COMPRISE

Cost effective, Multilingual, Privacy-driven voice-enabled Services

www.compriseh2020.eu

Call: H2020-ICT-2018-2020
Topic: ICT-29-2018
Type of action: RIA
Grant agreement Nº: 825081

WP Nº6:	Evaluation and demonstration for practical use cases
Deliverable Nº6.2:	Initial scientific evaluation
Lead partner:	USAAR
Version Nº:	1.0
Date:	28/02/2020

European
Commission

Document information	
Deliverable Nº and title:	D6.2 – Initial scientific evaluation
Version Nº:	1.0
Lead beneficiary:	USAAR
Author(s):	Thomas Kleinbauer (USAAR), Askars Salimbajevs (TILDE), Imran Sheikh (INRIA)
Reviewers:	Irina Illina (INRIA), Raivis Skadiņš (TILDE)
Submission date:	28/02/2020
Due date:	29/02/2020
Type[1]:	R
Dissemination level[2]:	PU

Document history			
Date	Version	Author(s)	Comments
27/01/2020	0.1	Thomas Kleinbauer	Content draft
15/02/2020	0.2	Thomas Kleinbauer, Askars Salimbajevs & Imran Sheikh	Initial version
20/02/2020	0.3	Thomas Kleinbauer, Askars Salimbajevs & Imran Sheikh	Revision based on the reviewers' comments
28/02/2020	1.0	Emmanuel Vincent & Zaineb Chelly Dagdia	Final version reviewed by the coordinator and by the project manager

[1] **R**: Report, **DEC:** Websites, patent filling, videos; **DEM:** Demonstrator, pilot, prototype; **ORDP:** Open Research Data Pilot; **ETHICS:** Ethics requirement. **OTHER:** Software Tools

[2] **PU:** Public**; CO:** Confidential, only for members of the consortium (including the Commission Services)

Document summary

Scientific evaluation is one of the two main contents of Work Package 6 (WP6), the other one being the development of prototype demonstrators. While the latter focuses on the practical use of COMPRISE results with commercial application in mind, the former highlights the project's achievement from a scientific point of view. This deliverable focuses on the evaluation part of WP6. It describes the results of the first combined evaluation.

A specific evaluation setup was developed and is presented that takes into account the current state of the project, both in terms of the progress made on the individual components as well as the overall integration.

We present a complex experimental setup that allows us to study the effect of multiple parameters, such as a) the effect of privacy-transforming speech data before adding it to a training set; b) the effect of adding different amounts of such data to the training set; c) the role of in-domain vs. out-of-domain data; and the benefit of training Machine Translation specifically for spoken language.

In addition to the experimental setup, the employed datasets and metrics are introduced that allow for reproducibility of the experiments, an important aspect of any scientific measurement.

The outcomes of the experiments are thoroughly analysed and discussed.

This is only the first combined evaluation of COMPRISE. For the future, we plan further modifications and additions to the experiments of this report.

Table of contents

1. Introduction

COMPRISE is a multi-faceted project that unites commercial interests with scientific perspectives. Work Package 6 exemplifies this synergy by concerning itself with two lines of work: the development of prototype demonstrators in tasks T6.2, T6.3, and T6.4, and scientific evaluation in T6.1. The evaluation of novel approaches is a key task in any scientific work, and consequently work packages WP2, WP3, and WP4 all address it at an individual level. In addition, this deliverable reports on the first *combined* evaluation in which some of the developments of the aforementioned work packages are evaluated as they interact with each other. This is an important aspect in COMPRISE for two reasons.

First, the *operating branch* consists architecturally of a pipeline that augments the classic dialogue system architecture with a Machine Learning component. Since automatic dialogue processing is inherently prone to errors, a long pipeline architecture bears the risk of accumulating noise and errors, thus eventually leading to an unsatisfactory user experience. A combined evaluation can give a realistic assessment of these effects.

A second reason why it is important to evaluate COMPRISE approaches, not just individually but also in a combined fashion, is the project’s focus on the *training branch*: one of COMPRISE's main goals is to enable the provider of voice-based systems to collect voice interaction data in a privacy-preserving manner and continuously retrain models using this data. However, as the data passes through a number of processing steps before getting stored on a cloud platform, e.g., speech-to-text, speech/text transformation, the effect of these steps with respect to the usefulness of the resulting data needs to be studied and measured carefully. This involves subsequent (weakly) supervised learning approaches that are the main incentive for collecting data in the first place.

These points describe the long-term vision for the scientific evaluation for the whole duration of the project. At the current time and for this first combined evaluation, we are interested in the effects that arise from having multiple COMPRISE components interact with each other. At the same time, some of the innovations of COMPRISE will also be usable in diverse external settings, i.e., even outside of dialogue system applications. For instance, text transformations could potentially be employed to securely aggregate patient records for medical studies. Such stand-alone tools play an important role for the dissemination of COMPRISE. Regarding the terminology used in this deliverable, however, since the underlying technology of such tools is not fundamentally different from their use as a component in a bigger system, we will use the terms “tool” and “component” interchangeably.

The integration of the components in COMPRISE is realised through the SDK developed in Work Package 4. Since this work package is still ongoing (it will, in fact, end at M30), not all components are currently fully integrated. For the first combined evaluation covered by this deliverable, we thus carefully chose a specific evaluation setup (see Section 2) that allows a first assessment of component inter-dependencies given the current state of development.

