An Experimental Methodology for an End-to-End Evaluation in Speech-to-Speech Translation

Olivier Hamon[(1,2)] and Djamel Mostefa[(1)]
(1) Evaluation and Language Resources Distribution Agency (ELDA)
55-57 rue Brillat-Savarin, 75013 Paris, France
(2) Laboratoire d'Informatique de Paris-Nord (UMR 7030) – Université Paris 13 & CNRS
99 av. J.-B. Clément, 93430 Villetaneuse, France
{hamon, mostefa}@elda.org

Abstract

This paper describes the evaluation methodology used to evaluate the TC-STAR speech-to-speech translation (SST) system and their results from the third year of the project. It follows the results presented in (Hamon et al., 2007), dealing with the first end-to-end evaluation of the project. In this paper, we try to experiment with the methodology and the protocol during the second end-to-end evaluation, by comparing outputs from the TC-STAR system with interpreters from the European parliament. For this purpose, we test different criteria of evaluation and type of questions within a comprehension test. The results reveal that interpreters do not translate all the information (as opposed to the automatic system), but the quality of SST is still far from that of human translation. The experimental comprehension test used provides new information to study the quality of automatic systems, but without settling the issue of what protocol is best. This depends on what the evaluator wants to know about the SST: either to have a subjective end-user evaluation or a more objective one.

1. Introduction

A Speech-to-Speech Translation (SST) 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. In TC-STAR[1], evaluations of individual components (ASR, SLT and TTS) are carried out and their performance is measured with methodologies and metrics specific to each component. Here, we focus on the evaluation of the SST as a whole by comparing the SST output speech with a human interpreter speech.

We first give a description of the tasks and languages, then we brink back the evaluation protocol, the methodology parts we modified, and how we set up the experiment. Finally, we present a part of the results obtained by the TC-STAR system and compare them to the human interpreter ones.

2. Tasks and Languages

For this second end-to-end evaluation of the TC-STAR project, we adopted the general features of the first end-to-end evaluation. Only the English-to-Spanish direction was considered, automatic systems being applied to data from audio recordings in English of the European Parliament Plenary Sessions (EPPS). The raw resources consist of 20 audio recordings of around three minutes each, from the parliamentary debates in English dating from June and July 2006. The total adds up to one hour of speech, namely around 8,000 words.

Professional interpreters from the European Parliament produce oral human translation in several European languages including Spanish.

Translations are done in real time, what allows us to evaluate human translation in the same way as automatic translation in order to compare the automatic and human speech translation performance. For this purpose, meaning preservation is checked between the audio input, in English, and the audio output, in Spanish.

For this second evaluation, 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 module made of a combination of several SLT engines, as a ROVER (Matusov et al., 2006);
- The TTS module developed by UPC (Bonafonte et al., 2006).

Therefore, if we exclude the transit from one module to another, the system is fully automatic: no manual modifications are done on the outputs of the modules.

For each audio sample in English, an automatic transcription is produced by several ASR systems and an ASR ROVER output is built up. This ASR output is automatically translated into Spanish by several SLT systems and a SLT ROVER output is also built up. Finally, the SLT output is synthesized in Spanish by the TTS module.

3. Protocol

In this experiment, we kept the same protocol as that used for the first end-to-end evaluation (Hamon et al., 2007) with few exceptions in order to experiment new methods.

[1] http://www.tc-star.org

The concepts of *adequacy* and *fluency* are based on machine translation (White et al., 1994) and calculated over a five-point scale which is filled in by several judges. We only change the content of the questions. In our case, we decided to select 20 judges who were not familiar with the speech-to-speech translation domain. They were native Spanish speakers and were able to do the subjective tests online, through a Web interface.
In order to process the evaluation, we extracted 20 samples containing around three minutes of English speech each. Each sample is a monologue.
The objective of this evaluation is twofold: on one hand, we want to look at how much of the meaning is preserved and, on another hand, we want to estimate the quality of the audio output.
Thus, we decided to ask questions built on the English speeches in order to work on what the speaker would mean (and, for instance, not what the interpreter understood and reformulated). These questions are first translated into Spanish and then, the translated questions are asked to human judges in order to observe the information loss or preservation in the target speech, in Spanish.
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, judges assess the Spanish samples according to the evaluation protocol described below;
- Finally, subjective evaluation 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 in the evaluation. Judges act like end-users, in as much as aim at observing to what extent the information is preserved and how much the quality is sufficient.
Thus, each judge receives four audio samples to evaluate: two from the TC-STAR system and two others from the interpreters. So, distribution is fair and audio samples are presented anonymously and distributed randomly among the judges.
With 40 audio samples (20 from the TC-STAR system and 20 from the interpreters) and 20 judges, each audio sample is evaluated twice: this helps to observe the agreement between judges, and most of all, it permits to compute a mean between judgments, in case some judges are mistaken.

