohlová is specifically supporting the contractual, reporting and management work on NEUREM3.

At CUNI, project administration is carried out by Libuše Kaprálová with the help of Tereza Vojtěchová who is responsible for managing annotators and student workers on short-term contracts. Recently, 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 Position of the team and international excellence

5.1 International Evaluations and Challenges

Open evaluations and challenges (also dubbed benchmark campaigns, 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[22]) 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.),
- 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.

5.1.1 Methods of Text and Speech Translation Evaluation

Shared Task Organization in Machine Translation Shared tasks serve as a key stabilizing component in research. The ability to evaluate world's best systems in common settings and with fixed data and evaluation methods is a priceless guide for research: the results of the shared task indicate, which approaches are the most promising and the whole community mixes and builds upon them for the next round of the task. Aware of the utility of shared tasks, we are very active in this area.

For machine translation, we are annually taking part in the organization of WMT [Bar+19; Bar+20; Akh+21]. This work on our part involves the preparation of the test materials (including reference translations), discussion and updates on the evaluation criteria (specifically, we need to respond to the improving quality of the systems and in the last few years, we have been gradually moving towards document-level style of evaluation), and then the analysis of the results.

Most years, we try to participate in the WMT news test shared task, too. In 2021, we contributed with minor improvements of the system described in Section 2.5.1 [Geb+21].

Shared Task Organization in Machine Translation Evaluation Manual evaluation (in its many possible forms) is the ultimately best way of evaluating machine translation. However, it is costly and difficult to reproduce, so it is suitable rather for annual shared tasks or other larger evaluation campaigns. For day-to-day development of systems, automatic evaluation methods are needed – but they need to correlate well with the human judgments.

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

In the WMT Metrics shared task [Ma+19; Mat+20b; Fre+21], we provide the opportunity to benchmark automatic metrics of machine translation quality. Our contribution to the evaluation comprises mainly the contribution to the design of evaluation criteria which again have to keep evolving as the metrics themselves are getting better.

Shared Task Organization in Speech Translation In the area of speech translation, we focus on two aspects: non-native speech and simultaneous speech. Again, we see the organization of shared tasks as a very valuable method for ensuring progress of the field.

We have joined the organizers of IWSLT. In 2020, we provided our own task on translation of non-native speech, building upon our test sets, incl. Antrecorp described in Section 2.3.1 [Ans+20].

In 2021, we contributed to the task on simultaneous translation [Ana+21], and we are expanding this for the upcoming year 2022, by providing human benchmark in speech translation (i.e. interpreters) as well as human evaluation (see Section 2.9.2).

A Toolkit for Evaluation of ASR, MT and Speech Translation In [Ans+21], we present our tool, SLTev, which is designed to serve as the standard for evaluating ASR, MT and speech translation in a comprehensive way. The three main scales, along which the output is evaluated, are: (1) translation or recognition quality, approximated by fixed versions of common metrics BLEU and WER, (2) latency, in the form of delay behind an ideal speech recognition or translation system, and (3) flicker, a measure useful for systems that gradually refine their outputs.

We have participated in the IWSLT 2021 workshop in the "offline speech tranlsation" track, where we submitted a system to translate from English speech to German text [Vyd+21a]. Our system achieved BLEU score of 33.79 which is competitive other to the best systems.

Pre-trained machine translation metric quality Automated MT metrics, which evaluate the quality of MT system output without costly human annotation, are key to developing high-quality neural MT systems. Recently, new metrics relying on large pre-trained language models (LLMs) have achieved high correlation with human MT evaluations. However, much like the LLMs on which they rely, these metrics are opaque and hard to interpret; thus, it is difficult to know in which situations these metrics will perform well or fail. Fortunately, many studies explore LLMs' strengths and weaknesses, although not in the context of machine translation.

In [HB21], we draw on this body of research to answer the aforementioned question: in what situations will LLM-based MT metrics succeed and fail? We examine one such metric, BERTScore, and find that it performs well not only in the presence of various linguistic phenomena, but also when candidate translations are poor-quality. However, much like its string-based predecessors, it still yields overly-high scores to incorrect candidate translations that have high lexical overlap with a reference translation.

Additionally, in [HM21], we analyze BERT, on which BERTScore relies, 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.

Neural Machine Translation Quality Even though sentence-centric metrics are used widely in machine translation evaluation, document-level performance is at least equally important for professional usage. In [ZVB20], we brought attention to detailed document-level evaluation focused on *markables* (expressions bearing most of the document meaning) and the negative impact of various markable error phenomena on the translation. For an annotation experiment of two phases, we chose Czech and English documents translated by systems submitted to WMT20 News Translation Task. These documents are from the News, Audit and Lease domains. We showed that the quality and also the kind of errors varies significantly among the domains. This systematic variance is in contrast to the automatic evaluation results.

5.1.2 Speaker diarization

VoxConverse challenge The VoxConverse speaker diarization challenge was part of the VoxCeleb speaker recognition challenge[23], which was organized by a group of researchers from the University of Oxford, Google research, Amazon research and Naver Corp.

Our system [Lan+21a] ranked 2nd in the challenge[24]. VBx diarization (see 2.2.7 for details) was the core of the diarization system developed by our group, but several other key components of the system were analyzed and improved for the challenge.

