Speech-to-Speech Translation Pipelines
for Conversations in Low-Resource Languages

Andrei Popescu-Belis[1,3], Alexis Allemann[1], Teo Ferrari[1], Gopal Krishnamani[2]

[1]HEIG-VD / HES-SO, 1401 Yverdon-les-Bains, Switzerland
[2]Bhaasha Sàrl, 1400 Yverdon-les-Bains, Switzerland
[3]EPFL, 1015 Lausanne, Switzerland
Correspondence: andrei.popescu-belis@heig-vd.ch

Abstract

The popularity of automatic speech-to-speech translation for human conversations is growing, but the quality varies significantly depending on the language pair. In a context of community interpreting for low-resource languages, namely Turkish and Pashto to/from French, we collected fine-tuning and testing data, and compared systems using several automatic metrics (BLEU, COMET, and BLASER) and human assessments. The pipelines included automatic speech recognition, machine translation, and speech synthesis, with local models and cloud-based commercial ones. Some components have been fine-tuned on our data. We evaluated over 60 pipelines and determined the best one for each direction. We also found that the ranks of components are generally independent of the rest of the pipeline.

1 Introduction

One of the most challenging applications of spoken language translation is real-time interpreting of human conversations. We consider the application to community interpreting, for ethnic minorities who need assistance to access services across a language barrier, e.g., for healthcare, asylum rights, or education. The case study presented here involves Bhaasha, a company that provides services for community interpreting, and the Data Science group of HEIG-VD, an academic partner. Due to a growing demand, the company aims to clarify whether a system for automated interpreting meets certain quality thresholds and can be offered when human interpreters are not available. While several online offers exist, these systems do not include the desired language pairs, or their quality is clearly insufficient, and privacy is not guaranteed.

We present the methods and the results of a joint project aimed at determining the best speech-to-speech translation pipeline made from off-the-shelf components, cloud-based services, or fine-tuned models, for two language pairs that are in high demand, but are insufficiently supported by existing systems: French-Turkish and French-Pashto. The paper is organized as follows. In Section 3, we present methods for collecting and annotating data representative of the intended context of use. In Section 4, we outline the design of translation pipelines, whose components can be smoothly interchanged. In Section 5, we evaluate all combinations of four ASR, three MT, and two TTS components, either local or cloud-based, also including two fine-tuned components and a speech-to-text translation one. We present evaluation scores from automatic metrics and determine the best combination of components per direction. We also show that the ranking of components is generally independent of the other modules in a pipeline. Finally, we present human scores over a subset of the data, showing that accuracy, fluency and intonation of the best pipelines are considered as 'good' or 'very good'.

2 State of the Art

Methods for speech-to-*text* translation (e.g. for subtitling) have been the subject of many recent publications, unlike methods for speech-to-*speech* translation, as the speech synthesis part is difficult to train. Moreover, spoken translation has been studied more often for monologues than for conversations. The three necessary components are automatic speech recognition (ASR, or speech-to-text, STT), machine translation (MT), and speech synthesis (or text-to-speech, TTS).

Research interests, however, have shifted from loosely coupled cascades of ASR and MT, to tighter coupling, and finally to recent end-to-end models

(Sperber and Paulik, 2020; Xu et al., 2023). For instance, an approach to multilingual speech-to-text translation through efficient transfer learning from a pretrained speech encoder (wav2vec) and text decoder (BERT), is proposed by Li et al. (2021). Dong et al. (2021) propose a Listen, Understand and Translate (LUT) approach to train end-to-end speech-to-text translation.

Bentivogli et al. (2021) compare the two paradigms – cascaded vs. end-to-end – and claim that the gap between them is almost closed. However, for low-resource languages, end-to-end systems are difficult to train due to the lack of data, while cascaded systems can use components trained with simpler tasks. Alternatively, massively multilingual systems such as Whisper (Radford et al., 2022) for ASR + MT claim that low-resource languages are improved thanks to higher-resource similar languages. For instance, cascaded approaches can take advantage of optimized low-resource MT components (Atrio and Popescu-Belis, 2022).

The IWSLT 2022, 2023 and 2024 evaluation campaigns (Anastasopoulos et al., 2022; Agarwal et al., 2023; Ahmad et al., 2024) featured various shared tasks, including speech-to-text and speech-to-speech translation for low-resource languages. A typical low-resource system presented at IWSLT 2023 is the Marathi to Hindi submission by Kesiraju et al. (2023a), including an end-to-end and a cascaded system. Various techniques for improving low-resource speech-to-text translation, in particular with initialization from a multilingual ASR system, have been proposed (Khosravani et al., 2021; Fu et al., 2023; Kesiraju et al., 2023b).