3.1 Adequacy Evaluation

Adequacy evaluation is a comprehension test on potential users which allows estimating the rate of intelligibility of the audio outputs. The level of adequacy is computed as the rate of answers that are found by the judges, for each audio they assessed. The final objective of the *adequacy* evaluation is to determine whether the meaning is preserved or not.

In order to check this meaning preservation, we prepare a comprehension questionnaire of 10 questions for each sample. 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 “reference answers”. It means those reference answers are used as a gold standard when answers drawn up by the judges are checked manually. Then, all questions and answers are translated into Spanish.
For this evaluation, we tried to classify questions into three categories, partly coming from information retrieval (Voorhees and Dang, 2005): *Factoid* (70% of questions), *Boolean* (20%) and *List* (10%). This could determine the quality of the system according to the type of question. Table 1 gives examples, out of the context, for each type of question (Q), associated to the respective reference answer (A). Answers are possible in the context of what the speaker says only.

Categorie	Question & Reference Answer
Factoid	Q: Who stated that the origins of this crisis dated back many years? A: Lord Penrose
Factoid	Q: What is the essence of humanity? A: Desire for freedom
List	Q: Who represented the British Government? A: A spokesman for the Treasury, the Financial Services Authority and the Government Actuary
List	Q: Which issues concern the E.U. as a community of values? A: Tolerance, anti-discrimination and equality
Boolean	Q: Is it the right time to say the text must be rewritten? A: No, it isn't
Boolean	Q: Does Europe need Bulgaria and Romania? A: Yes, it does

Table 1: Samples of Adequacy questions.

Finally, after being translated, questions and audio files are put into the interface and judges can start the evaluations.
First, they are informed about the evaluation procedure and its context. Then, they can listen to each of their assigned audio files and answer the respective questions. They are not informed about the provenance of the audio (i.e. either from the TC-STAR system or from one interpreter).
Once all questionnaires have been filled out by judges, a single assessor checks all the answers manually, looking at whether they are correct or not. To that end, the assessor, who is a Spanish native speaker, uses the reference answers and compares them to the answers provided by judges. He then gives scores to each answer, according to the following criteria (the values given for the scoring are provided between brackets):

- *Wrong* (0): the answer is not correct;
- *Incomplete* (1): the answer is not perfect, but could be considered as good;
- *Right* (2): the answer is most certainly correct.

We were inspired by criteria which are widely used in the evaluation of systems from the information retrieval domain (Magnini et al., 2004). However, to be more consistent with the previous end-to-end evaluation, we split the *Right* and *Incomplete* criteria to obtain two criteria of assessment [2]. Then, after presenting the corresponding results, we study the behaviors when three criteria are used.
This part of the protocol differs slightly from the previous evaluation, for which two criteria (*Right* or *Wrong*) were used. We decided to revise the method of assessment to be able to be stricter with the answers of the judges.
Finally, when all the answers are assessed, the adequacy score (i.e. the meaning preservation) is computed by audio, then by output.

3.2 Fluency Evaluation

Further to the meaning comprehension test, we carried out a quality test. This fluency test is more subjective and several questions related to features such as quality of the audio or utility of the output are asked to the judges. Each fluency score is the mean of a five-point scale answer.
For each sample and after each adequacy judgment, judges are asked to fill in a fluency questionnaire. They have to rate the sample they have just listened to according to the four fluency questions shown in Table 2.