To this end, the work reported on in this deliverable is tightly related to other project deliverables. Most notably, it is tied to D3.1 – “Initial multilingual interaction library” which presents research on combining speech-to-text and machine translation (submitted to the European Commission on February 28, 2020, Public), and to deliverable D4.1 – “SDK software architecture” which reports on the SDK software architecture that provides the backbone for combining COMPRISE components (submitted to the European Commission on November 29, 2019, Public). It should be noted, though, that the evaluation was carried out in *batch mode* where, instead of passing each user input through the full evaluation chain at once, each stage receives and processes all test inputs before any of them are passed to the next stage. There is no fundamental difference in terms of the achieved outputs, but the practical process is facilitated greatly.

The remainder of this deliverable is structured as follows. First, we introduce the general setup for the first combined evaluation. Next, we tend to the specification of datasets used to evaluate and introduce various evaluation metrics. To illustrate what exactly is measured and evaluated, the individual components are introduced in more detail. After that, we present the outcome of the evaluation and discuss our findings. Finally, we conclude and present our plans for the remainder of the project.

2. Evaluation setup

2.1 Overview

The overall architecture of the COMPRISE system is complex (see Figure 1). While full completion is still on its way, it is all the more important to run combined evaluations early on in order to correct possible misdirections on time.

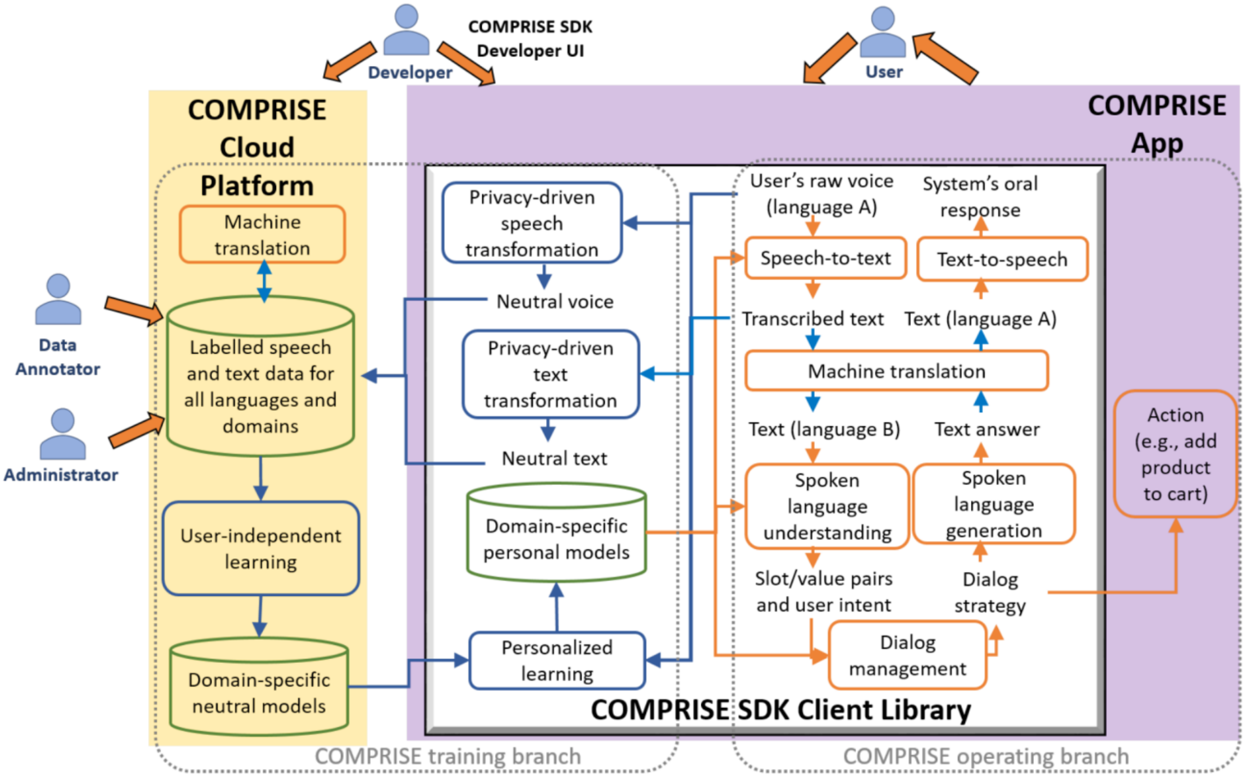
Developer
COMPRISE SDK Developer UI
User
COMPRISE Cloud Platform
COMPRISE App
Machine translation
Labelled speech and text data for all languages and domains
Data Annotator
Administrator
User-independent learning
Domain-specific neutral models
Privacy-driven speech transformation
Neutral voice
Privacy-driven text transformation
Neutral text
Domain-specific personal models
Personalized learning
User's raw voice (language A)
System's oral response
Speech-to-text
Text-to-speech
Transcribed text
Text (language A)
Machine translation
Text (language B)
Text answer
Spoken language understanding
Spoken language generation
Slot/value pairs and user intent
Dialog strategy
Dialog management
Action (e.g., add product to cart)
COMPRISE SDK Client Library
COMPRISE training branch
COMPRISE operating branch

Figure 1: The COMPRISE system architecture.