Large corpora exist for well-resourced language pairs, but not for the low-resource ones that we target. While datasets of recorded speech can be more easily found, datasets with transcriptions and translations are scarce or inexistent. The MuST-C (Di Gangi et al., 2019) and Multilingual TEDx corpora include speech and translation in English and 8 other European languages, but not Turkish or Pastho. CoVoST-2 (Wang et al., 2021) covers speech translation from several languages to/from English and includes Turkish. This resource was used to test the Whisper ASR + MT used here.

Recent developments of generative AI and large language models have enabled significant progress in speech translation and synthesis, but low-resource languages are still insufficiently supported. For instance, several companies advertise multilingual speech translation systems on the Web, as apps for smartphones, or as cloud-based services, mostly for for a few well-resourced languages. For instance, one of the major players, Google, offers the three components individually via APIs, but also bundles them into a pipeline that often appears in informal tests as one of the best translation apps for several language pairs. Other commercial offers include DeepL, Microsoft's Bing Translator, iTranslate, SayHi, Translate Now, Yandex, or Talking Translator. Many of the related apps have received reviews from their users, which provide a form of evaluation, although ratings for specific language pairs are rarely found. In our tests, we observed that these solutions are not ready for the low-resource languages studied here, nor for use in the setting of community interpreting.

3 Data Gathering and Formatting

To the best of our knowledge, there are no datasets with parallel conversational speech (i.e. interpreted in both directions) for Turkish-French and Pashto-French (tr-fr, ps-fr). Therefore, we collected new data which suits our project's needs.

The central idea of our parallel dataset is to include complete dialogues in situations encountered by Bhaasha's community interpreters. For each utterance, we have a reference transcript and an audio recording, in each of the three languages of the project: an excerpt is shown in Table 1. For each utterance, in each language, the dataset contains indexing information (dialogue codename, utterance number, and language), the transcript of the utterance, the name of the audio file with the utterance (similarly indexed), and whether it is used in the fine-tuning or the testing subsets. With this structure, the dataset can be used to fine-tune or to test speech translation pipelines in any translation direction.

3.1 Data Sources

Collecting such a dataset requires an abstraction over the complex reality of community interpreting, which involves three speakers: the two persons between whom the dialogue takes place, and the interpreter, who interprets consecutively the speech in both directions. However, it appeared early in the project that real dialogues mediated by interpreters *could not be recorded due to privacy reasons.*

Therefore, for most of our data, we settled on the following protocol. We gathered or wrote dialogues

dial.	utt.	lang	audio	transcription
B001	1	fr	B001-1-fr.wav	Bonjour Monsieur, qu'est-ce qui vous amène ?
B001	1	ps	B001-1-ps.wav	سلام ښاغلی، تاسو دلته څنګه راغلي یاست؟
B001	1	tr	B001-1-tr.wav	Merhabalar beyefendi, bugün neden buradasınız?
B001	2	fr	B001-2-fr.wav	J'ai mal à la tête, très mal depuis déjà plus de deux semaines.
B001	2	ps	B001-2-ps.wav	زه له دوو اونیو راهیسې دیر بد سر درد لرم
B001	2	tr	B001-2-tr.wav	2 haftadan fazla başım ağrıyor.
B001	3	fr	B001-3-fr.wav	Qu'est-ce que vous prenez pour calmer ces douleurs ?
B001	3	ps	B001-3-ps.wav	تاسو د دې درد څخه د خلاصون لپاره څه اخلي؟
B001	3	tr	B001-3-tr.wav	Ağrılar geçsin diye ne alıyorsunuz?

Table 1: Three utterances from our dataset: each one is available in three languages and two modalities.

similar to those handled by Bhaasha interpreters, writing them in French, in some cases with the help of the GPT-4 LLM. Then, we asked interpreters from Bhaasha to write translations of the entire dialogues into Turkish or Pachto, by postediting automatic translations from the Google or Microsoft online systems. Finally, we recorded interpreters reading aloud these translations, and added French audios read by different native speakers, manually segmenting all audios into utterances.

This text-centric protocol appeared to be much more efficient than an audio-based one in which interpreters listen to a source sentence (or read a sentence) and then utter the spoken translation in the target language, which is recorded and then transcribed. This solution was very demanding for interpreters, and was not entirely natural as it required interpreting both dialogue participants in the same direction. Enacting original new spoken dialogues appeared also to have too high transcription and translation costs. The current dialogues, although more fluent than real ones, are the best substitute that could be found within the frame of our project.