Test	Fluency questionnaire
Understanding	Do you think that you have understood the message? 1: Not at all 5: Yes, absolutely
Fluent Speech	Is the speech in good Spanish? 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 audio sample. 1: Very bad, unusable 5: It is very useful

Table 2: 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 highest (5).
Questions have been slightly modified between the previous experiment and this one, for the *Fluent Speech* and the *Overall Quality*, in order to improve the inter-judge agreement. It has been done regarding the comments from judges from this previous evaluation, and their agreement scores. In this way, two questions have been simplified whereas they were:

- For *Fluent Speech* criterion: *Is the system fluent?*
- For *Overall Quality* criterion: *Rate the overall quality of this translation system.*

Here, the notion of "system" disappears and the interest on the audio output is reinforced.
Finally, when all the samples have been rated by all the judges, the average values of each fluency rate are computed for both interpreters and TC-STAR system outputs. Scores can then be compared.

4. Results

After all the samples have been listened to, answered and rated, answers are checked by the assessor and validated or not, according to the three criteria of assessment. Then, the final results are computed for each output: we obtain scores for the adequacy and for the four fluency judgments. We first give the overall results using two criteria then we observe the results when three criteria are used in a specific section.

4.1 Judges Agreement

Each sample is evaluated twice by two different judges, so we can compute the inter-judges agreement.
In general, judges give similar answers: 75% of the 400 questions get the same assessment. It means that 25% of the questions raise problems, but some of them were not easy to answer. The agreement is slightly higher when judges answer questions from the interpreter samples (79%) than from the TC-STAR system (71%).
Finally, we observe that agreement is quite the same as in the previous experiment, which achieved 77% of agreement between judges.
For fluency, the agreement is quite low: 30% for *understanding*, 52% for *fluent speech*, 35% for *effort* and 45% for *overall quality*. However, it corresponds to the state-of-the-art and agreements are better than for the previous experiment (15% for *understanding* and above 30% for the other criteria, respectively). However, what is more interesting is the *1-agreement*, the ratio of scores that did not differ in more than 1 unit between the evaluation from the first judge and the evaluation from the second one. Table 3 presents those scores.

System	Understanding	Fluent Speech	Effort	Overall Quality
Overall	82.5%	80%	72.5%	90%
Interp.	85%	85%	60%	85%
Tc-star	80%	75%	85%	95%

Table 3: Fluency 1-agreements.

[2] We discuss in the next sessions the way to split the results. The assumption is that an incomplete answer may be considered as correct and not really show a problem of comprehension.

Judges provide similar rates about the quality of the samples, whether it is on the TC-STAR system or the interpreter ones. However, the *effort* criterion still causes problems, especially for the Interpreters' samples: *effort* 1-agreement is low regarding the 1-agreement of the other criteria. That is probably due to the difference of perception of judges, linked to the difficulty for interpreters to speak both smoothly and quickly, due to the real time translation constraints.

4.2 Adequacy

4.2.1 Overall Results

In order to compare with the previous end-to-end evaluation, we propose the results as if there were still two criteria, considering the *Incomplete* criterion as being *Right*. This corresponds to the definition we had for the previous evaluation, which was less strict about the correctness of the answers.

Table 4 presents the adequacy results for the interpreter and TC-STAR system speeches, indicating:

- The subjective results of the end-to-end evaluation ("Subj." column) done by the judges and checked by the assessor;
- An objective verification of the presence of the answers in each component ("Obj", "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, synthesized audio from TTS – corresponding to the "Obj." column – and speaker audio) are checked by the assessor.

System	Audio Output		SLT Output	ASR Output
	Subj.	Obj.		
Interp.	74% (50)	91% (72)	-	-
Tc-star	64% (58)	89% (83)	92% (83)	97% (91)

Table 4: Adequacy results. Scores are shown in percentage, a score of 100% means all the answers are correct. Scores from the previous experiment are shown between brackets.

Regardless of the type of evaluation (whether subjective or objective), interpreters' speeches obtain higher results than TC-STAR system speeches. Only 9% of the information has not been translated by the interpreters.