In this first combined evaluation, we focus on the initial parts of the operating and training branches, consisting of the combination of four components (see Figure 2):

- Privacy-driven speech transformation
- Speech-to-text (STT)
- Machine translation (MT)
- User-independent learning

Note that, while eventually the parts of the training branch outside the user's device will run on the COMPRISE Cloud Platform, we used a standard PC to perform the User-independent learning for the evaluation. This constitutes merely a technical difference owing to the fact that the cloud platform will not be fully available until M30 (see Work Package 5), but has no influence on the validity of the evaluation results.

One of COMPRISE's main aspirations is to create a privacy-preserving infrastructure for *collecting* user data and subsequently *using* them to learn better models for the operating branch. To guarantee privacy, however, all user speech and text data are transformed prior to storage. As these transformations are typically lossy, it is not self-evident whether using such transformed data for training will actually result in improved or degraded machine learning performance.

Therefore, the main objective of this evaluation is **to measure the effect of adding privacy-transformed speech to STT training data on**:

1. STT performance
2. MT performance

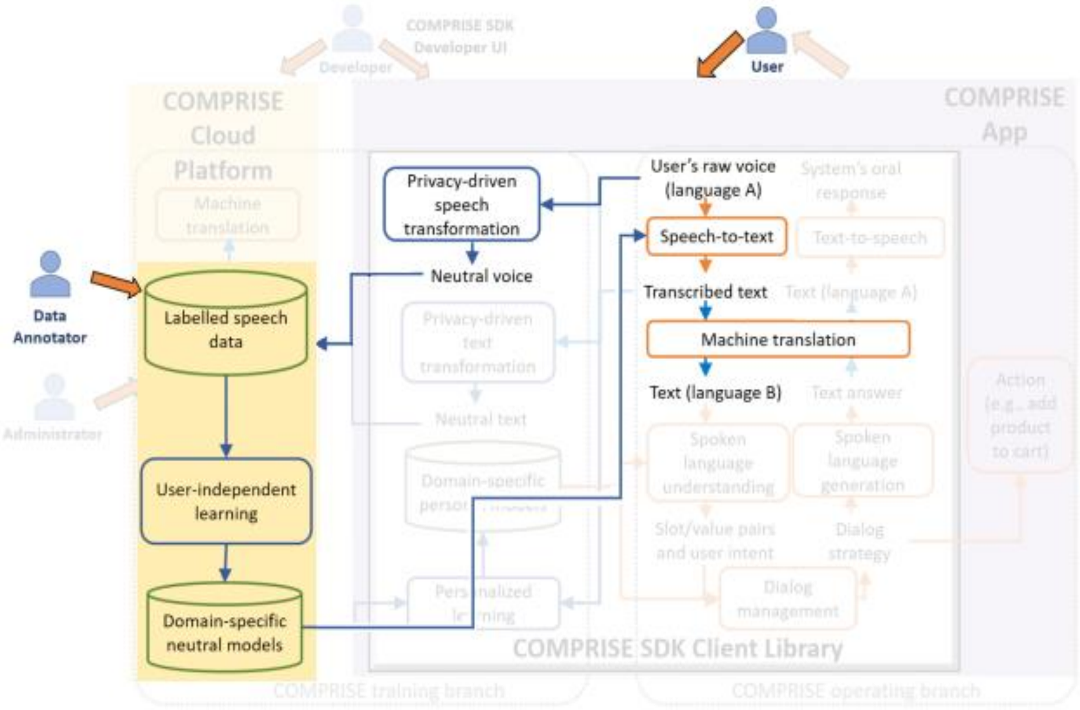

Figure 2: Components of the first combined evaluation.

2.2 Experimental design

The experiment we devised compares two setups — one with and one without privacy-preserving speech transformation — with respect to the inclusion of additional training data on top of an existing baseline dataset. The language chosen for all speech data is Latvian.

In the first setup, we add original, untransformed speech data to the baseline dataset by increments of 10 hours (0, 10, 20, 30, …, 100). We train a new STT model for each increment and calculate the STT performance on a test corpus. Next, we perform MT of the STT output and evaluate the quality of the translation. Figure 3 illustrates the process schematically.