- A novel voice activity detection (VAD) module was built as a combination of three different VAD systems based on energy, DNNs and ASR, respectively.
- A global speaker embedding reclustering was employed. The refinement module made use of the large amount of data available per output speaker cluster (as compared to the short input segments used for the initial x-vector embedding estimation), to estimate new speaker embeddings. This new embeddings, potentially more robust, were used in the re-clustering stage to merge (*same-identity*) speaker clusters.
- An overlapped speech detection and labeling strategy to deal with cross talk.

DIHARD III speaker diarization challenge: The third edition of DIHARD[25], organized by researchers from various institutions leaded by the Linguistic Data Consortium, was a follow-up of previous DIHARD challenges, which focus on diarization on hard and heterogeneous data. As a novelty, the third DIHARD challenge included for the first time telephone data among its various data domains.

The system developed by us had again VBx diarization as the main system. Besides, we developed and included for the first time an end-to-end diarization system, which

[23] https://www.robots.ox.ac.uk/~vgg/data/voxceleb/competition2021.html
[24] https://www.robots.ox.ac.uk/~vgg/data/voxceleb/interspeech2020.html
[25] https://dihardchallenge.github.io/dihard3/

boosted the performance significantly on the telephone data. Finally, we also explored new methods for overlap detection and system fusion [Lan+21b]. Our system ranked 5th in a virtual draw between the 3rd, 4th and 5th competitors[26].

BUT members were also asked to give one of the keynote talks at the challenge workshop[27]. Mireia Diez gave the talk "Variants of Bayesian HMMs for speaker diarization" focusing on the different Bayesian HMM diarization methods developed by us. This presentation is the most viewed video in the official DIHARD YouTube channel[28] with over 140 views.

5.1.3 Speaker recognition

Voices from a distance The "VOiCES from a Distance Challenge 2019"[29] is designed to foster research in the area of speaker recognition and automatic speech recognition (ASR) with the special focus on single channel distant/far-field audio, under noisy conditions. BUT team had taken part in the speaker recognition area and developed a robust system which ranked closely 2nd in the evaluation among more than 20 participants (see Figure 9). Challenge results were presented on a satellite workshop of Interspeech 2019, where participants also exchanged knowledge gained during development of systems designed for far-field and noisy data. Detailed results and system architecture was published in [Mat+19].

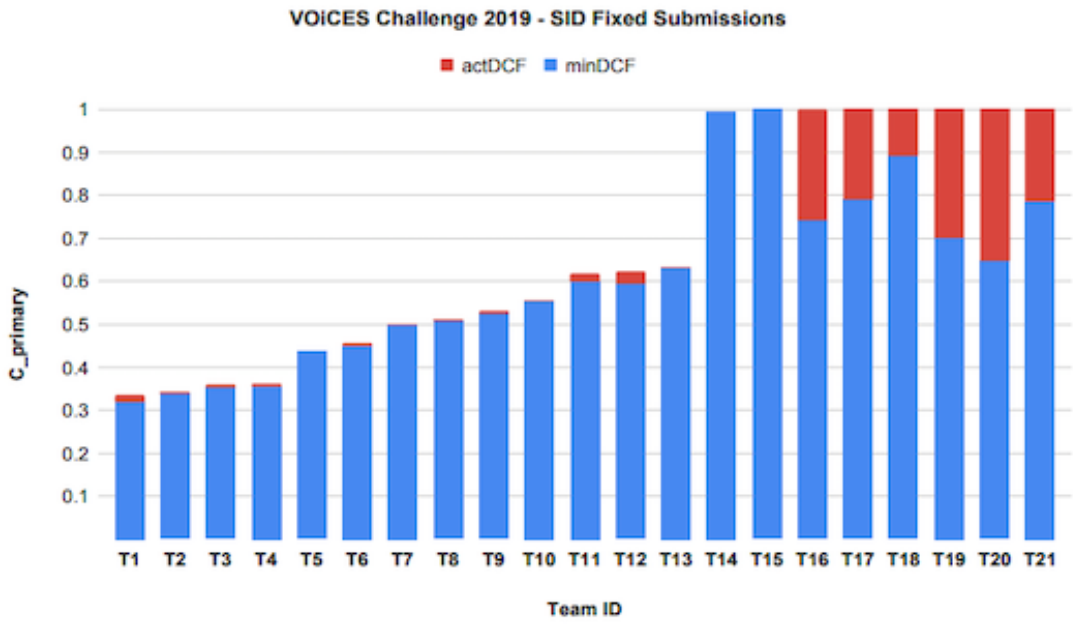


Figure 9: Results of the Voices from a Distance challenge. BUT is team number 2.

[26] https://dihardchallenge.github.io/dihard3/results
[27] https://dihardchallenge.github.io/dihard3workshop/program
[28] https://www.youtube.com/channel/UC2CGGwzc30mR2n-b-85wakQ/videos
[29] https://gab41.lab41.org/interspeech-2019-voices-from-a-distance-challenge-21ffd6ee5ef7

Voxceleb 2019 challenge The goal of this evaluation[30] was to test the performance of at the time state-of-the-art technology for speaker recognition in the "wild" data that was downloaded mostly from YouTube. The data contains various videos from celebrity interviews that are professionally made to videos taken by amateurs and low-end equipment and videos with large amount of background noise caused by different environments. We consider these evaluations as important as the voxceleb dataset serves as a favourite benchmark and the release of large training data accelerated research with deep neural models for speaker recognition.