The sources of the dialogues included in our project data are the following ones. Each dialogue is identified by a letter coding the generation method and an index number. Each separate sentence (utterance) appears on one line, and speaker turns can be made of one or more lines (as indicated in the metadata, see 3.2).

- 'G' series (G001–G013, 630 lines): dialogues generated with GPT-4.[1] This was the quickest technique and provided about half of our data.

Given a precise prompt in French,[2] GPT-4 generated a realistic in-domain dialogue of the desired size and style, which was improved to satisfactory levels with minimal human edits from the experimenters.

- 'C' series (C001–C004, 498 lines): excerpts from the CoVoST-2 corpus (Wang et al., 2021), Turkish-English parallel subset, mostly with spoken news (admittedly, not dialogues). We translated the data into French by postediting MT output, and added French audio from native speakers.

- 'B' series (B001–B006, 180 lines): dialogues created as French text by interpreters from Bhaasha, in the spirit of those that they encounter as community interpreters.

- 'P' series (P001–P004, 110 lines): four samples of learners' material in French, corresponding to our style and topics.

	Utterances		Durations		
	train	test	sec.	min.	words
fr	722	723	4,138	69	11,986
tr	722	723	4,985	83	8,295
ps	0	400	1,741	29	3,717
total	3,290		10,864	181	23,998

Table 2: Data used for fine-tuning and testing.

As summarized in Table 2, our dataset includes 28 dialogues with 1,445 utterances (lines). All utterances are available in French and Turkish, with

---

[1] https://chat.openai.com

[2] Prompts describe in detail a situation, matching closely those encountered by interpreters, e.g., "Write a dialogue at the welfare office with this topic: a young man has found a part-time job (50%) and wants to know what impact this will have on his welfare. He will have a long and costly commute. Generate 30 to 40 turns."

transcript and audio, but Pashto translation is partial, from lack of availability of Pashto interpreters. We randomly sampled 723 lines for testing and 722 for fine-tuning from the French side, and similarly from the Turkish side. We did not sample entire dialogues, to ensure better similarity between fine-tuning and testing data. For Pashto, all 400 lines were used for testing.

3.2 Exchange Format

The exchange format is kept simple, to ensure easy reuse. The dataset is contained in two text-based files and one folder with audio files:

- `dialogues.json` – metadata in JSON format, described below.
- `dataset.csv` – indexed transcripts and names of audio files, as shown in Table 1.
- `audios` – contains one audio file per utterance, named using indexes, from recordings on smartphones or laptops in silent environments (2 channels, 48 kHz, 32 bits).

As metadata, we include for each dialogue identified by its codename: long name or brief description, creation method (including prompt to GPT-4 for the G series), date of recording, and number of utterances. For each language, we indicate whether it is an original or translated version, how it was translated, and the total duration of audios. Finally, we indicate the grouping of utterances in speaker turns using their index numbers.

4 Speech-to-Speech Translation Pipelines

4.1 Components

We considered all possible combinations of the following ASR, MT and TTS components. Table 3 below provides the exact names and URLs of all of them. We evaluated ASR/MT/TTS cloud-based commercial components from Google and Microsoft, as well as the following open-weight models run locally. For ASR, we tested Whisper from OpenAI (Radford et al., 2022) in 'transcribe' mode, i.e. in the same language, and MMS from Meta (Pratap et al., 2023). For MT, we tested the multilingual NLLB-200 model with 3.3B parameters (NLLB Team et al., 2022). But for TTS, no competitive local model could be found for our languages. Moreover, we fine-tuned Whisper and NLLB-200 with 1.3B parameters on the training subset of the fr-tr data (see Table 2), resulting in the models prefixed with 'ft' below.

4.2 Architecture

We built a flexible application to support experimentation, but also real-time demonstration. Hence, the application includes a frontend and a backend, and is hosted on the Kubernetes infrastructure of the Swiss AI Center[3] with S3 MinIO storage. The dataset is managed using DVC.

The frontend of the application is developed using the React framework, while the backend is built in Python with FastAPI, providing several HTTP endpoints to enable the use of different versions of the ASR, MT, and ST modules. The frontend orchestrates the sequence of calls across the various stages of the speech-to-speech translation pipeline. These endpoints allow responses to be generated using either local models running on GPUs or remote models accessed via third-party APIs. Additionally, for every request, the backend stores copies of the audio and model outputs at each stage in a S3 bucket, which facilitates analysis and human evaluation.