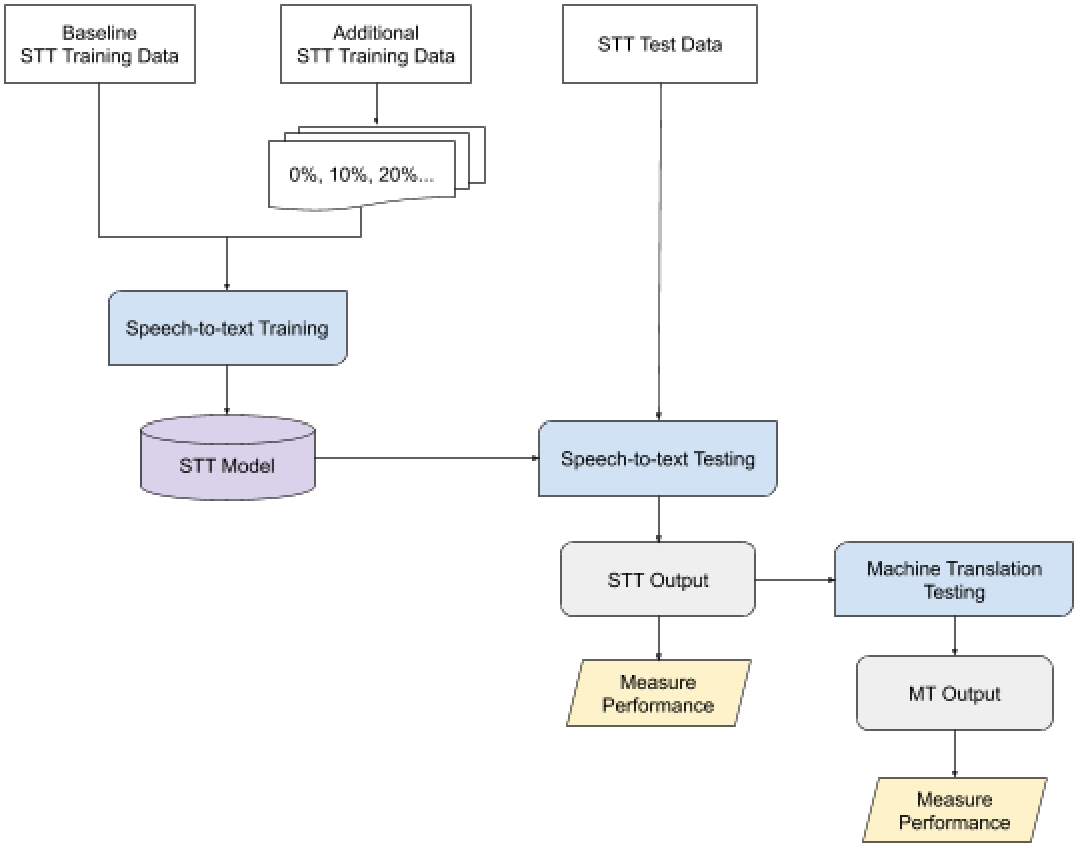

Figure 3: The first experimental setup of the evaluation.

The second setup is similar to the first one. The only difference is that the additional training data is privacy-transformed prior to adding it to the baseline data. Training and evaluation follow the same procedure as in the first setup. This allows us to not only measure the impact of using privacy-transformed data in addition to untransformed data, both immediately on STT performance and on downstream MT quality, but also to gain an intuition about which role the quantitative ratio between untransformed and transformed training data plays. In both experiments, we add the respective portions of privacy-transformed speech training data to the baseline dataset in the same order, i.e., STT models are trained on exactly the same speech recordings in either setup.

Figure 4 displays the schema of the second experimental setup. Note the insertion of the privacy-preserving voice transformation, marked in red, as opposed to the first setup.