The challenge was split into two parts: (i) Fixed training condition, where the participants were allowed to use only pre-defined voxceleb training set and (ii) Open training condition, where the participants were allowed to use any data. BUT has ranked first in both tracks and the detailed description of submitted systems can be found in [Zei+19].

SdSV Challenge In April 2020, we took part (and we co-organized) Short-duration Speaker Verification Challenge.[31] This evaluation, as the name suggests, was focused on short duration segments and the languages were Farsi and English. The data for the evaluation was collected by our former colleague Hossein Zeinali (now with the Amirkabir University of Technology, Iran) in cooperation with our team [ZČB19].

In this evaluation, we focused mainly on text-dependent speaker recognition, which is the verification task when the content of the uttered phrase (passphrase) is part of the verification. Text dependent speaker recognition is a specific problem, where the generative models like i-vectors are still state-of-the-art. During the evaluation, we focused on combination of the system based on i-vectors and a modern approach based on neural networks (x-vectors). Resulting system, which is a fusion of both approaches ranked at first place. Detailed description of all subsystems and analysis of results are available in system description [Bur+20] and Interspeech paper[Loz+20]. Overall results of the challenge are visualized in Figure 10.

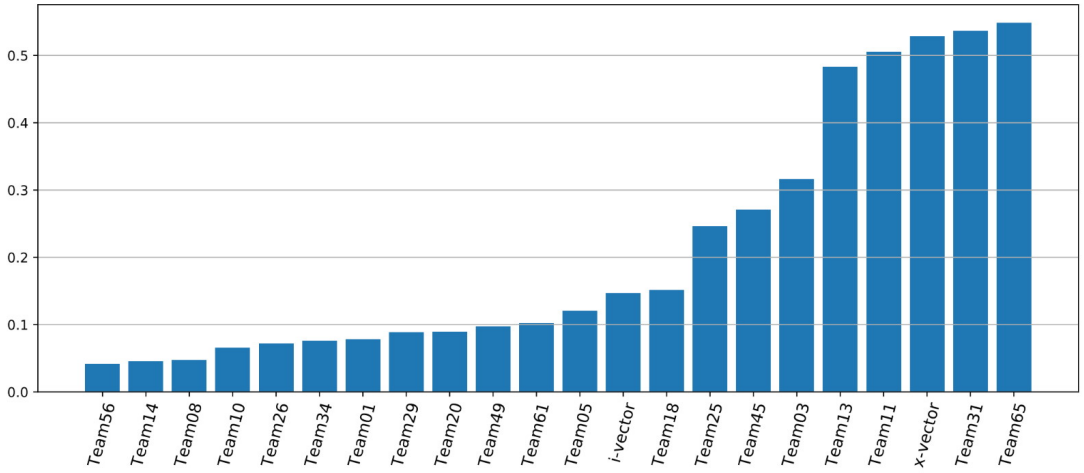


Figure 10: Short-duration Speaker Verification (SdSV) Challenge 2020 BUT is team number 56.

[30] https://www.robots.ox.ac.uk/~vgg/data/voxceleb/competition2019.html
[31] https://sdsvc.github.io/2020/results/

NIST SRE As is a tradition at Speech@FIT since 2005, we took part in NIST speaker recognition evaluations (SRE) in 2019, 2020 and 2021.

The NIST SRE2019 was focused on conversation telephone speech (CTS), but this time it included also audio-visual track, where participants were asked to make use of the two modalities (audio and visual).

BUT participated in the consortium of academic and industrial partners (Phonexia - Brno, Omilia Conversationla Intelligence - Athens, Greece, CRIM - Montral, Canada, Speechlab - Shanghai Jiao Tong University, China, and Audias-UAM - Universidad Autonoma de Madrid, Spain). Our consortium has affirmed our top position in speaker recognition among 50 other research laboratories from all over the world. Our primary submitted system was a fusion of four systems based on x-vectors. For the audio-visual part, we treated the two systems (speaker and face recognition) as independent and fused their outputs via a simple logistic regression. We achieved very good performance which confirmed that the two modalities are highly complementary. We summarized the architecture of our systems and analysis of our findings in [Ala+19] and [Ala+20].

The NIST SRE2020 CTS challenge is a still ongoing evaluation focused on addressing the language variability and other domain mismatches caused by various telephone networks. Currently our submissions rank among the top ten in the leaderboard. As the topic of the evaluation suggests, we have concentrated our efforts in domain adaptation – working mostly in the space of embeddings and PLDA backend. We have also explored various calibration schemes to address large variability in verification scores. In this evaluation, we are part of two consortiums and we cooperate both with private companies (Phonexia and Omilia) and academia (CRIM).

NIST SRE2021 was again focused on audio-visual tasks including the scenario of having conversational telephone speech tested against speech extracted from video (mostly Youtube). In this evaluation we ranked sixth and seventh in audio and audio-visual part, respectively. Although not reaching the top performance, we still have a solid place among the top performers. In this evaluation, we focused on experimentation with various embedding architectures, their duration fine-tuning and domain adaptation in embedding space.

5.2 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"[32].

According to Research.com, Lukáš also ranks as the 4th most cited Czech researcher in the area of Computer Science[33].

5.3 Impact of publications

The team leaders and senior researchers on the team have a significant citation track, the following table presents their h-indices from both SCOPUS and Google Scholar.