The difference between subjective and objective evaluations is quite strong (but similar for both TC-STAR system and interpreters): Judges did not find 25% of the information for the TC-STAR system and 27% for the interpreters' speeches.

In the same way, we can see that 3% is lost by the ASR module, 5% by the SLT module and 3% by the TTS module. It seems that some questions were difficult to answer out of context.

As in (Hamon & al., 2007), we decided to compare the TC-STAR system and the interpreters in a fair manner by only selecting questions for which answers are in interpreters' samples and corresponding to the objective evaluation.

We make the assumption that interpreters select important information because of their hard task of real-time oral translation. We then get a new subset of 182 questions, for which information has been kept by the interpreters. As with earlier experiments, the outcome of the study is presented in Table 5.

If we consider interpreters' translation as perfect (100% of the questions could be answered), then the TC-STAR system obtains rather good results.

System	Audio Output		SLT Output	ASR Output
	Subj.	Obj.		
Interp.	80% (67)	100%	-	-
Tc-star	66% (63)	91% (86)	93% (86)	97% (95)

Table 5: Fair Adequacy results. Scores are shown in percentage, a score of 100% means all the answers are correct. Scores from the previous experiment are shown between brackets.

In any case, results seem to be lower than for the previous evaluation and actual scores of interpreters' quality. The subjective loss is really deep for the TC-STAR system: judges do not find the information in the translated speech easily.

Finally, this is the SLT module that loses the most in terms of information, and interestingly enough, the TTS module also loses information and quality decreases.

4.2.2 Comparison with the Previous Experiment

Actually, comparison is indicative, while questions and answers are not the same for both evaluations, and data is checked on different contexts. Anyway, this permits to give an idea of system improvement.

Globally, results seem to be better in this evaluation than in the previous one. But that should be put into context, since interpreters also get better results. This is mainly explained by the increase in terms of performance of the TC-STAR system but also by the fact that questionnaires seem to be less difficult for this experiment.

In fact, improvement of the TC-STAR system is not so good. While improvement on interpreters data is of 48% in absolute for the subjective evaluation (and 36% for the objective one), it is 10% for the TC-STAR system (and 7% for the objective one).

So, even if scores are better, we can not say the TC-STAR system improves for this second end-to-end evaluation, since improvement is weaker than for the interpreters. That could be due to either the use of the SLT ROVER or the change of topics and context of data.

4.2.3 Criteria of Answers Assessment

After the general results presented above, we show the results according the three new criteria (*Wrong*, *Incomplete* and *Right*) for the meaning comprehension test and we try to observe the differences and the utility of the new method. In order not to lose information, we did not combine answers given by two different judges on the same questions.

Table 6 presents statistics for answers assessed as *Wrong* (W), *Incomplete* (I) and *Right* (R), and the combinations of results. The overall set contains 800 answers (there are 200 questions for interpreters and 200 questions for the TC-STAR system answered by two different judges each).

System	(R)	(I)	(W)	(R+I)	(I+W)
Interp.	236 (59%)	58 (15%)	106 (26%)	294 (74%)	164 (41%)
Tc-star	197 (49%)	59 (15%)	144 (36%)	256 (64%)	203 (51%)

Table 6: Adequacy results for each criteria of assessment.

For both outputs, *Incomplete* answers get the same rates, around 15%, what represents a rather small proportion of answers. In any case, decomposing the assessments in three criteria gives more accurate information about the systems performance. By taking (R+I) assessments as correct answers, performance is acceptable, while by taking only (R), assessments performance is quite low. That makes a strong difference on the perception of the results. However, (R+I) results are closer to the objective evaluations than (R) ones. It means that combining (R) and (I) criteria corresponds to a better assessment (i.e. determining the quality of a speech-to-speech system output).

It also means that we should not be extremely strict with the assessment of the answers. And as shown with the judges' agreement, errors and doubts may very well come occur when judges answer questions.

In any case, this is also the aim of the end-to-end evaluation, and the difficulty of subjective judgments: We would like to know the quality of outputs from an end-user point of view. That is probably most interesting, since it tests the usability of the system.

