

End-to-End Evaluation of a Speech-to-Speech Translation System in TC-STAR

Olivier Hamon, Djamel Mostefa, Khalid Choukri

Evaluations and Language Resources Distribution Agency (ELDA)
55-57 rue Brillat-Savarin, 75013 Paris
France
{hamon,mostefa,choukri}@elda.org

Abstract

The paper describes an evaluation methodology to evaluate speech-to-speech translation systems and their results. The evaluation scheme uses questionnaires filled in by human judges for addressing the adequacy and fluency of audio translation outputs and was applied in the second TC-STAR evaluation campaign. The same evaluation methodology is carried out both on the outputs of an automatic system and on human translations produced by professional interpreters. The obtained results show that the professional interpreters obtain better results on fluency. But surprisingly, the automatic systems results on adequacy are higher than for the human translator. But this has to be reconsidered by the fact that humans have to translate in a real time, and select important information, while an automatic system tends to translate all the information.

Introduction

The TC-STAR project¹ is a long-term effort to advance research in the core technologies of Speech-to-Speech Translation (SST). The project targets a selection of speech domains (speeches and broadcast news) and three languages: UK English, European Spanish and Mandarin Chinese. To assess the advances made in all SST technologies, annual competitive evaluations are organized. In the second evaluation campaign of TC-STAR which took place in February and March 2006, an end-to-end evaluation was carried out.

A Speech-to-Speech Translation system is composed of an Automatic Speech Recognizer (ASR) chained to a Spoken Language Translation (SLT) module and to a Text-To-Speech (TTS) component in order to produce the speech in the target language.

Evaluations of individual components (ASR, SLT and TTS) have been done many times in the past and are also evaluated in TC-STAR. Methods, protocols and metrics for the evaluation of independent ASR, SLT and TTS modules have already been established. Performance of speech recognition systems is typically described in terms of word error rate (WER). For SLT automatic metrics such as BLEU (Papineni et al. 2001) or mWER (Nießen et al., 2000) are used. Performance of SLT systems can also be measured by human judges who can usually measure fluency, adequacy, etc. of the translations. A numerical indication of the perceived quality of TTS synthesized audio data is provided by Mean Opinion Scores (ITU-T, 1993). The MOS is generated by averaging the results of a set of standard subjective tests.

To evaluate the performance of a complete speech-to-speech translation system, we need to compare the source speech used as input to the translated output speech in the target language. The proposed methodology enables to measure the *fluency* and the *adequacy* of the translated output.

We first give a quick overview of the TC-STAR evaluation results. Then we describe tasks and languages of the experiments, and the end-to-end evaluation methodology and protocol. Finally, the results obtained by the TC-STAR system and the human interpreter are depicted and analyzed.

Overview of the TC-STAR evaluation results

In this section we introduce the TC-STAR evaluation results regarding the individual component (ASR, SLT, TTS) evaluated within the project. It aims at defining the scope of the end-to-end evaluation through the state-of-the-art of each component evaluation. Well-know evaluations are used and details can be found in (Mostefa et al., 2006).

Three languages are involved by using recordings of the Parliament's sessions in English and Spanish, recordings of Spanish Parliament's sessions in Spanish and broadcast news recordings in Chinese.

Automatic Speech Recognition Evaluation

The results in term of Word Error Rate (WER) for English and Spanish and Character Error Rate (CER) for Chinese are shown in Table 1.

	English (WER)	Spanish (WER)	Chinese (CER)
Top 1	8.2%	10.2%	9.8%
Tc-star	6.9%	8.1%	

Table 1: ASR results in WER/CER

For English, the best results are obtained by the TC-STAR system combination with a Word Error Rate of 6.9%, while the best individual system has obtained a WER of 8.2%. For Spanish, the best performance of an individual system is 10.2%. Again a ROVER combination of all hypotheses is performed and gives the best results with a WER of 8.1%. For Chinese, a common submission from two sites has performed with a CER of 9.8%.

Spoken Language Translation Evaluation

Three kinds of text data are used as input:

- The output of the automatic speech recognition systems (ASR);
- The manual transcriptions (Verb.) including spontaneous speech phenomena, such as corrections, false-starts, etc.;
- The Final Text Editions (FTE) provided by the European Parliament. Some sentences are

¹ <http://www.tc-star.org>

rewritten and text data do not include transcription of spontaneous speech phenomena.