[32] https://www.aminer.org/ai2000/sr

[33] https://research.com/scientists-rankings/computer-science/cz

We are proud that some of the soon-to-graduate PhD students actually have excellent bibliometric results.

person	SCOPUS	G. scholar
Lukáš Burget	41	56
Martin Karafiát	28	38
Ondřej Bojar	19	36
Pavel Pecina	16	26
Karel Veselý	16	19
Murali Karthick Baskar	9	10
Anna Silnova	8	12
Jindřich Libovický	8	11
Jindřich Helcl	8	9
Santosh Kesiraju	5	8
Petra Galuščáková	5	7

Some papers published with NEUREM3 support have already gathered quite some attention and citations, for example [Žmo+19] published in prestigious IEEE Journal of Selected Topics in Signal Processing has already reached 67 citations on Google Scholar. [Die+20] published in IEEE Transactions on Audio, Speech and Language processing, gathered 21 citations. The work of Popel at al. [Pop+20], published in Nature Communications received 73 citations on Google Scholar so far.

5.4 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.[34].
- 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).

5.5 Organization of international events

5.5.1 Interspeech 2021

The first 3 years of NEUREM3 were marked by preparation and execution of Interspeech 2021[35] in Brno. Interspeech (owned by the International Speech Communication Association - ISCA[36]) is the most comprehensive and important conference in the area of speech processing, held annually. BUT team won the organization of the conference back in 2017, with Lukáš Burget serving as the lead Technical Chair and Honza Černocký (the managing head of BUT speech group) as its General chair.

Interspeech took place in Brno from August 30 to September 3, 2021. Due to the epidemiological situation, it combined full-time and online form: more than 300 scientists

[34] https://www.interspeech2021.org/best-student-paper-shortlist
[35] https://www.interspeech2021.org/
[36] https://www.isca-speech.org/iscaweb/

(mainly from Europe) visited Brno, and more than 1,600 other participants who virtually failed to arrive due to the epidemiological situation and travel restrictions of governments and companies and universities - mainly from the USA and Asia. The technical program committee headed by Lukáš selected 992 scientific articles from more than two thousand submitted for presentation.

Figure 11: Logo of Interspeech 2021.

5.5.2 Co-organization of WMT Shared Tasks

NEUREM3 allows us to stay among the organizers of WMT (conference on Machine Translation),[37] the annual flagship of shared tasks centered around machine translation.

Specifically, we co-organized the News translation task in 2019, 2020 and 2021 [Bar+19; Bar+20; Akh+21] and the Metrics task in the same years [Ma+19; Mat+20b; Fre+21].

In News task, we ensure uninterrupted inclusion of Czech since 2007, focusing primarily on the translation into Czech. Thanks to our long-term efforts in this area, Czech is treated as one of the critical languages for MT research; even the most important teams evaluate their approaches on Czech. An anecdotal evidence of this was provided by Kyunghyun Cho in his keynote (see page 57) who mentioned that the population of his home city Seoul is comparable to the total number of Czech speakers and yet he needed to ensure that neural MT works for this "strange language".

Thanks to our efforts, Czech is also one of the first languages where MT performance comparable to humans was reached, see Section 2.5.1.

In both the News task and Metrics task, we are one of the driving forces behind methodological updates of translation quality evaluation. This is necessary with the continuous progress of MT systems, to ensure that the evaluation methods (both manual and automatic) are discerning and informative. Over the NEUREM3 years, we have gradually moved to document-level manual evaluation.

As part of the News task, we also started organizing "test suites", i.e. inviting fellow researchers to provide customized test sets which focus on particular aspects of translation quality, eps. those missing in the common evaluations. Test suites have now become one of the standard WMT paper tracks, an "unshared" task where every team is doing something else.

5.5.3 Co-organization of WAT Shared Tasks

To ensure a good visibility of our Hindi Visual Genome (see Section 2.7.2), we make it the dataset of one of the tracks at WAT. Over the years, other research communities take the

[37] https://www.statmt.org/wmt19/ to wmt21

inspiration and contribute by creating Malayalam[38] and Hausa (under review) variants of the dataset. This increases the possible set of (typically low-resource) languages on which multi-modal approaches to translation can be tested.

5.5.4 SummDial Session at SIGDIAL

Ondřej Bojar was one of the organizers and panel moderator of the "Summarization of Dialogues and Multi-Party Meetings (SummDial)"[39] special session at the SIGDIAL 2021 conference.[40]

The SummDial special session on summarization of dialogues and multi-party meetings was held virtually within the SIGDial 2021 conference on July 29, 2021. SummDial @ SIGDial 2021 aimed to bring together the speech, dialogue, and summarization communities to foster cross-pollination of ideas and fuel the discussions/collaborations to attempt this crucial and timely problem.

SummDial session consisted of research papers and a panel discussion.

Six papers were accepted by SummDial, four of which were also accepted to SIGDIAL main conference. Eventually, one of these papers, "Coreference-Aware Dialogue Summarization" [LSC21] received the best paper award of SIGDIAL.

Details on the SummDial session were submitted in the form of a paper to ACM SIGIR Forum and are now under review.

5.5.5 AutoMin Shared Task, collocated with Interspeech 2021

On March 22, 2021, we announced the call for participation to our AutoMin Shared Task[41], a shared task on automatic summarization of meeting transcripts. At the same time, we have provided prospective participants with a small example of the data to be used in the task. The full training data was released for participants on May 15. The test sets followed in June and early July, giving participants up to one month for the processing of the test data.