A frontend interface allows human users to inspect or demonstrate the system. The interface enables on-the-fly change of components in the pipeline, depending on the desired source and target languages. A laptop with a regular microphone can be used for demos.

4.3 Evaluation Metrics

We use Word Error Rate (WER) to score ASR components, with the JiWER Python package.[4] We use four automatic metrics for MT: three of them, available from the Sacrebleu library (Post, 2018),[5] use various form of edit distance between candidate and reference translations: BLEU, ChrF, and Translation Error Rate (TER). The fourth metric, COMET (Rei et al., 2022)[6], compares source and target embeddings using a large language model (wmt22-comet-da), and is applicable to French-Turkish as well as French-Pashto. We found that there is a strong correlations between these metrics: using each system as a data point, average pairwise Pearson correlation is 0.89 for fr-tr and 0.97 for fr-ps, with four metrics. Therefore, we use below two representative and least correlated metrics, namely BLEU and COMET.

Stage	Type	Name	URL
	cloud	Google STT	https://cloud.google.com/speech-to-text/v2
	cloud	Microsoft STT	https://speech.microsoft.com/portal
ASR	local	OpenAI Whisper-large-v3	https://huggingface.co/openai/whisper-large-v3
	local	Fine-tuned Whisper	https://huggingface.co/openai/whisper-large-v3
	local	Meta MMS	https://huggingface.co/facebook/mms-1b-all
ASR+MT	*local*	*Whisper Translate*	https://huggingface.co/openai/whisper-large-v3
	cloud	Google MT	https://cloud.google.com/translate
	cloud	Microsoft MT	https://www.microsoft.com/en-us/translator
MT	local	NLLB-200 3.3B	https://huggingface.co/facebook/nllb-200-3.3B
	local	Fine-tuned NLLB (1.3B)	https://huggingface.co/facebook/nllb-200-1.3B
	local	*HelsinkiNLP*	https://huggingface.co/Helsinki-NLP
	cloud	Google TTS	https://cloud.google.com/text-to-speech
TTS	cloud	Microsoft TTS	https://speech.microsoft.com/portal
	local	*YourTTS*	https://github.com/Edresson/YourTTS

Table 3: Components used in our speech-to-speech translation pipelines (in italics, preliminary studies only).

5 Results

5.1 ASR Scores (WER)

The WER scores for the ASR components are given in Table 4 (lower is better). The rankings are consistent across French and Turkish, although the differences between systems are not. For Turkish, the fine-tuning of Whisper on our data brings a visible improvement (from 0.14 to 0.09), while the untuned Whisper performs on par with the Microsoft cloud-based service. The Google service and the Meta local model follow at some distance.

French ASR	WER
ft_whisper_transcribe	**0.04**
whisper_transcribe	0.06
microsoft_stt	0.08
google_stt	0.23
meta_mms	0.24
Turkish ASR	
ft_whisper_transcribe	**0.09**
whisper_transcribe	0.14
microsoft_stt	0.15
google_stt	0.31
meta_mms	0.40
Pashto ASR	
microsoft_stt	**0.45**
google_stt	0.89
whisper_transcribe	0.92

Table 4: WER for French, Turkish, and Pashto.

5.2 MT Scores (BLEU and COMET)

The scores of written MT for Turkish and French (both directions) for all combinations of modules, with four metrics, are shown in Table 8 in the Appendix. Similarly, the MT scores for Pashto and French are shown in Table 9. These tables indicate the best ASR+MT pipelines, with substantial agreement between metrics:

- **tr-fr:** fine-tuned Whisper (or not fine-tuned) with Google MT (or Microsoft MT).

- **fr-tr:** fine-tuned Whisper (or not fine-tuned) with Microsoft MT (or Google MT).

- **ps-fr:** Microsoft ASR with Microsft MT (or Microsoft MT).

- **fr-ps:** fine-tuned Whisper (or not fine-tuned) with Google MT.

To perform a systematic analysis of the intrinsic quality of each module and of the effects of their combinations, we propose the following approach, applied to each translation direction.