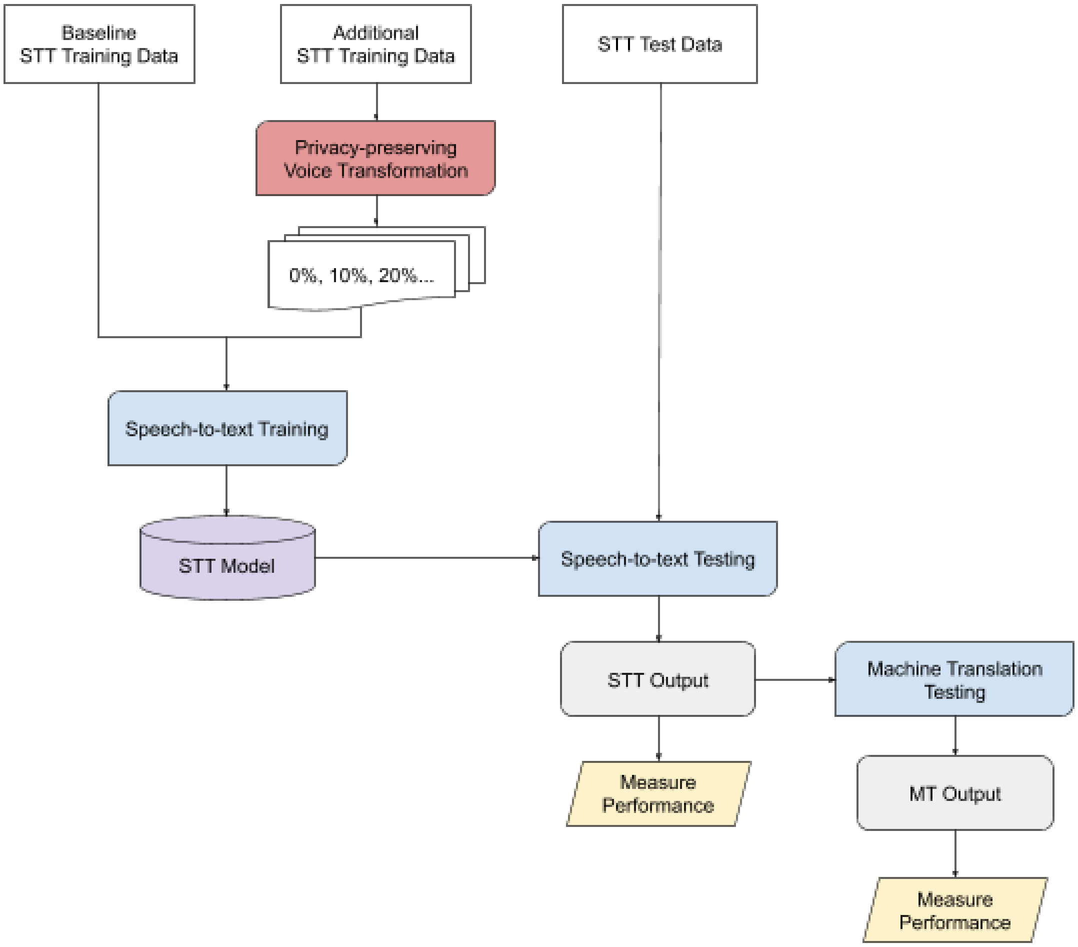

Figure 4: The second experimental setup of the evaluation.

2.3 Component details

In the following, we specify the details of the various components used in both experimental setups.

2.3.1 Speech-to-text / User-independent learning

The open-source Kaldi toolkit (Povey et al., 2011) is used to train and evaluate Latvian STT models. The two crucial parts in the standard speech recogniser architecture are the *acoustic model* which encodes pronunciation information and the *language model* which basically encodes grammar information.

We train end-to-end Factorized Time Delay Neural Network (TDNN-F) (Povey et al., 2018) acoustic models with Lattice-Free Maximum Mutual Information (LF-MMI) in a flat-start manner (Hadian et al, 2018). The model architecture and hyper-parameters are copied from the recipe for the Wall Street Journal (WSJ) dataset (Paul & Baker, 1992), which has a similar size (80 hours).

Because Latvian has highly phonemic orthography, word pronunciation is modelled by treating each grapheme as a separate phoneme.

For STT language modelling, we employ a sub-word 4-gram language model. This model is trained on a 40 M sentence text corpus collected from Latvian web news portals and has a sub-word unit vocabulary which is generated using the Byte-Pair Encoding (BPE) method. N-grams are pruned to about 100 MB so that the decoding process can fit in 2 GB of RAM. Correct sub-word unit combination is ensured by a modified decoding graph (Smit et al., 2017).

2.3.2 Voice Transformation

For the privacy transformation in the second setup, we use the voice transformation tool developed within COMPRISE as described in Deliverable D2.1 – "Baseline speech and text transformation and model learning library" of Work Package 2 (submitted to the European Commission on August 30, 3019, Public). This tool applies the VoiceMask voice conversion technique, which is inspired by Qian et al. (2017) and Qian et al. (2018). It is based on Vocal Tract Length Normalisation (VTLN) (Cohen et al., 1995; Eide & Gish, 1996). After using standard signal processing methods to compute spectral envelope, pitch, and aperiodicity features, VoiceMask modifies the spectral envelope through frequency warping. To provide privacy, this method is based on the composition of a quadratic function and a bilinear function using two different parameters. The inverse of this transformation is much more difficult to compute, and therefore more resistant to attacks.

2.3.3 Machine Translation

For MT evaluation, we use Neural Machine Translation (NMT) systems developed in Task T3.1 and described in Deliverable D3.1 – "Initial multilingual interaction library" of Work Package 3 (submitted to the European Commission on February 28, 2020, Public). NMT systems are trained on the Latvian-English WMT 2017 training dataset (Bojar et al., 2017) using the *Marian* NMT toolkit (Junczys-Dowmunt et al., 2018). The models are based on the self-attentional Transformer architecture (Vaswani et al., 2017) and use the Marian base model configuration for the model hyper-parameters. The words in the training dataset were split into sub-word units using byte-pair encoding. For this task, we used *SubwordNMT*[3] (Sennrich et al., 2016).

For this first scientific evaluation, we used two NMT systems:

- A baseline MT system trained on original parallel data (written language only).
- An adapted MT system trained on both original and synthetic parallel data. Synthetic data imitates STT output and allows the system to translate spoken language with higher accuracy.

2.4 Datasets

For the training of the STT models, we use two speech corpora:

- The 100 h Latvian Speech Recognition corpus (Pinnis et al, 2016) as baseline STT training dataset.

[3] https://github.com/rsennrich/subword-nmt

- A 100 h subset of the Latvian Parliament Speech corpus (Salimbajevs, 2018) as additional data that is appended to the STT training dataset.

A subset of 100 hours was taken from the second corpus to make the total length of both corpora comparable.

Testing was performed on two evaluation datasets:

- The Tilde Balss S2MT test set.
- An in-domain test set.

The Tilde Balss S2MT test set is a subset of data collected by the Tilde real-time Latvian STT engine and was created specifically for the evaluation of speech-to-text MT. It consists of 1,159 utterances (queries and short messages) and has been manually translated to English.

Because the speech recordings which are appended to the baseline training dataset come from a particular domain (Latvian Parliament session recordings), it was decided that the results should also be evaluated on data from the same domain. This in-domain test set contains 439 utterances (1 hour) from recordings of debates in the Parliament of Latvia from 2014 to 2016, containing contributions from about 300 different speakers. The recording time period does not overlap with the aforementioned Latvian Parliament Speech corpus, so as to guarantee that all utterances in the training and test sets are distinct. However, we could not exclude a possible overlap in the speakers.

2.5 Metrics

In order to assess the effect of adding either untransformed or transformed training data at different increments, we use standard metrics for both the STT performance and the MT quality.