The task organization included the preparation of the training and test data, including their de-identification to preserve privacy of participants of the underlying meetings. Once the automatic summaries were collected, we organized manual evaluation of the outputs, to complement standard automatic evaluation metrics for summarization (ROUGE). The proceedings of AutoMin Shared Task are unfortunately still in preparation, due to a delayed write-up of the overview paper.

[38] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3533
[39] https://elitr.github.io/automatic-minuting/summdial.html
[40] https://www.colips.org/conferences/sigdial2021/wp/
[41] https://elitr.github.io/automatic-minuting/

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 machine translation - Hari Vydana launched this activity making use of his extensive know-how of end-to-end ASR systems (see Section 2.6.3) and both teams cooperated on organizing and participating in the IWSLT challenge (see Section 5.1.1).

Jonas Kratochvil, 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 Vesely and Ekaterina Egorova, and he subsequently made use of this know-how when building speech translations systems back at CUNI.

6.2 Professional elevation of team members

several team-members completed their PhD in the lifetime of NEUREM3:

- Hari K. Vydana defended his Ph.D thesis "Salient Features for Multilingual Speech Recognition in Indian Scenario" at IIIT Hyderabad shortly after joining the NEUREM3 team in January 2019.
- Santosh Kesiraju defended his Ph.D thesis "Generative models for learning document representations along with their uncertainties" at the same University in January 2021.
- Two PhD students supervised or co-supervised by Lukáš Burget defended at BUT since the start of NEUREM3, both contributed to project goals despite bot being supported financially from the project: Lucas Ondel [Ond+19; Yus+21] and Ondřej Novotný [Nov+19a] (and several speaker reocgnition publications, as he was involved in several evaluation systems).
- Jindřich Libovický defended his Ph.D. thesis "Multimodality in Machine Translation" in June 2019.
- Jindřich Helcl submitted his Ph.D. thesis "Non-Autoregressive Neural Machine Translation" in November 2021, the defence is scheduled for February 2022.

6.3 Community building

6.3.1 Czech speech / NLP days

Working on a flag-ship Czech speech/NLP project, we feel obliged (and actually are also very happy) to be the leaders and organizers of the Czech speech and NLP community. On Monday 21st October, 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)[42] was invited as Keynote speaker.

The planned to continue this event in 2020 but these plans were disrupted by the Covid pandemics.

In 2021, it was replaced by Interspeech 2021 (see Section 5.5.1) that attracted most of the Czech speech and NLP communnity. We plan to continue these events in 2022 and beyond.

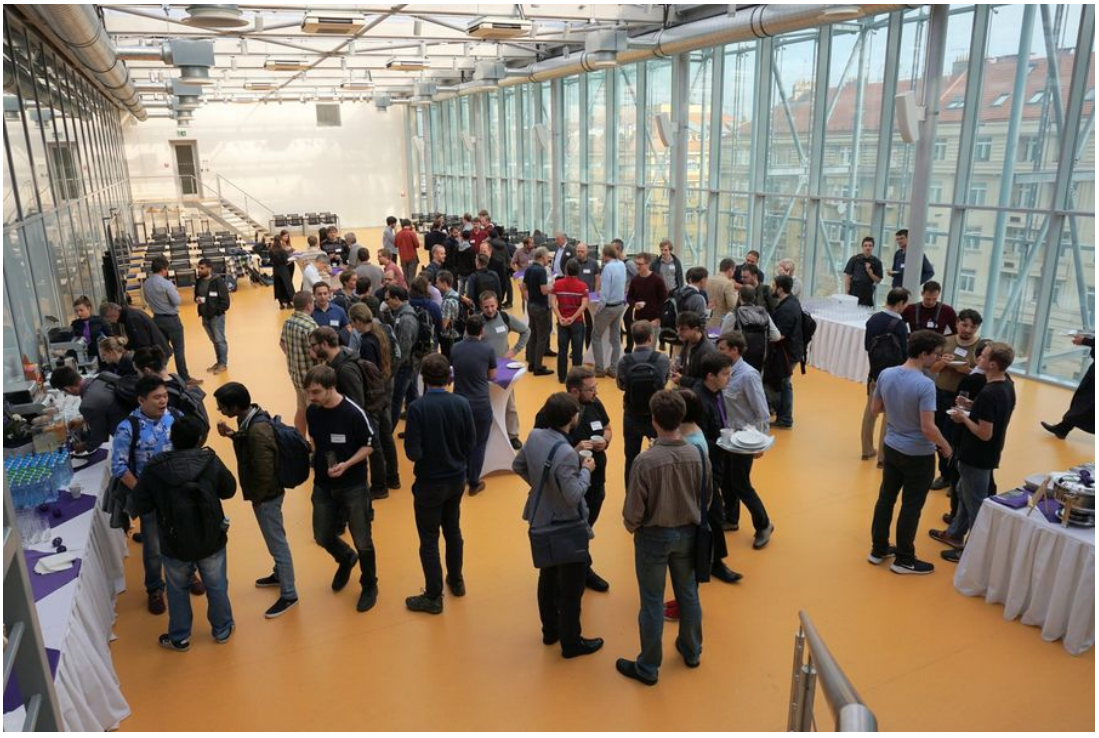


Figure 12: Czech speech/NLP community gathering at CIIRC CTU Prague.

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. In 2019, several team members were at the founding of AICzechia[43] 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 (see web page of AICzechia, in Czech), that now forms part of the National AI strategy adopted by the Czech Government in May 2019[44].