5.2.1 Turkish and French

We first present a detailed analysis of **Turkish →French** pipelines, and then summarize conclusions for the other direction, and then for Pashto and French. We organize the COMET scores in two ways. First, as shown in Table 5, pipelines are grouped by MT systems, and the groups are ranked by average COMET. Inside each group, ASR components are ranked too. We find that the ranking of ASR is the same inside the first and second best

ASR	MT	COMET	AVG
whisper	google_mt	**89.60**	87.47
ft_whisper	google_mt	89.40	
microsoft_stt	google_mt	87.87	
google_stt	google_mt	87.28	
meta_mms	google_mt	83.20	
whisper	microsoft_mt	88.42	86.04
ft_whisper	microsoft_mt	88.06	
microsoft_stt	microsoft_mt	86.64	
google_stt	microsoft_mt	86.49	
meta_mms	microsoft_mt	80.61	
whisper	ft_nllb-1.3B	86.14	83.86
ft_whisper	ft_nllb-1.3B	85.30	
microsoft_stt	ft_nllb-1.3B	84.21	
google_stt	ft_nllb-1.3B	84.21	
meta_mms	ft_nllb-1.3B	79.42	
whisper	nllb-3.3B	85.92	83.62
ft_whisper	nllb-3.3B	84.99	
microsoft_stt	nllb-3.3B	84.02	
google_stt	nllb-3.3B	83.81	
meta_mms	nllb-3.3B	79.38	

Table 5: COMET scores for Turkish-to-French speech-to-text translation, grouped by MT system, and ranked by average COMET over each group.

ASR	MT	COMET	AVG
whisper	google_mt	**89.60**	87.52
whisper	microsoft_mt	88.42	
whisper	ft_nllb-1.3B	86.14	
whisper	nllb-3.3B	85.92	
ft_whisper	google_mt	89.40	86.94
ft_whisper	microsoft_mt	88.06	
ft_whisper	ft_nllb-1.3B	85.30	
ft_whisper	nllb-3.3B	84.99	
microsoft_stt	google_mt	87.87	85.69
microsoft_stt	microsoft_mt	86.64	
microsoft_stt	ft_nllb-1.3B	84.21	
microsoft_stt	nllb-3.3B	84.02	
google_stt	google_mt	87.28	85.45
google_stt	ft_nllb-1.3B	86.49	
google_stt	microsoft_mt	84.21	
google_stt	nllb-3.3B	83.81	
meta_mms	google_mt	83.20	80.65
meta_mms	microsoft_mt	80.61	
meta_mms	ft_nllb-1.3B	79.42	
meta_mms	nllb-3.3B	79.38	

Table 6: COMET scores for Turkish-to-French speech-to-text translation, grouped by ASR system, and ranked by average BLEU over each group.

groups, which are those with Google MT and Microsoft MT (with a large difference between them). The ranking of the first two ASR systems is permuted when we move to the third and fourth groups. Therefore, the following stable ranking is found for Turkish ASR. Fine-tuning Whisper turns out to be beneficial to BLEU scores but not to COMET ones.

```
Whisper > Fine-tuned Whisper >
Microsoft ASR > Google ASR >
Meta MMS ASR
```

Second, as shown in Table 6, pipelines are grouped by ASR system, and the groups are ranked by average COMET; inside each group, MT components are ranked too. We find that the ranking of MT is almost always the following one (except in one group where the second and third ranks are permuted):

```
Google MT > Microsoft MT > Fine-
tuned NLLB 1.3B > NLLB 3.3B
```

For the **French → Turkish pipelines**, a similar analysis shows that the stable ranking of ASR components in each grouping based on MT is the following one (the ranking of the last two components is reversed in half of the groups):

```
Fine-tuned Whisper > Whisper >
Microsoft ASR > Google ASR ≈
Meta MMS ASR
```

Conversely, when grouping by ASR, fine-tuned Whisper ahead of the others in BLEU score, but the untuned Whisper is slightly ahead on COMET. They are followed by Microsoft ASR, and then at some distance by Meta MMS and Google ASR, which are quit close. When grouping by ASR, the stable ranking of MT components is the following one, with some uncertainty over the fine-tuned NLLB, and a reversal of the first two ranks with COMET:

```
Microsoft MT > Google MT >
Fine-tuned NLLB 1.3B > NLLB 3.3B
```

5.2.2 Pashto and French

Similar to the above strategy, the scores for the **Pashto → French** pipelines are either grouped by MT systems to observe the rankings of ASR in each group, or grouped by ASR systems to observe the rankings of MT. The actual scores are shown in Table 9 in the Appendix. We make the following observations using COMET scores. When grouping by MT system (Google or Microsoft), the rank-

23

ing of ASR is always the same: `Microsoft ASR > Whisper > Google ASR`. In terms of the actual average score per ASR, Microsoft ASR is much better than Whisper or Google ASR, which do not seem usable here. When grouping by ASR, the ranking of MT is the same for the first two ASR systems, but is reversed for the last one, likely due to the poor quality of input to MT: `Google MT > Microsoft MT`. In terms of average per MT, Google MT is also slightly ahead of Microsoft MT, as in the observations grouped per ASR.