4.2.4 Type of Questions

When the questions are built up, we try to respect the proportion of Factoid, List and Boolean questions. At the end of the process, there were, for the subjective evaluation, 144 Factoid (127 for the objective evaluation), 15 List (14 for the objective evaluation) and 41 Boolean (41 for the objective evaluation).

Thus, Factoid questions are especially truncated when the selection is made for the objective evaluation. That would mean their respective contexts are not dealt in the same way by the interpreter. The fact is understandable as regards the List questions: when an interpreter hears something like an enumeration, he pays attention to translate correctly each point of the enumeration, because, in general, this is particular and important points of the discourse. For instance, in the part of the sentence:

> *[…] as they have been in their condemnation of racism, xenophobia, anti-Semitism, homophobia and indeed other hate speech and hate crimes*

the focus is made on the enumeration done by the speaker and the sentence would loose its consistency without these terms.

What is particularly odd is the difference between handling Factoid and Boolean questions. All the Boolean questions are selected for the objective evaluation, whereas more than 10% of the Factoid questions are not used for it. *A priori* the decrease effect should be the same. The single hypothesis we could have to this phenomenon is that the Factoid questions require exact and detailed answers. On the contrary, Boolean questions contain already details in the question and then answers can be detected in the output easily.

Results according to each type of questions are presented in Table 7. The "Subj./Obj." criteria are done according to the fact that the evaluation is made by judges and checked by the assessor, or directly made by the assessor with the help of the reference answers. The "Fair/Unfair" criteria are related to the fact that the evaluation is made on the selected question for which answers may be present in the interpreter output or not.

System	Factoid	List	Boolean
Interp. (Subj / Unfair)	69%	69%	90%
Tc-star (Subj / Unfair)	63%	54%	72%
Interp. (Subj / Fair)	76%	75%	90%
Tc-star (Subj / Fair)	66%	55%	72%
Interp. (Obj / Unfair)	89%	92%	100%
Tc-star (Obj / Unfair)	89%	83%	93%
Tc-star (Obj / Fair)	92%	82%	93%

Table 7: Adequacy results for each type of question.

Boolean questions are easier to answer than Factoid and List ones, as we could expected, since Boolean questions contain more information. Moreover, interpreters' scores are higher than TC-STAR system's ones and this is not really surprising.

The gap between the interpreters' results and the TC-STAR system results increases when the evaluation is *Fair*, instead of *Unfair*, whatever the type of question is: it shows the real difference between the systems. But this gap is reduced when the evaluation is objective rather than subjective.

This last point reveals how important the evaluation made by judges is: their perception and comprehension of the information remains different to that of the "real quality" of a system.

4.3 Fluency Results

Table 8 presents the fluency results for the interpreter and the TC-STAR system samples and shows the results for 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	Understanding	Fluent Speech	Effort	Overall Quality
Interp.	3.85	4.08	3.38	4.03
Tc-star	2.43	2.03	1.63	2.05

Table 8: Fluency results.

For the interpreters, at first sight, the scores are good and the averages are above 3 points for all the fluency questions, but the results are not as good as one may expect. This is explained by the working conditions of the interpreters who have to translate in real time. As we denoted in the previous experiment, there are some noises (background recordings, speaker's noises, etc.) and contexts (speaker hesitations) which cause difficulties to understand and to follow the speaker.
For the TC-STAR system, the quality is much lower than the interpreter one. Even if the Understanding is slightly higher, the audio quality is constraining for the judges, in particular represented by the Effort of listening.
Actually, the interpreters fail regarding Effort and Understanding, while the TC-STAR system fails in what concerns Effort but also on the Fluent Speech and the Overall Quality. That corresponds to the results of the previous evaluation too. In the same way, all scores for both interpreter and TC-STAR system are higher than those of the previous evaluation.

4.4 Data Analysis

Many analyses can be done on both the interpreter and TC-STAR speeches, with respect to the Adequacy or the Fluency criteria. Indeed, most of the errors could cause reduction of the quality. We try here to outline issues from both kinds of speech, in addition to those already found in the previous evaluation (Hamon et al, 2007).