The results in term of BLEU scores (Papineni et al. 2001) for English-to-Spanish (EnEs), Spanish-to-English (EsEn) and Chinese-to-English (ZhEn) are shown in Table 2.

	EnEs			EsEn			ZhEn	
	ASR	Verb.	FTE	ASR	Verb.	FTE	ASR	Verb.
Top 1	35.97	46.61	49.81	39.41	52.54	48.16	16.07	12.39
Tc-star	-	47.53	50.74	-	52.55	48.07	-	-

Table 2: BLEU scores [%] for the SLT evaluation

The SLT ROVER combination of automatic translation has been performed in addition to individual systems outputs and achieves the best results for most of the tasks. Moreover, a human evaluation has been carried out for the Spanish-to-English direction in which each segment has been evaluated in relation to both adequacy and fluency measures (White et al. 1994). Two evaluations per segment were carried out by 130 native Spanish evaluators. A total of 20,360 segments were evaluated. Table 3 shows the results of the SLT human evaluation for the Top 1 system, the ROVER and also for one of the two reference translations used for the automatic evaluation.

Task	Top 1			ROVER		Human Ref.	
	ASR	Verb.	FTE	Verb.	FTE	Verb.	FTE
Fluency	3.06	3.39	3.63	3.32	3.46	4.31	4.56
Adequacy	3.13	3.54	3.79	3.55	3.72	4.31	4.44

Table 3: Fluency and adequacy judgements on a scale from 1 to 5 (5: best; 1: worst)

Regarding the general performance of the automatic system, results are good, as the scores are all above 3. However, the difference between the human reference and the automatic systems is still considerable. The ROVER system is not better than the Top 1 system although the scores are very close.

Text to Speech Evaluation

TTS evaluation is made of subjective tests to evaluate the prosody, the acoustic synthesis, the expressive speech, the voice conversion and the global TTS component, done by 20 native speakers.

Table 4 presents the results for the overall quality test of the TTS component on the English, Spanish and Chinese languages. The evaluation is done in a translation scenario, using the ASR and SLT outputs.

	English	Spanish	Chinese
Top 1	3.41	4.33	3.84
Natural Voice	4.79	4.66	4.44

Table 4: Overall quality (1-5) for TTS evaluation

Natural voice obtains the best performance compared to the best TC-STAR system. However, for Spanish, the best system is pretty close to the natural voice.

In the following sections we describe the end-to-end procedure and results.

End-to-End Evaluation

Tasks and Languages

Although three different directions are performed in TC-STAR (English-to-Spanish, Spanish-to-English, Chinese-to-English) we only consider the English-to-Spanish direction for time and cost constraints.

The evaluation data consist of audio recordings in English of the European Parliament Plenary Sessions (EPPS). The raw resources consist of audio recordings of September-December 2005 in English of the parliamentary debates. The focus is on the Parliament Members speaking in English and therefore we have selected only politicians' speeches. These recordings are used as input data for the TC-STAR system.

The evaluation data is made of 20 segments of around 3 minutes each. So in total, the evaluation set is composed of one hour of speech and around 8,000 running English words.

Thanks to the multilingualism in Europe, the European Parliament is translating and broadcasting in real time, each Plenary Session in many languages, including Spanish. Therefore, the corresponding Spanish audio translation made by professional interpreters is available. This human translation audio data is evaluated in the same way as the automatic translation.

The evaluated TC-STAR system includes the following modules:

- the ASR module made of a combination of several ASR engines (Lamel et al., 2006), using the Recognizer Output Voting Error Reduction (ROVER) method (Fiscus, 1997);
- the SLT component provided by RWTH (Matusov et al., 2006);
- the TTS module developed by UPC (Bonafonte et al., 2006).

These three components are trained on data including training corpora built from the EPPS recordings.

For each audio sample in English an ASR output is produced, then the ASR output is automatically translated into Spanish and finally, the SLT output is synthesized in Spanish by the TTS module using the alignment between SLT and ASR to get the prosodic features from the source language. The transit from one component to another is done manually but no modifications on the different outputs are made and so the system can be considered as fully automatic.