Finally, for the **French → Pashto** pipelines, when grouping by MT, the rankings of ASR differ, although the best system is the Fine-tuned Whisper in both cases. For Google MT, Whisper is the second best, although it is ranked fourth when using Microsoft MT. However, given the poor quality of this last MT system, the rankings may not be reliable. In terms of average per ASR, the Fine-tuned Whisper is first, followed by Whisper and by Microsoft ASR, and then by Meta MMS and Google ASR. (As they concern French ASR, these rankings are similar to those for fr-tr.) When grouping per ASR, the ranking of MT is always the same: `Google MT > Microsoft MT`, with large differences between the two (8–9 COMET points).

5.3 End-to-end Scores (BLASER)

The BLASER 2.0 scores (Dale and Costa-jussà, 2024) of the speech-to-speech translation pipelines are given in Table 7. They were computed for Turkish and French, as no models are available for Pashto. We selected two representative ASR + MT pipelines: Whisper + NLLB is entirely local and not fine-tuned, while Google + Google is the commercial cloud-based offer from Google. We combined each of them with two cloud-based speech synthesis solutions, respectively from Google and Microsoft, as no local TTS was satisfactory, We computed BlaserQE and BlaserRef scores for each of the four pipelines. For each sentence, BlaserQE compares the embeddings of the *source* and of the *candidate* translation in the audio modality, while BlaserRef also considers the embedding of the *written reference* translation.

The BLASER 2.0 scores indicate that using Google TTS is *always* slightly better than Microsoft TTS. The difference between these systems for fr-tr is statistically significant at the 1% level (as measured by a t-test) with the BlaserRef metric, regardless of the ASR + MT part. As for tr-fr, the difference is significant at the 1% level (t-test) only

ASR+MT	Whisper+NLLB		Google+Google	
TTS	Google	MS	Google	MS
fr-tr systems				
WER	0.06		0.23	
BLEU	25.76		22.72	
COMET	89.37		87.49	
BlaserQE	3.07	3.02	3.06	3.01
BlaserRef	3.24	3.16	3.25	3.18
Meaning	4.43		4.27	
Correctness	4.55		4.69	
Intonation	4.55	4.55	4.50	4.57
tr-fr systems				
WER	0.14		0.31	
BLEU	38.46		43.43	
COMET	85.92		87.26	
BlaserQE	3.18	3.16	3.28	3.26
BlaserRef	3.19	3.18	3.34	3.32

Table 7: Results of automatic evaluation with BLASER 2.0 and of human evaluation of speech-to-speech translation. For comparison purposes, we reproduce the WER, BLEU and COMET scores. MS stands for the Microsoft speech synthesis component.

when combined with Google ASR + MT. Moreover, the Google-only pipeline scores significantly better than both local ones.

5.4 Human Evaluation

As a pilot experiment, we showed 21 utterances to two human judges, native speakers of Turkish, one of them being an interpreter. We presented them with source audio and translations from French to Turkish by the same four pipelines as in the previous section. For each utterance, they were asked to grade three aspects: (1) how well the original meaning is communicated by the translation; (2) how correct is the wording of the translation; and (3) how good is the intonation of the translation. The first two aspects are akin to the traditional *adequacy* and *fluency* dimensions, but here no transcript is seen by evaluators. The third one is aimed specifically at speech synthesis. To speed up evaluation, when the ASR + MT pipeline is the same but the TTS is different, we ask evaluators to rate only once the meaning and correctness, and to rate separately the two different TTS outputs. At the top of the interface, which includes links to the audios and a drop-down menu for each rating, we briefly defined each aspect. The possible values for ratings are the following ones (originally in French):

- Meaning: (1) not at all; (2) the general idea; (3) some elements; (4) almost entirely; (5) entirely.

- Correctness: (1) very incorrect; (2) quite incorrect; (3) medium; (4) quite correct; (5) very correct.

- Intonation: (1) not understandable; (2) a little understandable; (3) medium; (4) well understandable; (5) perfectly understandable.