In many audio outputs, the interpreters hesitate and make repetitions. That is probably due to the delivery of the speaker: there is no feedback when speakers talk and most of the time this is a fast speech, since speakers have a short time to utter their speech. Interpreters have some difficulties to follow speakers and then give fewer details. So, interpreters are *de facto* forced to select information, and, inevitably, they restrict the comprehension of the topic and disturb the listener (a judge in our case). For instance, a speaker gives many details in his speech while speaking quickly: as a consequence, the interpreter limits the translation and does not provide any details in the end, in order to resume the translated speech at a “quieter” moment (i.e. at the end of the speech or when the speaker breathes/makes a pause).

In the same way, some questions are general and the absence of major details prevents the judge from answering those questions. For instance, in one audio sample, a speaker talks about information published in the German newspaper “Der Spiegel”. But the interpreter avoids drastically the name of the newspaper, even if the rest of the information is translated. Then, judges could not answer the, even informative, question (in English) *“Which main German newspaper published a report denying the link between the World Cup and an increase in trafficking and forced prostitution?”*, since there was no link with the audio output to find the corresponding information.

Interpreters interpolate or reformulate the source coming from the speaker. In the same way, they summarize the speech. For instance, we found in the audio output five sentences from the speaker summarized into two sentences by the interpreter.

We also found the case for which an interpreter has to wait for the end of the speaker sentence (in English) to be able to translate it into Spanish, otherwise it is not possible to understand. As a consequence, the quality of the current sentence is lower (the interpreter has to speed up), but above all, the next sentence is also damaged since the speaker continues to speak during the translation, etc.

An interesting point concerns the impact of the prosody of the TC-STAR system on the comprehension of the output speech. That is probably one of the most surprising facts, the TTS output being (normally) the exact synthesis of the SLT output and TTS systems getting good results for synthesis (Mostefa et al., 2007). Actually, the explanation is rather simple and is due, in part, to the quality of the translation. Indeed, when the quality of the SLT output is quite low, the prosody breaks the flow and the output speech is less understandable. For instance, the sentence:

> *pero que no sería necesariamente el caso y no hay ninguna reflexión sobre Letonia permítanme añadir , y esto es sólo un ejemplo mientras que si esa empresa estaba fuera de la Unión Europea cada Estado miembro* ****comprobar**** *concienzudamente y que es un problema*

for which the following sentence is an attempt of translation in English:

> *but that wouldn't be necessarily the case and there is no reflection about Latvia let me add , and this is just an example whereas if that company was outside the European Union every member state* ****check**** *thoroughly and that it is a problem*

is a low quality translation of the source sentence:

> *but that wouldn't necessarily be the case and that's no reflection on Latvia let me add and this is just an example whereas if that company was outside the European Union every Member state would check thoroughly and that's a problem*

When looking at the context, reading the translation is still understandable. However, the problem arises when the sentence is synthesized and thus listened to. The TTS module stops its utterance just before the verb "comprobar" (i.e. "to check"), which introduces a long pause and gives the impression of starting a completely different new sentence afterwards, thus disassociating the subject from the verb (which is in fact in infinitive mode) and completely misleading and even confusing the listener/judge. As a consequence, the question (*"In which condition would Member States examine thoroughly a financial services company?"*) requires an answer which is, theoretically, in between the two sentences "perceived" by the judge and thus, he cannot find the searched information unless he makes some "strange" deduction. Moreover, even if the assessor had the reference answer in front of him, he decided to define the answer as "impossible to answer" regarding the audio, which also implies that the objective assessments are different for both SLT and TTS scores. For instance, in this particular case, the TTS score is lower than the SLT one.

Another case that shows a typical error that can be attributed to the TTS module is the following: named entities are not always well synthesized. For instance, in the translated sentence:

necesitamos acciones como Sophie Veld dicho de la Comisión y necesitamos actuar como han dicho muchos de la Presidencia finlandesa

translation of:

we need action as Sophie Veld said from the Commission and we need action as many have said from the Finnish Presidency

the name *"Sophie Veld"* is translated correctly and can be easily read into the text file, but the name is badly synthesized. Indeed, instead of the name, one can listen to something like *"comoso biebeld"* with a small distortion right in the middle:

- word *"como"* is combined with the phoneme *"so"*, beginning of the name *"Sophie"*, and a pause is inserted between the two created "words";
- phoneme "ph" is synthesized in "b" (what maybe due to the distortion);
- the *"v"* of *"Veld"* is pronounce *"b"*, like in Spanish.