[42] https://kyunghyuncho.me/

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

[44] https://www.vlada.cz/assets/evropske-zalezitosti/umela-inteligence/NAIS_kveten_2019.pdf

The team members are also active in the efforts of regional organizations supporting AI in Prague[45] and Brno[46].

The team members are involved in the preparation of a National Center of Competence proposed by the Technology Agency of the Czech Republic (TACR NCK) — while NEUREM3 concentrated on the basic research, the studied techniques are also highly interesting for potential industrial exploitation. The TACR NCK proposal include top Czech speech and NLP labs, as well as innovative local industry.

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 a newly accredited Master program "Information Technology and Artificial Intelligence"[47] 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 is finishing its LangTech project (2017-2022)[48] where both Ondřej Bojar and Pavel Pecina are involved. The project has modernized most of the classes on NLP taught by CUNI and it is 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"[49]. 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; from this academic year, it is renamed to Perception because it now includes also computer vision. As a result, prg.ai Minor facilitates "cherry-picking" of the best AI subjects across the schools, and through this comparison and increased visibility motivates also the teachers to improve their classes. As a result of this activity, all participants of prg.ai Minor have a broader view of AI research scene in Prague and can better choose their next study steps or research directions. Prg.ai Minor is designed for enrolled students of Charles University and Czech Technical University at all levels (Bc, MSc, PhD). Prg.ai is limiting the number of enrolled students so that the participating schools are not unexpectedly flooded, and it has also ensured smooth administrative handling for such school "visits".

[45] https://prg.ai/; with Ondřej Bojar being one of the initiators of the whole idea
[46] https://www.brno.ai/
[47] https://www.fit.vut.cz/applicants/degree-programme/.en#mgr
[48] https://ufal.mff.cuni.cz/grants/langtech
[49] https://prg.ai/minor/

7 Plans for the next period

7.1 Overall plans

The project will continue in the five broad areas as defined in Section 1.2. The five tasks defined at the outset of the project

- Task 1. Multi-linguality in ASR, NLP and MT.
- Task 2. Multi-modality in ASR, NLP and MT.
- Task 3. Rich input, intermediate and output representations in neural ASR, NLP and MT systems.
- Task 4. Hierarchies and automatic inference of units.
- Task 5. Text to text and speech to text translation based on non-parallel and heterogeneous training data, robustness towards noise.

are still relevant and there is a lot to do in all of them, therefore, they will be further elaborated. In addition, we have defined three new tasks:

- **Task 6. Personality and individual adaptation** - the work in speaker recognition, target speaker extraction, individual adaptation and relation to the personality of author/writer/speaker merits an independent task.
- **Task 7. Semantic processing** beyond speech, NP and MT which is already reflected in currently elaborated research topic "Towards Semantics" (Section 2.8).
- **Task 8. Human performance and human factors** - this field of research, already gathering substantial attention (Section 2.9) is also ready for the definition of a new Task.

7.2 Detailed plans

In the area of *speaker embeddings*, we plan to analyze the basic properties of speaker embeddings. For example, it is common to try to produce speaker embeddings whose distribution is domain invariant. However, it has not been analyzed to what extent this is actually possible. An obvious example that contradicts the idea of domain invariant speaker embeddings is that speaker recognition is inherently harder in some domains than in others. Accordingly, embeddings from such two domains cannot follow the same distribution. As another example, it is not understood how much information about the speaker identity is lost when converting a variable duration utterance into a vector of fixed dimension. Further, our work in [Wan+19] on removing phonetic information from speaker embeddings could be useful not only due to the improved speaker recognition performance but also from a privacy point of view. Therefore an analysis on to what extent various attributes can be removed from speaker embeddings without hampering the speaker recognition performance is also an important future work.

In the area of *Neural signal processing for speaker recognition*, we want to focus on multi-channel processing and NN-based beamforming techniques and end-to-end training of such architectures. In the single-channel domain, we also want to explore learnable

feature extraction NN layers and explore the benefits of fine-tuning such feature extraction towards a specific application. We also want to explore advanced data augmentation and simulation by exploiting large-scale pre-trained models for speech synthesis or voice-conversion and explore the robustness and performance of systems trained with additional synthetic speakers on top of original training data.

In the area of *target speech extraction*, we would like to explore the possibility of unsupervised adaptation of the method on realistic data in the future. This is needed as current methods of target speech extraction and speech separation often fail on realistic data, with spontaneous speech, noise, and reverberation.

In the area of *integration of variational auto-encoders and spatial clustering*, we would like to test the method to more realistic types of noises in our future work. This might require better noise modeling, e.g. based on non-negative matrix factorization. The explicit speaker latent variable in the model also opens the way for using additional speaker information, which has not been explored currently.

In the area of *speaker diarization*, we plan to further extend VBx as well as to explore and extend the new EEND diarization systems.

In the area of *Low-resource speech recognition*, we plan to continue working on developing ASR for the air-traffic domain. In the modeling part, we plan to experiment with wav2vec 2.0 features. We are also working on a contextual adaptation for ASR with rapidly changing context information, and we work on semi-supervised training. And, we also currently work on Czech ASR for meeting recognition that should be customizable for user domains.

In the area of *voice conversion*, we plan to incorporate our ideas into more complicated models such as ones inspired by advances in the text-to-speech area. We also plan to use pre-trained features e.g. wav2vec 2.0 which has recently shown its potential in use for voice conversion.