Protocol

We try in this experiment to have a different approach regarding the already reported works. In (Somers and Sugita, 2003), judges were asked to paraphrase what they had understood or heard and were informed that the audio was synthesized. The answers were judged on a seven-point scale. This evaluation was applied to SLT and TTS outputs, while a morpheme error rate (MER) was applied to the ASR output. The seven-point scale evaluation was taken over within the LC-STAR project (Banchs et al.,

2006) for an utterance-based evaluation. In both cases, judges produced and evaluated audio outputs, taking into account that no judge evaluated the output produced by him/herself. Our approach here is slightly different.

In machine translation, the two basic concepts for human evaluation are usually *adequacy* and *fluency* based on a five-point scale. Rather than checking these criteria with expert translators, we decided to use *adequacy* and *fluency* questionnaires, filled in by human judges who are not particularly familiar with the speech-to-speech domain and who are not bilingual. Since the target public of a speech-to-speech system is made of non bilingual people, it seems more appropriate to carry out a subjective evaluation with native speakers of the target language (Spanish) without any constraint on the education level in the source language (English).

To process the evaluation, we have extracted 20 samples containing around three minutes each of English speech. Each sample is spoken by only one speaker.

Since we look for meaning preservation during the whole process, we have chosen to ask questions build on the English speeches. These questions are first translated into Spanish and then asked to the human judges, in order to observe the information loss or preservation.

Using this protocol, the evaluation is carried out in three steps:

- First, a questionnaire is established for each English sample; and then translated into Spanish;
- Then, evaluators assess the Spanish samples according to the evaluation protocol described below;
- Finally, subjective evaluations results (answers given by judges) are checked by a single person.

We also try to compare the TC-STAR system with the professional interpreters, and to do that correctly, judges were not informed about the presence of audio data from interpreters. Judges act like end-users, in as much as aim at observing to what extent the information is preserved and if the quality is sufficient. *De facto* we consider this evaluation as *user-oriented*.

Adequacy Evaluation

We believe it is very difficult for a human being to make a judgment on adequacy, based only on listening to synthesized speech in the target language and the source speech. Moreover such tests could be carried out only by perfectly bilingual speakers. Instead, we use a functional test in which comprehension is rated. Thus, adequacy evaluation is a comprehension test on potential users which allows to measure the rate of intelligibility. The level of adequacy is computed as the percentage or number of questions that are assessed correctly.

Our first goal is to know whether the speech-to-speech translation system is able to keep the meaning through the modules or not. The second goal, presented in the next section, is to know whether the audio quality of the output is sufficient.

Fluency Evaluation

This evaluation concerns a judgment test with several questions related to fluency as well as the utility of the system. Each fluency score is the mean of the five-point

scale answers. We focus only on the audio quality for two reasons: first the evaluation concerns the final audio output and the overall end-to-end system, second the text quality is already evaluated within the SLT human evaluation of individual components.

Evaluation Set up

Evaluation Set up for Adequacy

To carry out the subjective evaluation we have recruited 20 native Spanish speakers, between 18 and 40 years old, with no hearing impediments. In opposition to previous experiments (Banchs et al., 2006) they are not experts in speech synthesis and they are paid for the task. The reason is to have a realistic scenario. Subjects are required to have access to high-speed Internet connection and good listening equipment.

Each evaluator evaluates three different samples and at least one TC-STAR sample and one interpreter sample, plus either another TC-STAR or interpreter sample. If we do not make any distinction between the TC-STAR and the interpreters' data, the repartition of the evaluated data is done in such a way that no judge evaluates twice the same sample. We do not want judges to make comparison between a sample produced by the TC-STAR system and its corresponding human translation.

Since there is a total of 40 audio samples (20 Spanish samples from the TC-STAR system and 20 samples from the interpreters), some of them are evaluated twice. Samples are assigned in no particular order.

In order to check meaning preservation, we prepare a comprehension questionnaire of 10 questions for each sample. To do that, the manual transcriptions of the source English speech are used to prepare the 10 questions set per sample. We hold onto the answers to all 200 questions and use them as the "reference answers". The goal of those reference answers is to assess the evaluations done.