Since the answer of the question *"Who said that we need action from the Commission and from the Finnish Presidency?"* is the name *"Sophie Veld"*, nobody managed to find the correct answer from the audio output, regardless of being a judge or an assessor listening to the audio.

Occasionally, judges make deductions/guesses from the translated speech, and answer a question correctly. This is clearly the case when the topic is about quantities or general knowledge (or, sometimes, named entities). For instance, the Party of European Socialists Women is translated by the TCSTAR system as "BSE" instead of something like "PS" (and moreover, the TTS module could not synthesized "BSE" correctly). Even if the "P.S." acronym was not in the audio, the judges answered the question *"How many signatures did P.S. Women collect for its petition in two months?"* correctly because the audio contains the sentence *"la recopilación de más de veinte tres mil firmas en dos meses"* (automatic translation of the sentence *"we collected more than twenty-three thousand signatures in two months"*). So judges managed to answer the question without the information on who collected the signatures.
In this regard, judges should be better informed about the evaluation task in order to avoid this kind of "under-evaluation".

Finally, and generally speaking, an objective validation still remains slightly subjective, and results should be taken carefully. Some questions may be ambiguous and whatever the output observed is from ASR, SLT, or TTS, the quality of answers is limited by the understanding of the speech or the text. This can be difficult, even with the reference answer available.

5. Conclusion

An evaluation of a speech-to-speech translation system has been presented. A methodology has been reused and modified in order to experiment different methods of evaluation. Similar results on a different data set have been obtained, with different judges and different questionnaires. This allows us to conclude that we have performed a rather robust evaluation.
The TC-STAR speech-to-speech system has been compared with interpreters of the European parliament, demonstrating the trench between an automatic system and humans. However, it also shows that people are able to understand audio in outputs from an automatic system in a certain context, and can answer questions about their meaning. Even if the audio quality is lower than what would be wished, translation of a politician speech could be understood, at least the outline, in a certain way.
The methodology of the Adequacy evaluation (the subjective part) has been studied in more detail, regarding the type of questions asked and the number of criteria for the assessment of answers. Splitting the number of criteria from two to three shows different results, and gives two different interpretations of them. However, it is closer to the objective evaluation when two criteria are used. Moreover, studying the system according to the type of questions asked allows finding other sources of errors and helps to diagnose the output.
The analysis of the end-to-end output is costly and becomes very time-consuming, since many parameters are involved, starting with the different modules. The advantage of the methodology proposed here is that it helps developers (among other people) to diagnose issues related to a SST system.

6. Acknowledgments

This work was supported by the TC-STAR project (grant number IST-506738). We would like to thank Victoria Arranz and Fernando Villavicencio 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 judges.

7. References

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*, Barcelona, Spain, pp. 31-36.

Hamon O., Mostefa D., Choukri K. (2007). End-to-End Evaluation of a Speech-to-Speech Translation System in TC-STAR. In P*roceedings of MT Summit XI*, Copenhagen, Denmark.

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,* Barcelona, Spain, pp. 123-128.

Magnini B., Vallin A., Ayache C., Erbach G., Peñas A., De Rijke M., Rocha P., Simov K., Sutcliffe R. (2004) Overview of the CLEF 2004 Multilingual Question Answering Track. In *Working Notes of the Workshop of CLEF 2004*, Bath, 15-17 september 2004.

Matusov E., Ueffing N., Ney H. (2006). Automatic Sentence Segmentation and Punctuation Prediction for Spoken Language Translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Trento, Italy, pp. 158-165.

Mostefa D., Hamon O., Moreau N., Choukri K. (2007). Technological Showcase and End-to-End Evaluation Architecture, Technology and Corpora for Speech to Speech Translation (TC-STAR) projects. *Deliverable D30*, May 2007.

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.

Voorhees E. and Dang H. (2005). Overview of TREC 2005 question answering track. In *Proceedings of TREC2005*.

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