In the area of *language modeling for ASR, NLP and OCR*, we plan to investigate further into efficient schemes of data augmentation for high error rate scenarios. We believe that it should be possible to obtain further improvements from utilizing real recognizer errors. This direction should also provide further insight into the information kept in the internal representations of the LMs. In the area of utilizing embedding uncertainty information, we would like to explore alternative NN architectures for extracting embeddings with uncertainty, alternative approaches to training such networks (i.e. different training objectives). Also, we would like to apply a probabilistic embedding framework to the speaker verification task.

In the area of *OOV detection and recovery*, we plan to jointly train the tasks of word prediction and spelling prediction. Jointly solving OOV detection and OOV recovery would allow the two tasks to benefit from additional information, and would also eliminate the problem of time alignment of detected OOVs and phoneme output which we face in the approaches that use two separate systems.

In *multi-lingual machine translation*, we will focus on experimental exploration of the area of with the goal to propose and evaluate variations of NMT model architectures, training data layout and training methods to achieve gains in quality or efficiency.

In the area of *multi-source MT*, we will propose methods to benefit from the additional source languages, if available at the moment, and gracefully handle the situation when the additional parallel source is unavailable. More specifically, we focus on the MT

part of the cascaded simultaneous speech translation, with multiple independent ASR sources. We will primarily experiment with English and German as source languages and Czech as a the target. We plan to investigate combination of currently existing methods for multi-source text translation [ZK16; Fir+16; DCK17] with methods for simultaneous speech translation [Nie+18; Boj+21] and our [Mac+20]. Namely, we plan to use supervised training on multi-parallel or multiple pairs of bi-parallel data. Further details on the plans and the current achievements are described in [Mac21].

In the area of *low-resource translation*, we plan to experiment with generative adversarial networks to create synthetic sentences to augment existing data sets. We will also investigate the role of terminology in domain-specific translation and focus on latent representations of the terms.

In the area of *multi-modality and eye-tracking of translation*, we plan to train multimodal models on the same tasks that the participants in our eye-tracking experiment were required to perform. The aim is to compare the nature of representations learnt by the models with the collected cognitive data.

In the area of *machine, human, and superhuman translations*, we plan to investigate the impact of evaluators' translation proficiency on the translation quality assessment: we aim to investigate the differences in the translation quality assessment between various groups of annotators: non-translators, translation students, and professional translators. Furthermore, we are focusing on the process of creating so-called superhuman translations, i.e., we are looking for the optimal method that will lead to high-quality human translations. We plan to have the superhuman translations evaluated by annotators in the context of other human and machine translations. We also intend to evaluate the ability of existing automatic MT metrics to use superhuman translations as references; when using superhuman references, automatic metrics should still agree with human assessments of candidate translation quality. This is potentially challenging for such metrics, as many rely either directly or indirectly on string overlap to judge translation. However, high-quality superhuman translations may make syntactic or lexical choices that result in low string overlap with reasonable translations from existing MT systems. Thus, we should verify that when using superhuman references, automatic metrics give moderate scores to reasonable translations, while still promoting higher-quality translations.

In the area of *summarization*, we have so far taken more of the role of facilitator and organizer. We will continue with further data refinement and further shared task instances but we will also move to the modelling phase, see below. The dataset we prepared for AutoMin 2021 has been so far made available only to the registered participants, upon signing a non-disclosure agreement. We took this approach to protect any sensitive information that might have appeared in the meeting transcripts and minutes. The dataset was already de-identified, so it was GDPR-compliant for AutoMin 2021, but one or two participants suggested that even when anonymized, some texts may contain sensitive or inappropriate parts. We are now in the process of yet another manual revision round, targeted at removing any such content and we will release the dataset fully publicly.

In the area of *compositionality and disentanglement*, we focus on studying emergent languages and transfer the findings to more general deep learning models.

Design and development of methods of *automatic meeting summarization* is impossible without automatic metrics of output quality, and these in turn need a reliable manual

golden truth. The AutoMin 2021 manual evaluation was limited to Likert-scale scoring and the results were rather little discerning. We will refine this scoring using an annotation interface which we already developed (described in an LREC 2022 submission, now under review). It is clear from our preliminary analysis of AutoMin data, that humans adopt diverse strategies when taking notes and reference-based style of evaluation is thus problematic. Instead, we plan to develop source-based methods. The first step is to explicitly align parts of the transcripts to items in the proposed minutes, a task which also resembles topic segmentation. We already started with this and according to our expectations, sentence embeddings using large LMs are promising for content-heavy sentences but fail for short segments. With automatic evaluation ready, we will start devising summarization models. Transformer will serve as the primary baseline but we expect it will be a challenge to train it to identify the implicit structure of an informal meeting or construct it for meetings that went not quite organized. The problem will be to find out how to best harness transfer learning across domains and summarization styles because the training data for minuting will remain very scarce.

7.3 Planned ERC proposals

The proposal(s) for the European Research Council (ERC) grants are a required output of the project. These will be submitted in the final two years of the project with possible topics ranging from ASR in different dialects, multi-modal ASR, over high-quality speech translation, e.g. benefiting from human interpretation up to the psycholinguistics and neurolinguistics areas, where we will try to design more human-inspired deep neural networks and compare their behaviour on complex multi-modal language processing tasks with that of humans.