For example here is a part of an English sample:

"Of course the nature of our societies has changed dramatically over these years, economically, socially and technologically."

The question:

"What has changed dramatically over these years?"


is asked and the reference answer is:

"The nature of our societies".

Then all questions and answers are translated into Spanish. Subjective tests are carried out through the Internet. A specific interface (Figure 1) was developed. Thus, evaluators could do the assessment at home.

An informative web page explains the TC-STAR system as well as the evaluation procedure. Furthermore, each evaluator listens to one minute of synthesized speech to become familiar with the voice and completed a training session. The training audio is part of the 40 selected audio samples, but is not reused for the evaluation session.

Within the interface, the evaluator can play the sound corresponding to either TC-STAR speech or interpreter speech, but is not informed about the provenance of it (interpreters *versus* automatic system).

Evaluation 2 on 3
Haz clic en el boton  para escuchar el sonido.

Cuántas directivas se ocupan de la venta y producción de bienes para la oferta y para servicios?
56

Cuántas directivas se ocupan de la compra, la mercadotecnia y la presentación del fertilizante para venta?
16

Quién debe hacer un pago hacia el presupuesto central?
Los Estados Miembros

Cuál es el asunto central para crear una estructura propia para Europa en el siglo 21?
La Recaudación

Fue un error ligar la política común agrícola con la perspectiva financiera?
Sí

Figure 1: Snapshot (in Spanish) of the web interface for adequacy questionnaire

Evaluators are instructed to:

- Read the questionnaire;
- Listen to the whole sample;
- Listen to it a second time. They were allowed to stop the recording so as to write down the answers for the adequacy questionnaire.

Evaluation Set up for Fluency

At the end of the adequacy evaluation, judges are asked to fill the fluency questionnaire for each sample. They have to rate the following standard fluency questions (presented in Table 5) taking into account the audio sample he/she has just listened to (the questions are in fact in Spanish but are here translated here in English):

Test	Fluency questionnaire
Under-standing	Do you think that you have understood the message? 1: Not at all 5: Yes, absolutely
Fluent Speech	Is the system fluent? 1: No, it is very bad! 5: Yes, it is perfect Spanish
Effort	Rate the listening effort 1: Very high 5: Low, as natural speech
Overall Quality	Rate the overall quality of this translation system 1: Very bad, unusable 5: It is very useful

Table 5: Fluency questionnaire

A five-point scale is provided for each question. Only extreme marks (1 and 5) are explicitly defined, ranging from the lowest level (1) to the higher (5). After all the evaluations are done, the average of the fluency rate is computed for each of the four fluency questions. Then the mean is computed for the interpreter and the TC-STAR system speeches, thus giving two scores per fluency question, and therefore, an overall of 8 scores.

Results

After all the evaluations are done and the answers are compiled, a native Spanish speaker has compared the answers of the evaluators with the reference answers in Spanish. The evaluator is asked to be “flexible”, as the reference answers are not exactly the same as the evaluator’s answer. For example, the reference answer to the question:

“¿Qué publicación concierne al grupo del ponente?”

(“Which publication is the speaker’s group concerned about?”)

is:

“La publicación del código de conducta para las organizaciones no lucrativas”

(“The publication of the code of conduct for not-for-profit organisations”),

while the evaluator’s answer is:

“del código de conducta sobre las organizaciones sin ánimo de lucro”

(“of the code of conduct about not-for-profit organisations”)

It is obvious that the answers can only be marked as correct or inexact by a human being and not automatically. Each evaluator answers in a different form (style, whether it is the whole sentence or a single word, etc.) even if the answer submitted is correct. Furthermore, synonyms can be used, or paraphrases, etc. and the translation of the interpreter can also be a different translation from that of the TC-STAR system or the translation of the answers, even if all the answers are correct. Therefore, the examiner, who has both the reference answer and the system answer in front of him, has to pay attention to the meaning of the questions, etc. The results for the two evaluations are presented below. For adequacy, two scores are available for the audio output: the results of the evaluation done by the judges and the same results obtained by an expert who has the reference answers in front of him. For both adequacy and fluency evaluation, we present three scores for TTS, SLT and ASR in order to observe from which component the information and the quality are lost.

