End-to-End Evaluation in Simultaneous Translation

Olivier Hamon[1,2], **Christian Fügen**[3], **Djamel Mostefa**[1], **Victoria Arranz**[1],
Muntsin Kolss[3], **Alex Waibel**[3,4] **and Khalid Choukri**[1]
[1]Evaluations and Language Resources Distribution Agency (ELDA), Paris, France
[2] LIPN (UMR 7030) – Université Paris 13 & CNRS, Villetaneuse, France
[3] Univerität Karlsruhe (TH), Germany
[4] Carnegie Mellon University, Pittsburgh, USA
{hamon|mostefa|arranz|choukri}@elda.org,
{fuegen|kolss|waibel}@ira.uka.de

Abstract

This paper presents the end-to-end evaluation of an automatic simultaneous translation system, built with state-of-the-art components. It shows whether, and for which situations, such a system might be advantageous when compared to a human interpreter. Using speeches in English translated into Spanish, we present the evaluation procedure and we discuss the results both for the recognition and translation components as well as for the overall system. Even if the translation process remains the Achilles' heel of the system, the results show that the system can keep at least half of the information, becoming potentially useful for final users.

1 Introduction

Anyone speaking at least two different languages knows that translation and especially simultaneous interpretation are very challenging tasks. A human translator has to cope with the special nature of different languages, comprising phenomena like terminology, compound words, idioms, dialect terms or neologisms, unexplained acronyms or abbreviations, proper names, as well as stylistic and punctuation differences. Further, translation or interpretation are not a word-by-word rendition of what was said or written in a source language. Instead, the meaning and intention of a given sentence have to be reexpressed in a natural and fluent way in another language.

Most professional full-time conference interpreters work for international organizations like the United Nations, the European Union, or the African Union, whereas the world's largest employer of translators and interpreters is currently the European Commission. In 2006, the European Parliament spent about 300 million Euros, 30% of its budget, on the interpretation and translation of the parliament speeches and EU documents. Generally, about 1.1 billion Euros are spent per year on the translating and interpreting services within the European Union, which is around 1% of the total EU-Budget (Volker Steinbiss, 2006).

This paper presents the end-to-end evaluation of an automatic simultaneous translation system, built with state-of-the-art components. It shows whether, and in which cases, such a system might be advantageous compared to human interpreters.

2 Challenges in Human Interpretation

According to Al-Khanji et al. (2000), researchers in the field of psychology, linguistics and interpretation seem to agree that simultaneous interpretation (SI) is a highly demanding cognitive task involving a basic psycholinguistic process. This process requires the interpreter to monitor, store and retrieve the input of the source language in a continuous manner in order to produce the oral rendition of this input in the target language. It is clear that this type of difficult linguistic and cognitive operation will force even professional interpreters to elaborate lexical or synthetic search strategies.

Fatigue and *stress* have a negative effect on the interpreter, leading to a decrease in simultaneous interpretation quality. In a study by Moser-Mercer et al. (1998), in which professional speakers were asked to work until they could no longer provide acceptable quality, it was shown that (1) during the first 20 minutes the frequency of errors rose steadily, (2) the interpreters, however, seemed to be unaware of this decline in quality, (3) after 60 minutes, all subjects made a total of 32.5 meaning errors, and (4) in the category of nonsense the number of errors almost doubled after 30 minutes on the task.

Since the audience is only able to evaluate the simultaneously interpreted discourse by its form,

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 345–353,
Athens, Greece, 30 March – 3 April 2009. ©2009 Association for Computational Linguistics

the *fluency* of an interpretation is of utmost importance. According to a study by Kopczynski (1994), *fluency* and *style* were third on a list of priorities (after content and terminology) of elements rated by speakers and attendees as contributing to quality. Following the overview in (Yagi, 2000), an interpretation should be as natural and as authentic as possible, which means that artificial pauses in the middle of a sentence, hesitations, and false-starts should be avoided, and tempo and intensity of the speaker's voice should be imitated.

Another point to mention is the time span between a source language chunk and its target language chunk, which is often referred to as *ear-voice-span*. Following the summary in (Yagi, 2000), the ear-voice-span is variable in duration depending on some source and target language variables, like speech delivery rate, information density, redundancy, word order, syntactic characteristics, etc. Short delays are usually preferred for several reasons. For example, the audience is irritated when the delay is too large and is soon asking whether there is a problem with the interpretation.

3 Automatic Simultaneous Translation

Given the explanations above on human interpretation, one has to weigh two factors when considering the use of simultaneous translation systems: *translation quality* and *cost*.

The major disadvantage of an automatic system compared to human interpretation is its translation quality, as we will see in the following sections. Current state-of-the-art systems may reach satisfactory quality for people not understanding the lecturer at all, but are still worse than human interpretation. Nevertheless, an automatic system may have considerable advantages.

One such advantage is its considerable short-term memory: storing long sequences of words is not a problem for a computer system. Therefore, compensatory strategies are not necessary, regardless of the speaking rate of the speaker. However, depending on the system's translation speed, latency may increase. While it is possible for humans to compress the length of an utterance without changing its meaning (summarization), it is still a challenging task for automatic systems.