For STT, a standard measure for the performance of a specific system is the Word Error Rate (WER). Given a sequence of words, both as a recording of someone speaking these words as well as a reference transcription, the output of the STT component is compared to the reference transcription. In the ideal case, all words are the same. If that is not the case, we can distinguish between three cases: the STT component erroneously 1) inserted or 2) deleted a word, or 3) substituted a word for another one.

With that in mind, the WER is defined as:

$$\mathrm{WER} = \frac{I + D + S}{N}$$

where *I* is the number of inserted words, *D* the number of deleted words, and *S* the number of substituted words. *N* is the number of words in the reference transcript.

For MT, we use the BLEU metric (Papineni et al., 2002) which is the usual metric in the research community. Generally speaking, evaluating the quality of any translation is challenging because there is often more than one good translation for any given source text. This means that, unlike in the STT case where there is exactly one correct reference *transcription* of any spoken sentence, in the MT case there might be many correct reference *translations* for a given input sentence. The BLEU metric takes this into account by comparing the output of an MT component with multiple reference translations. However, it is also applicable in cases when a single reference translation is provided.

The computation of the BLEU metric is based on n-gram overlap between the translation produced by the system to be evaluated and the given reference translation(s), with the intuition that the more a translation overlaps with reference translations, the more likely it is to constitute a good translation itself. The BLEU metric is defined as follows:

$$\mathrm{BLEU} \;=\; \mathrm{BP} \cdot \exp\left(\sum_{i=1}^{n} w_i \log p_i\right)$$

where w_i is a weight factor, commonly set to $1/n$, and p_i is a modified precision score. BP is a penalty for short system translations, defined as

$$\mathrm{BP} \;=\begin{cases} 1 & \text{if } c \;>\; r, \\ e^{(1-r)/c} & \text{if } c \;\le\; r. \end{cases}$$

Here, *c* is the total length of the automatically translated corpus, and *r* is the effective reference length of the test corpus.

3. Results

First, STT quality evaluation of models trained using different amounts of additional data was performed on the Tilde Balss S2MT test set. The results are presented in Table 1 and Figure 5.

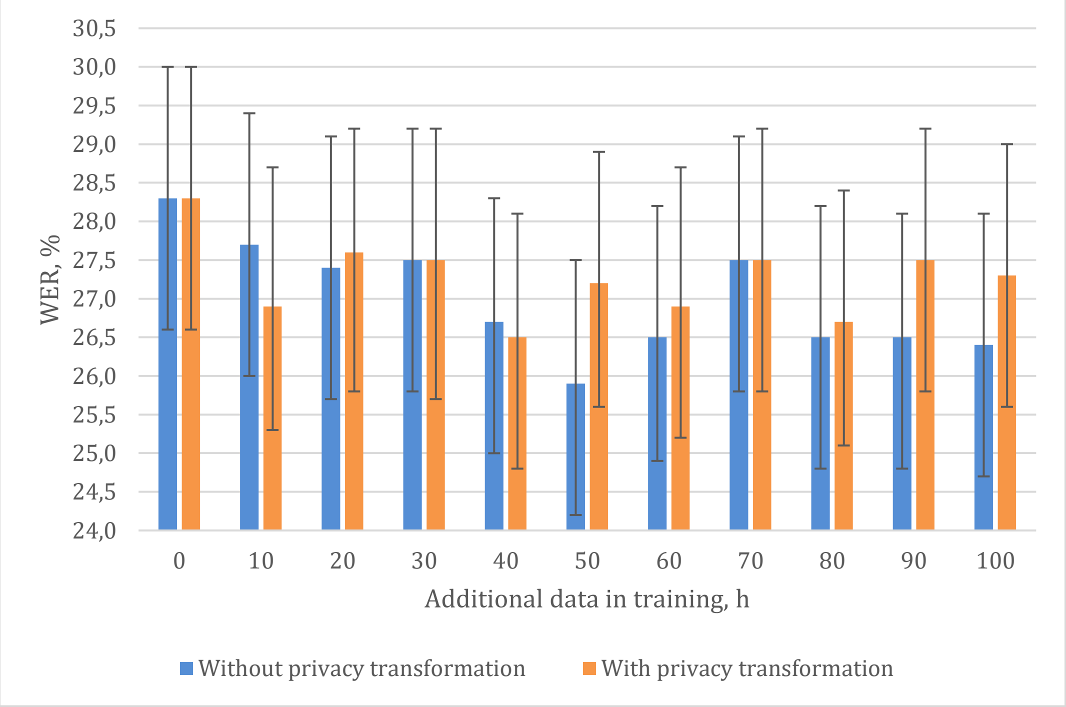

Figure 5: Plot of the values in Table 1 for illustration purposes.

Table 1: WER achieved with untransformed vs. transformed additional data on the Tilde Balss S2MT test set.