Evaluator Agreement

Since 20 samples are evaluated twice by different evaluators, the comparison of both scores allowed us to compute the inter-evaluator agreement. For the 200 adequacy answers available for the comparison, 153 are identical (i.e. either the two evaluators found the correct answer or neither of them found the correct answer) while 57 are different. Thus, we can say that, in general, the evaluators answer identically to the same question (for 77% of the questions), while some others questions raise problems. The agreement is quite the same if we take separately the interpreter (75%) or the TC-STAR system (79%).

For Fluency the agreement is very good: for the *understanding* factor, 15% of the evaluations are identical (which is quite a low percentage), but 95% of the evaluations obtain a score that did not differ in more than 1 unit between the first evaluation and the second evaluation (denoted as *1-agreement* hereafter). For the *fluent speech*, the *effort* and the *overall quality* the percentage of identical evaluations is above 30%, while the percentage of evaluations that differ in more than 1 unit is about 80%. Judges make very similar judgments. But, as shown in Table 6, the 1-agreement differs according to the interpreter or the TC-STAR system.

System	Under-standing	Fluent Speech	Effort	Overall Quality
Overall	95%	75%	75%	80%
Interp.	100%	64%	91%	73%
Tc-star	89%	89%	56%	89%

Table 6: Fluency inter-judge 1-agreement

According to the fluency criteria, judges are coherent either with the interpreter or with the automatic system, but differences are strong, especially for the fluent speech and effort tests. This could show the subjectivity level of each criterion. Actually, no definition is given to the judges for the four criteria, as we have performed a user-oriented evaluation. So each judge can have his own definition of the *Understanding* of a text or could find an audio more fluent than another judge, etc. Partly, that's why we chose to have more than one judge per audio sample.

Adequacy Results

Overall Evaluation

Table 7 presents the adequacy results for the interpreter and TC-STAR system speeches. It indicates:

- the subjective results of the end-to-end evaluation ("Subj." column) done by the same assessors who made the subjective evaluation. Scores are shown in percentage, a score of 100% means all the answers are correct, while a score of 0% means all the answers are incorrect;
- an objective verification of the presence of the answer ("Obj." column): the audio files have been validated by a native Spanish to check whether they contain the answers to the questions (the question are created from the English source and the answer might not be present in the Spanish translation). It shows the "evaluation of the evaluators", or the capacity of the evaluators to find the correct answers. Scores are shown in percentage, a score of 100% means all the answers are present in the sample, while a score of 0% means all the answers are not present in the sample;
- an objective verification of the presence of the answers at each component of the end-to-end process ("SLT output" and "ASR output" columns), in order to determine in which component of the TC-STAR system, the information is lost. To do that individual outputs

from each component (recognition output from ASR, translated output from SLT, and synthesized audio from TTS –corresponding to the "Obj." column) have been checked.

System	Audio output		SLT output	ASR output
	Subj.	Obj.		
Interp.	50%	72%	-	-
Tc-star	58%	83%	83%	91%

Table 7: Adequacy results

The scores of the interpreter audio samples are much lower for the subjective evaluation than for the objective one. Evaluators have found 50% of the correct answers in the interpreter's speech, while 72% of the answers could have been found. So there, 28% of the information has not been translated by the interpreter.

At first glance, the TC-STAR system performs better than the interpreter and the percentage of correct answers found in the samples is 58%, 8% more than for the interpreter.

The evaluators did not find 25% of the answers they could have found. This could be due either to the quality of the audio output or to the subjectivity of the tests. Results show that the TC-STAR system tends to translate more information than the interpreter.

Error Analysis

The results can be interpreted by the fact that the TC-STAR system sticks very closely to the source data while professional interpreters usually resume and reformulate when producing the translated speech. For example, they usually eliminate what they consider to be meaningless. But the TC-STAR system translates everything without analysing what is more or less important. In fact we can elaborate different hypothesis to explain the better results of the automatic translation system:

- the questions are too difficult or context dependant;
- due to a limited time frame and real time translation, interpreters filter the information to translate speaker discourse;
- interpreters reformulate or paraphrase speaker discourse, which causes some ambiguity.