Average ratings for each aspect by the two judges are given in Table 7. The estimated quality by the human judge is overall between 4 and 5 for all aspects and systems. Communicated meaning is scored around 4, i.e. 'almost entirely', which is the lowest of the three scores, likely due to the combination of errors from ASR and MT. Grammatical correctness, depending almost exclusively on the ASR + MT pipeline, is also between 'quite correct' and 'very correct', here with a slight advantage to the Google components (4.9 vs. 4.5). This could be due to NLLB being a multilingual MT system, which has a lower fluency for Turkish than the Google's dedicated system. Intonation, either generated by Google TTS or by Microsoft TTS, scores close to 4.5, i.e. between 'well' and 'perfectly understandable'. There is no significant difference between the two systems, despite a slightly higher BLASER score for Google TTS. The human ratings give an idea of the calibration of automatic metrics, with BLEU scores of around 25 and COMET scores of nearly 90 being already perceived as good quality.

6 Conclusion

We have produced data and assembled numerous speech-to-speech translation pipelines, for interpreting Turkish ↔ French and Pashto ↔ French conversations. Specifically, we have produced three hours of data in settings compatible with community interpreting, and used half of it for fine-tuning two Turkish ↔ French ASR and MT systems, and the other half for evaluation. We scored over 60 pipelines of ASR, MT and TTS systems, either based on open-weight models run locally, or on commercial cloud-based services. We identified the best-performing pipeline in each direction, and found that the ranking of components was consistent, regardless of the other components of pipelines. We used four automatic evaluation metrics (WER, BLEU, COMET and BLASER),

along with pilot human evaluations. The implementation of an online system with a push-to-talk interface, along with an offline version allowing batch processing, now paves the way towards usability testing of automatic interpretation solutions, which will also need to take into consideration factors such as privacy, cost, and deployment strategy.

References

Milind Agarwal et al. 2023. Findings of the IWSLT 2023 evaluation campaign. In *Proceedings of the 20th Int. Conf. on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada. Association for Computational Linguistics.

Ibrahim Said Ahmad et al. 2024. Findings of the IWSLT 2024 evaluation campaign. In *Proceedings of the 21st Int. Conf. on Spoken Language Translation (IWSLT 2024)*, pages 1–11, Bangkok, Thailand. Association for Computational Linguistics.

Antonios Anastasopoulos et al. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th Int. Conf. on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland. Association for Computational Linguistics.

Àlex R. Atrio and Andrei Popescu-Belis. 2022. On the interaction of regularization factors in low-resource neural machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 111–120, Ghent, Belgium. European Association for Machine Translation.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus direct speech translation: Do the differences still make a difference? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2873–2887, Online. Association for Computational Linguistics.

David Dale and Marta R. Costa-jussà. 2024. BLASER 2.0: a metric for evaluation and quality estimation of massively multilingual speech and text translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16075–16085, Miami, Florida, USA. Association for Computational Linguistics.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Qianqian Dong, Rong Ye, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2021. Listen, understand and translate: Triple supervision decouples end-to-end speech-to-text translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12749–12759.

Yu-Kuan Fu, Liang-Hsuan Tseng, Jiatong Shi, Chen-An Li, Tsu-Yuan Hsu, Shinji Watanabe, and Hung-yi Lee. 2023. Improving cascaded unsupervised speech translation with denoising back-translation. *arXiv:2305.07455*.

Santosh Kesiraju, Karel Beneš, Maksim Tikhonov, and Jan Černocký. 2023a. BUT systems for IWSLT 2023 Marathi - Hindi low resource speech translation task. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 227–234, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Santosh Kesiraju, Marek Sarvaš, Tomáš Pavlíček, Cécile Macaire, and Alejandro Ciuba. 2023b. Strategies for improving low resource speech to text translation relying on pre-trained ASR models. In *Proceedings of Interspeech 2023*, page 2148–2152. ISCA.

Abbas Khosravani, Philip N Garner, and Alexandros Lazaridis. 2021. Learning to translate low-resourced Swiss German dialectal speech into Standard German text. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 817–823. IEEE.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Multilingual speech translation from efficient finetuning of pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 827–838, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv:2207.04672*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *arXiv:2305.13516*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv:2212.04356*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421, Online. Association for Computational Linguistics.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. CoVoST 2 and massively multilingual speech translation. *Proceedings of Interspeech 2021*, pages 2247–2251.

Chen Xu, Rong Ye, Qianqian Dong, Chengqi Zhao, Tom Ko, Mingxuan Wang, Tong Xiao, and Jingbo Zhu. 2023. Recent advances in direct speech-to-text translation. *arXiv 2306.11646*.