Additional speech data (hours)	Without privacy transformation (WER %)	With privacy transformation (WER %)
0	28.3 (26.6-30.0)	28.3 (26.6-30.0)
10	27.7 (26.0-29.4)	26.9 (25.3-28.9)
20	27.4 (25.7-29.1)	27.6 (25.8-29.2)
30	27.5 (25.8-29.2)	27.5 (25.7-29.2)
40	26.7 (25.0-28.3)	26.5 (24.8-28.1)
50	25.9 (24.2-27.5)	27.2 (25.6-28.9)
60	26.5 (24.9-28.2)	26.9 (25.2-28.7)
70	27.5 (25.8-29.1)	27.5 (25.8-29.2)
80	26.5 (24.8-28.2)	26.7 (25.1-28.4)
90	26.5 (24.8-28.1)	27.5 (25.8-29.2)
100	26.4 (24.7-28.1)	27.3 (25.6-29.0)

The results are quite noisy which may be attributed to a mismatch between the domain of the original training set and the additional data. Still, it is possible to make three main observations:

- Additional data improves speech recognition quality.
- Adding untransformed data helps to achieve better WER.
- The difference between adding untransformed and privacy-transformed data is small (2% relative between the best results of both methods).

To address the issue of high noise in the first evaluation, we also calculated the WER on an in-domain test set which is from the same domain as additional speech data. The results are presented in Table 2 and Figure 6.

Table 2: WER achieved with untransformed vs. transformed additional data on the in-domain test set.

Additional speech data (hours)	Without privacy transformation (WER %)	With privacy transformation (WER %)
0	13.6 (12.4-14.7)	13.6 (12.4-14.7)
10	12.8 (11.7-14.0)	12.7 (11.5-13.8)
20	12.5 (11.4-13.7)	13.3 (12.2-14.3)
30	12.5 (11.4-13.6)	13.1 (12.0-14.3)
40	12.4 (11.3-13.5)	13.1 (11.9-14.2)
50	11.9 (10.9-13.0)	13.0 (11.9-14.1)
60	11.9 (10.8-12.9)	13.0 (11.9-14.2)
70	12.1 (11.0-13.1)	12.6 (11.5-13.7)
80	12.2 (11.1-13.2)	12.7 (11.6-13.9)
90	11.8 (10.8-12.8)	12.4 (11.4-13.5)
100	11.8 (10.8-12.8)	12.8 (11.7-14.0)

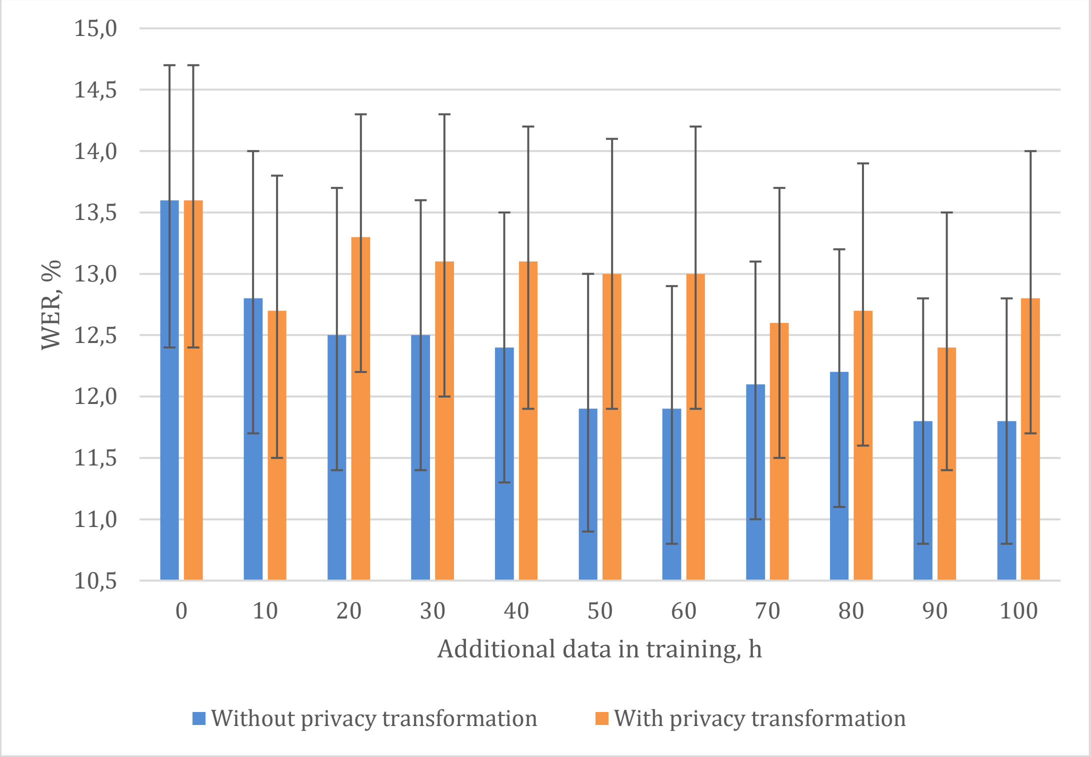

Figure 6: Plot of the values of Table 2 for illustration purposes.

This time, significantly less noise is observed while the conclusions from the previous experiment are confirmed: both types of additional data improve the WER and the difference between adding transformed and untransformed data is small (5% relative between the best results of both methods).

There is a noticeable WER improvement after adding the first 10 hours of privacy-transformed data which seems suspicious. Interestingly, adding the same 10 h of untransformed data only has a similar effect when evaluating on in-domain test data, but not on the general test set. As an additional experiment, we decided to take the last 10 h of additional data instead of first 10 h and to retrain the system. This time the WER improvement was smaller and fitted with other results. Therefore, we believe this result can be explained by irregularities in the additional data, some subsets of which are more beneficial than others.