For the TC-STAR system, the overall loss of information is 17%. It is quite easy to identify where the loss of information occurs by checking each individual component output. A native Spanish speaker read each question, and has checked to see whether the answers are present within the TTS audio (corresponding to the same objective evaluation done for the interpreter audio samples), then within the SLT translated text, and finally within the ASR recognized text. The results of those comparisons correspond to "SLT output" and "ASR output" columns in Table 7. The overall loss for the ASR component is 9% and 8% more for the SLT component. The subjective evaluation causes a 25% loss in addition. For the TC-STAR system, there are three audio files that contain 100% of the correct answers while there is only one for the interpreter. 14 audio samples contain more

than 80% of the required information (4 samples more than for the interpreter).

Again, those differences can easily be explained: interpreters filter and reformulate the information while the TC-STAR system does not. All the information goes through the whole chain of the automation TC-STAR system, without being selected.

Evaluation Based on Interpreter Information

Following the previous point, we decided to compare more precisely the interpreter and the TC-STAR system and we have then selected only questions for which answers are available in the interpreter samples. This is done under the assumption *a priori* that the human translated samples include the main and “important” information.

Our intention is to compare the overall quality of the speech-to-speech translation with the overall quality of the interpreters, excluding the noise factor of the missing information. Thus, we get a new subset of 144 questions for which all the information has been kept by the interpreter.

As previously mentioned, the same study is done on the three components, and the results are shown in Table 8:

System	Audio output		SLT output	ASR output
	Subj.	Obj.		
Interp.	67%	100%	-	-
Tc-star	63%	86%	86%	95%

Table 8: Adequacy results for the subset of the answers which are available in the human interpreters’ translations

Again, the preserved information decreases, but the results are better for this subset than for the whole question set. If we consider that the objective information loss is zero for the interpreter, the TC-STAR system is quite good in comparison since the system loses only 14% of the original information.

However, the subjective loss is 37%, while the subjective loss for the interpreter is 33%: the interpreter quality is slightly better than the TC-STAR system, contrary to the results shown in Table 7.

The 14% objective loss is decomposed in 5% for ASR and 9% more for SLT. As for previous experiments (Somers and Sugita, 2003), results show that MT is the “noisiest channel” since the component is the one which loses the most of information. Added to this point, we also find the overall meaning can be disturbed by the topic of the questions: the flexibility of the expert who did the objective evaluation could be arguable and another evaluation criterion should be implemented, as in question-answering domain, like the “wrong” criteria which means the answer is not complete or not sufficient. Then the question-answering metrics could be applied.

Fluency Results

Overall Evaluation

Table 9 presents the fluency results for the interpreter and TC-STAR system samples and shows the results of the four fluency questions. A score of 1 means the speech is of bad quality while a score of 5 means the speech is good.

System	Under-standing	Fluent Speech	Effort	Overall Quality
Interp.	3.45	3.48	3.19	3.52
Tc-star	2.34	1.93	1.55	1.93

Table 9: Fluency results

For interpreter, at first sight the scores are good, most of the audio samples have a result of 4 or 5, and the average is above 3 for all the fluency questions. The four means are quite identical, and so we wonder whether it is necessary to ask four questions instead of asking only one general question. Another solution would be to distribute the questions, thus separately evaluating the audio files (first all the comprehension questions, then all the fluency questions, etc.).

However, the interpreter’s results are not as good as one can expect. This is explained by the work condition of the interpreters who have to translate in real time. Moreover, recordings contain in the background the original speech in English as a light snarl (in fact due to the headphones of the interpreters).

For the TC-STAR system, a quarter of the samples is not fluent and is very hard to listen to. Excluding some exceptions, the audio samples are better for the interpreter than for the TC-STAR system.

We have tried to compare some of the samples in order to understand why the TC-STAR system obtains better results than the interpreter for some samples. (in particular three audio files). For those files we have established some deviations between the two outputs. Some are affecting the performance of human interpreter, while some others are affecting the performance of the TC-STAR system.