A   Appendix

ASR	MT	tr-fr				fr-tr			
		BLEU	ChrF	TER	COMET	BLEU	ChrF	TER	COMET
ft_whisper	ft_nllb-1.3B	39.79	62.13	52.97	85.30	30.77	60.22	54.49	89.80
ft_whisper	google_mt	**55.78**	**72.88**	**37.05**	**89.40**	**36.93**	**65.19**	50.31	**90.95**
ft_whisper	microsoft_mt	*45.38*	*67.66*	*43.53*	*88.06*	**38.11**	**65.51**	**49.40**	*90.20*
ft_whisper	nllb-3.3B	38.06	58.95	54.54	84.99	26.99	57.63	59.98	89.24
google_stt	ft_nllb-1.3B	35.77	58.63	54.62	85.30	22.38	54.37	62.61	87.10
google_stt	google_mt	43.43	66.49	46.32	87.28	22.72	58.69	62.12	87.49
google_stt	microsoft_mt	37.14	63.37	51.04	86.49	26.24	60.55	60.98	88.26
google_stt	nllb-3.3B	34.65	55.96	57.39	83.81	20.06	53.33	65.61	87.24
microsoft_stt	ft_nllb-1.3B	38.46	59.88	54.82	84.21	27.45	57.65	57.95	88.79
microsoft_stt	google_mt	*51.19*	*69.31*	*40.74*	87.87	32.98	62.92	53.77	90.07
microsoft_stt	microsoft_mt	42.67	65.03	47.43	86.64	*34.47*	62.75	53.37	89.37
microsoft_stt	nllb-3.3B	36.48	57.72	55.78	84.02	24.81	55.71	62.66	88.62
meta_mms	ft_nllb-1.3B	32.70	55.93	59.08	79.42	24.88	56.14	59.86	86.81
meta_mms	google_mt	41.92	65.07	49.33	83.20	23.41	59.35	61.41	86.94
meta_mms	microsoft_mt	34.26	60.11	56.54	80.61	28.02	60.42	60.55	87.29
meta_mms	nllb-3.3B	31.15	53.55	61.14	79.38	20.00	53.53	65.78	86.40
whisper	ft_nllb-1.3B	40.16	62.15	52.36	86.14	28.64	58.95	56.03	89.62
whisper	google_mt	**54.27**	**71.57**	**37.52**	**89.60**	34.27	*63.96*	*52.46*	**90.93**
whisper	microsoft_mt	44.97	67.07	44.67	*88.42*	*36.02*	*64.02*	*51.43*	*90.48*
whisper	nllb-3.3B	38.46	59.54	53.69	85.92	25.76	56.65	61.41	89.37

Table 8: MT scores of all tested combinations of modules for Turkish and French (both directions). The two best scores in each column are in **bold** and the next two in *italics*. The pipelines are ordered alphabetically by name of ASR and then of MT.

ASR	MT	ps-fr				fr-ps			
		BLEU	ChrF	TER	COMET	BLEU	ChrF	TER	COMET
ft_whisper	google_mt	-	-	-	-	**64.22**	**76.06**	**29.78**	**87.03**
ft_whisper	microsoft_mt	-	-	-	-	20.49	43.92	69.42	76.69
google_stt	google_mt	4.23	18.38	88.81	54.02	44.16	63.36	42.97	82.21
google_stt	microsoft_mt	6.61	20.69	87.89	54.26	19.61	41.97	72.91	74.30
meta_mms	google_mt	-	-	-	-	42.17	61.40	44.77	81.56
meta_mms	microsoft_mt	-	-	-	-	18.67	40.70	73.16	73.20
microsoft_stt	google_mt	**25.96**	**47.43**	**64.43**	**77.50**	56.37	70.51	36.16	84.30
microsoft_stt	microsoft_mt	*21.36*	*42.87*	*70.84*	*75.13*	19.67	42.83	70.06	75.66
whisper	google_mt	9.11	27.85	90.45	57.21	*56.87*	*71.81*	*35.23*	*85.82*
whisper	microsoft_mt	8.39	27.07	91.60	55.09	19.44	43.27	70.35	76.22

Table 9: MT scores of all tested combinations of modules for Pashto and French (both directions). The best score in each column is in **bold** and the second one in *italics*. The pipelines are ordered alphabetically by name of ASR and then of MT. The ASR system from Meta does not support Pashto, and we did not have enough data to fine-tune Whisper for Pashto.