Finally, an MT evaluation was performed by asking each system to transcribe the Tilde Balss S2MT test set and then translating the output with two different MT models, as described in Section 2. The results are presented in Table 3.

Table 3: Results of the MT evaluation.

Additional speech data (hours)	Without privacy transformation (BLEU)		With privacy transformation (BLEU)	
	Baseline MT	Adapted MT	Baseline MT	Adapted MT
0	**11.0**	12.8	**11.0**	12.8
10	10.6	12.7	10.4	12.4
20	10.9	12.5	10.8	**13.0**
30	10.8	12.7	10.9	12.7
40	10.7	12.5	10.8	12.5
50	10.9	12.8	10.9	12.8
60	10.8	**13.1**	10.8	12.9
70	10.6	12.6	10.7	12.6
80	10.7	12.8	10.6	12.8
90	10.6	12.7	10.7	12.6
100	10.8	12.3	10.5	12.7

Similarly to the evaluation of the WER on the same test set, the results are noisy. We do not find a significant impact on the BLEU score when comparing the two STT setups. We hypothesise that this is due to the fact that the relative differences in WER between the two setups were too small to matter with respect to BLEU. Still, the experiment allows us to draw some conclusions:

- The adapted MT model outperforms baseline MT on STT output translation by almost two BLEU points.
- While we have seen that using original untransformed data in STT training allows us to achieve better WER, the difference is too small to affect the translation as no improvement in BLEU is observed.

4. Conclusion

This document presented the first combined evaluation of the COMPRISE project. Although the project and the innovations developed therein have not reached completion yet, we were eager to test our ideas early on and especially with respect to their interdependencies. Therefore, we devised a first evaluation experiment that combines some of the already available components of COMPRISE, albeit in their first iterations. This ambitious experiment allowed us to compare various aspects of the project, including the effect of one kind of privacy-preserving speech transformation on subsequent STT and MT performance, as well as to quantitatively study the impact of different amounts of data added during the training.

We found that using in-domain data resulted in a clear benefit for STT quality. With a test set from another domain, the benefits of adding more training data suffered from noise artifacts. Generally speaking, however, we did not observe a dramatic decline in WER when applying the voice transformation prior to training.

A similar effect can be observed for MT where both the untransformed and transformed data additions lead to similar outcomes in terms of BLEU score. However, they were observed at different amounts of added data, and in both cases not at the full 100-hour mark. This interesting effect deserves further study. We also found that the MT technology developed specifically for spoken language outperformed the baseline solely trained on written language.

Even at this early stage, the evaluation has now given us important insights. The findings of the experiments will be directly relayed to Work Packages 2, 3, and 4. With a repeatable, combined evaluation setup in place, we now have the means to optimise COMPRISE tools not just for stand-alone performance, but also as parts of a more complex processing chain that introduces additional challenges.

As COMPRISE develops, more evaluation opportunities will become available. Therefore, we will update this deliverable in six months (M21) with newer results. Depending on the exact progress made until then, this may involve re-running the evaluation described above as well as adding additional evaluation conditions.

5. Bibliography

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., & Turchi, M. (2017). Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers* (pp. 169–214).

Cohen, J., Kamm, T., & Andreou, A. G. (1995). Vocal tract normalization in speech recognition: Compensating for systematic speaker variability. *The Journal of the Acoustical Society of America*, *97*(5), pp. 3246-3247.

Eide, E., & Gish, H. (1996). A parametric approach to vocal tract length normalization. In *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference* (Vol. 1, pp. 346-348).

Hadian, H., Sameti, H., Povey, D., & Khudanpur, S. (2018). End-to-end speech recognition using lattice-free MMI. In *Proceedings of Interspeech* (pp. 12-16).

Papineni K., Roukos S., Ward T., and Zhu W.-J. (2002) BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL* (pp. 311-318).

Paul, D. B., & Baker, J. M. (1992). The design for the Wall Street Journal-based CSR corpus. In *Proceedings of the workshop on Speech and Natural Language* (pp. 357-362).

Pinnis, M., Salimbajevs, A., & Auziņa, I. (2016). Designing a speech corpus for the development and evaluation of dictation systems in Latvian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (pp. 775-780).

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., . . . Vesely, K. (2011). The Kaldi speech recognition toolkit. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 1–4).

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., & Khudanpur, S. (2018). Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Proceedings of Interspeech* (pp. 3743-3747).

Salimbajevs, A. (2018). Creating Lithuanian and Latvian speech corpora from inaccurately annotated web data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (pp. 2871-2875)

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the ACL* (pp. 1715–1725).

Smit, P., Virpioja, S., & Kurimo, M. (2017). Improved subword modeling for WFST-based speech recognition. In *Proceedings of Interspeech* (pp. 2551-2555).

Qian, J., Du, H., Hou, J., Chen, L., Jung, T., Li, X. Y., . . . & Deng, Y. (2017). Voicemask: Anonymize and sanitize voice input on mobile devices. *arXiv preprint arXiv:1711.11460*.

Qian, J., Du, H., Hou, J., Chen, L., Jung, T., & Li, X. Y. (2018). Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems* (pp. 82-94).