8 Project outputs

8.1 Software

VBHMM x-vectors Diarization (aka VBx) was released in February 2020[50]. Since then, it has received more than 130 stars in GitHub, more than 35 forks and we have answered more than 40 questions asked by the community (between public issues in the repository and personal emails sent to us) mostly related to how to use the toolkit. Furthermore, the model and software have been widely adopted by the community: not only we have used it with great success in challenges as mentioned in 5.1.2 but also several other teams have used VBx as part of their systems:

- two out of four teams in the VoxCeleb Challenge 2020[51] made use of VBx.
- six out of seventeen teams in the Third DIHARD Challenge (2020)[52] made use of VBx, including the first and second teams according to the ranking.
- three out of seven teams in the VoxCeleb Challenge 2021[53] made use of VBx.

In all cases, more teams participated in the challenges but they did not submit their systems descriptions so we cannot know if they also used VBx.

Since the publication of the original VBx codebase and the follow-up versions, more than 60 peer-reviewed papers have cited at least one of the papers where we describe the model [Lan+22; Die+19]. Of those, more than 30 have used VBx either as a baseline or as part of their proposed diarization solutions.

SLTev is a tool for evaluating (simultaneous) speech recognition and translation. It received the **outstanding demo award** at EACL 2020. [Ans+21]. See page 48 for more details.

ASR and MT transformer models The code for training models related to our work in end-to-end ASR and SMT (Section 2.6.3) were released:

- Transformer_ASR/Transformer_E2E-ST [54]
- Transformer_MT code [55]

and the joint training ASR-MT code will be opesourced soon.

Mashcima a library that produces synthetic images of monophonic handwritten music[56] presented in [MP21].

[50] https://github.com/BUTSpeechFIT/VBx
[51] https://www.robots.ox.ac.uk/~vgg/data/voxceleb/competition2020.html
[52] https://dihardchallenge.github.io/dihard3/
[53] https://www.robots.ox.ac.uk/~vgg/data/voxceleb/interspeech2021.html
[54] https://github.com/BUTSpeechFIT/ASR_Transformer.git
[55] https://github.com/BUTSpeechFIT/MT_Transformer.git
[56] https://github.com/Jirka-Mayer/Mashcima

8.2 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, the data-sources created or generated include:

- ASR data for Czech: Large Speech Corpus for Czech and Speech test set with additional relevant texts (Section 2.3.1).
- Synthesized Training Data for Handwritten Music Recognition (Section 2.7.3).
- COSTRA: Corpus of Complex Sentence Transformations, see Section 2.8.4
- Data for Meeting Summarization shared task Section 2.8.6.

Moreover, we are preparing a release of MultiSV data-set for multi-channel speaker recognition training and testing.

8.3 Publications

From 2019 till 2021, NEUREM3 has produced 96 publications of which 12 in peer reviewed journals 49 at top conferences and 35 at local workshops, challenge and evaluation workshops, etc. The following sections list them as per years:

2019 Publications

[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 Computa-

tional 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. ISBN: 0000000000. 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 International 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.

[VB19] 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.

[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

[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. 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. 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

[Akh+21] 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. 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+21b] 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.

References (not project's outputs)

[Fir+16] Orhan Firat et al. "Zero-Resource Translation with Multi-Lingual Neural Machine Translation". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, 2016, pp. 268–277. DOI: 10.18653/v1/D16-1026. URL: https://aclanthology.org/D16-1026.

[Mül+16] Markus Müller et al. "Lecture Translator - Speech translation framework for simultaneous lecture translation". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. San Diego, California: Association for Computational Linguistics, 2016, pp. 82–86. DOI: 10.18653/v1/N16-3017. URL: https://aclanthology.org/N16-3017.

[ZK16] Barret Zoph and Kevin Knight. "Multi-Source Neural Translation". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, 2016, pp. 30–34. DOI: 10.18653/v1/N16-1004. URL: https://aclanthology.org/N16-1004.

[Joh+17] Melvin Johnson et al. "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation". In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 339–351. DOI: 10.1162/tacl_a_00065. URL: https://aclanthology.org/Q17-1024.

[Nie+18] Jan Niehues et al. "Low-Latency Neural Speech Translation". In: *INTERSPEECH*. 2018.

[Pop18] Martin Popel. "CUNI Transformer Neural MT System for WMT18". In: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, 2018, pp. 482–487. DOI: 10.18653/v1/W18-6424. URL: https://aclanthology.org/W18-6424.

[Tan+18] Chuanqi Tan et al. "A survey on deep transfer learning". In: *International conference on artificial neural networks*. Springer. 2018, pp. 270–279.

[Con+19] Alexis Conneau et al. "Unsupervised cross-lingual representation learning at scale". In: *arXiv preprint arXiv:1911.02116* (2019).

[DCK20] Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. "A Survey of Multilingual Neural Machine Translation". In: *ACM Comput. Surv.* 53.5 (2020). ISSN: 0360-0300. DOI: 10.1145/3406095. URL: https://doi.org/10.1145/3406095.

[Boj+21] Ondřej Bojar et al. "ELITR Multilingual Live Subtitling: Demo and Strategy". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, 2021, pp. 271–277. DOI: 10.18653/v1/2021.eacl-demos.32. URL: https://aclanthology.org/2021.eacl-demos.32.

[Mac21] Dominik Macháček. *Multi-lingual Machine Translation, doctoral pre-thesis.* Charles University, Prague, 2021. URL: https://ufal.mff.cuni.cz/~zabokrtsky/pgs/thesis_proposal/dominik-machacek-proposal.pdf.