Error analysis for Interpreter

On the performance of human interpreter, we analysis two main situations:

- recording noises: interpreter audio does not contain his/her speech but the original speaker is slightly audible in the backchannel. Moreover, various noises are also present in interpreter samples, like people knocking on tables, breathing, etc. The same noises are not present in the TC-STAR audio, which makes listening to it much easier.
- speaker hesitations: the interpreters often hesitate. Contrary to the TC-STAR system, the interpreter listens to the speaker even while he is speaking, and it can be very difficult to speak smoothly. For instance, in some samples, the interpreter hesitates too much, pauses often, makes many false starts, lengthens words, makes errors with the beginning of certain words, etc.

Therefore, these audio file obtain the lower scores out of all the fluency questions. However, the hesitations can also be due to the quality of the English speaker (with the same errors as described above: hesitations, lengthens words, etc.), thus contributing to the amount of errors made by the interpreter and distorting the listening level of the evaluators.

Error analysis for the TC-STAR System

On the performance of TC-STAR system, we analysis three main situations:

- problems of grammatical agreement (for genders, numbers and verb tenses) as well as of sentence syntax: lack of agreement can pose a real problem for comprehension. For example, in the following sentence, we can see a combination of errors, which increases the difficulty of the understanding task of the evaluator:
Source sentence (in English):

that this debate is attended by French , German , Austrian , Belgian , British and and other colleagues .

Translated and synthesized sentence (in Spanish):

*que este debate es *la* que *asistieron* francés alemán *austríaca* belga y británico y colegas de otros .*

The lack of agreement between “debate” (masculine) and “la” (feminine) means having to make a bigger effort to understand the Spanish text. This is worsened by the fact that the verb is in 3rd person plural while the noun (beginning of the subject noun phrase) is in 3rd person singular but does not contain a determinant to indicate this. Moreover the syntactic structure has been changed.

The originally passive construction “is attended by” has been transformed into the inversed-order active-voice “es la que asistieron” (“is that which attended”) in Spanish, which would be actually more common than a passive construction in Spanish, but has probably caused a higher difficulty for the system to achieve a correct final gender and number agreement. Last but not least, there is also a change in syntactic structure between the English “other colleagues” and the Spanish “colegas de otros” (“colleagues of others”), which changes the meaning at the end of the sentence.

Although these may look like minor problems when considered individually, they pose serious problems for the evaluators when put together. Evaluators need to make a much bigger effort to follow and understand the translation as they are easily distracted by trying to find the association between the non-agreeing words or by trying to restructure the sentence so as to achieve a correct connotation.

- Break between two sentences: the interpreter pauses between two sentences while the TC-STAR system inclined to concatenate the sentences. It is therefore more difficult for an

evaluator to follow the course of the speech, with this particular sequence of sentences: the speech is not clear if two sentences are spoken as a single sentence.

- Less natural connection words: the TC-STAR system produces unnatural connection between words, as the words: “well”, “so”, etc. It also could frustrate the listener, because if the word does not have a good prosody, it seems to contribute significantly to the gist of the sentence while in reality it does not. The listener has to make an additional effort to “delete” the word in order to understand the meaning of the sentence.

Consequences

Unfortunately, the problems for interpreter and TC-STAR system can also pose a threat to the adequacy evaluation. But, as for the adequacy evaluation, we could easily compare the audio output fluency scores with the SLT fluency score, and the WER ASR score, as (Somers and Sugita, 2003) compare their utterance score with the Morpheme Error Rate of the ASR. We do not have the SLT component score for the interpreter, so we take the score of the corresponding reference translation. Finally, we take the mean of the four fluency scores for the audio output, and we obtain the figures shown in Table 10:

System	Audio output (1-5)	SLT output (1-5)	ASR output (WER) [%]
Interp.	3.41	4.31	-
Tc-star	1.93	3.38	6.9

Table 10: Fluency comparison

There is degradation for each new component, but it is strongly marked for the audio output, especially for the synthesis from TC-STAR. If we consider each score represents the quality of the system it is certainly the TTS component which increases the noise within the final audio.

Conclusions and Future Work

An evaluation methodology of speech-to-speech translation systems has been presented. A large quantity of information and results has been compiled and analyzed.

We have tried to compare interpreters and speech-to-speech translation system by using questionnaires filled in by human judges.

The TC-STAR system has promising results in terms of information restoration with 86% of information recovered.

Nevertheless, the system is not very fluent and some information is lost due to the lack of fluidity. Regarding the adequacy the protocol has to be revised, especially the questionnaire. The question-answering domain should provide a lot of information, about the type of questions asked, the difficulty of the questions and the evaluation itself. We used two criteria (“correct” and “inexact”) to evaluate judges’ answers, and we should probably introduce another one (“wrong” or “incomplete”). In the same way, questions should be probably weighted according to their difficulty (as an example, factual

questions seem to be harder to answer than Boolean questions – 51% of correct answers for factual questions against 62% for other questions) and they should not be only a reformulation of what the speaker says.

The fluency questionnaire should be restricted, using fewer questions. For instance, fluent speech and effort tests seem to be harder to evaluate for judges. Furthermore, deviations have to be restricted as much as possible. Unfortunately, the interpreter audio recordings can not be improved (except if other interpreters are taken to compare each other), but the TC-STAR system can.

Finally, it would be better to have more judges for each audio in this kind of evaluation, due to the high level of subjectivity.

During the third year of the TC-STAR project a new End-to-End evaluation was carried out and results are understudy.

Evaluation packages

The material used in TC-STAR is publicly available through ELRA's catalog of language resources (<http://catalog.elra.info>). Each evaluation package includes resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the first evaluation campaign. The aim of these evaluation packages is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself.

Acknowledgments

This work was supported by the TC-STAR project (grant number IST-506738). We would like to thank Marie-Neige Garcia for her work on the preparation of the End-to-End evaluation, and Fernando Villavicencio and Victoria Arranz for their help. We are also very grateful to all the participating sites of the evaluation who have built the TC-STAR system and the human evaluators.

References

- Banchs R., Bonafonte A., Pérez J., 2006. "Acceptance Testing of a Spoken Language Translation System" in Proceedings of LREC'06, Genoa, Italy.
- Bonafonte A., Agüero P., Adell J., Pérez J., Moreno A., 2006. "Ogmios: The UPC Text-to-Speech Synthesis System for Spoken Translation", in Proceedings of the TC-STAR Workshop on Speech-to-Speech Translation, pp. 31-36, Barcelona, Spain.
- Fiscus J. G., 1997. "A Post-Processing System To Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)", in Proceedings of IEEE ASRU Workshop, pp. 347-352, Santa Barbara.
- ITU-T (1993) "A method for subjective performance assessment of the quality of speech voice output devices", Draft ITU-T Recommendation P.85, COM 12-R 6
- Lamel L., Gauvain J.L., Adda G., Barras C., Bilinski E., Galibert O., Pujol A., Schwenk H., Zhu X., 2006. "The LIMSI 2006 TC-STAR Transcription Systems", in Proceedings of the TC-STAR Workshop on Speech-to-Speech Translation, pp. 123-128, Barcelona, Spain, June 2006.

Matusov E., Zens R., Vilar D., Mauser A., Popovic M., Ney H., 2006. "The RWTH Machine Translation System", in Proceedings of the TC-STAR Workshop on Speech-to-Speech Translation, pp. 31-36, Barcelona, Spain.

Mostefa D., Hamon O. and Choukri K., 2006. "Evaluation of Automatic Speech Recognition and Spoken Language Translation within TC-STAR: results from the first evaluation campaign", In Proceedings of LREC'06, Genoa, Italy.

Nießen S., Och F., Leusch G., and Ney H., 2000. "An evaluation tool for machine translation: Fast evaluation for machine translation research". In Proceedings of LREC 2000, pp 39-45, Athens, Greece.

Papineni K. et al. (2001). *BLEU: a Method for Automatic Evaluation of Machine Translation*. Research Report, Computer Science IBM Research Division, T.J.Watson Research Center, RC22176 (W0109-022).

Somers H. and Sugita Y., 2003. "Evaluating Commercial Spoken Language Translation Software", in Proceedings of the Ninth Machine Translation Summit, pp. 370-377, New Orleans.

White J. S. and O'Connell T. A. (1994). The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. *Proceedings of AMTA Conference, 5-8 October 1994*, Columbia, MD, USA.