Human simultaneous interpretation is quite expensive, especially due to the fact that usually two interpreters are necessary. In addition, human interpreters require preparation time to become familiar with the topic. Moreover, simultaneous interpretation requires a soundproof booth with audio equipment, which adds an overall cost that is unacceptable for all but the most elaborate multilingual events. On the other hand, a simultaneous translation system also needs time and effort for preparation and adaptation towards the target application, language and domain. However, once adapted, it can be easily re-used in the same domain, language, etc. Another advantage is that the transcript of a speech or lecture is produced for free by using an automatic system in the source and target languages.

3.1 The Simultaneous Translation System

Figure 1 shows a schematic overview of the simultaneous translation system developed at Universität Karlsruhe (TH) (Fügen et al., 2006b). The speech of the lecturer is recorded with the help of a close-talk microphone and processed by the speech recognition component (ASR). The partial hypotheses produced by the ASR module are collected in the resegmentation component, for merging and re-splitting at appropriate "semantic" boundaries. The resegmented hypotheses are then transferred to one or more machine translation components (MT), at least one per language pair. Different output technologies may be used for presenting the translations to the audience. For a detailed description of the components as well as the client-server framework used for connecting the components please refer to (Fügen et al., 2006b; Fügen et al., 2006a; Kolss et al., 2006; Fügen and Kolss, 2007; Fügen et al., 2001).

3.2 End-to-End Evaluation

The evaluation in speech-to-speech translation jeopardises many concepts and implies a lot of subjectivity. Three components are involved and an overall system may grow the difficulty of estimating the output quality. However, two criteria are mainly accepted in the community: measuring the information preservation and determining how much of the translation is understandable.

Several end-to-end evaluations in speech-to-speech translation have been carried out in the last few years, in projects such as JANUS (Gates et al., 1996), Verbmobil (Nübel, 1997) or TC-STAR (Hamon et al., 2007). Those projects use the main criteria depicted above, and protocols differ in terms of data preparation, rating, procedure, etc.

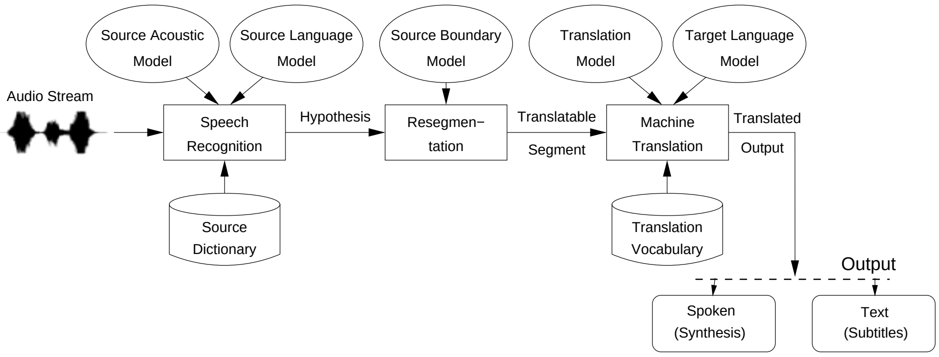

Figure 1: Schematic overview and information flow of the simultaneous translation system. The main components of the system are represented by cornered boxes and the models used for theses components by ellipses. The different output forms are represented by rounded boxes.

To our opinion, 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. To that aim, we reused a large part of the evaluation protocol from the TC-STAR project(Hamon et al., 2007).

4 Evaluation Tasks

The evaluation is carried out on the simultaneously translated speech of a single speaker's talks and lectures in the field of speech processing, given in English, and translated into Spanish.

4.1 Data used

Two data sets were selected from the talks and lectures. Each set contained three excerpts, no longer than 6 minutes each and focusing on different topics. The former set deals with speech recognition and the latter with the descriptions of European speech research projects, both from the same speaker. This represents around 7,200 English words. The excerpts were manually transcribed to produce the reference for the ASR evaluation. Then, these transcriptions were manually translated into Spanish by two different translators. Two reference translations were thus available for the spoken language translation (SLT) evaluation. Finally, one human interpretation was produced from the excerpts as reference for the end-to-end evaluation. It should be noted that for the translation system, speech synthesis was used to produce the spoken output.

4.2 Evaluation Protocol

The system is evaluated as a whole (black box evaluation) and component by component (glass box evaluation):

ASR evaluation. The ASR module is evaluated by computing the Word Error Rate (WER) in case insensitive mode.

SLT evaluation. For the SLT evaluation, the automatically translated text from the ASR output is compared with two manual reference translations by means of automatic and human metrics. Two automatic metrics are used: BLEU (Papineni et al., 2001) and mWER (Niessen et al., 2000).

For the human evaluation, each segment is evaluated in relation to *adequacy* and *fluency* (White and O'Connell, 1994). For the evaluation of adequacy, the target segment is compared to a reference segment. For the evaluation of fluency, the quality of the language is evaluated. The two types of evaluation are done independently, but each evaluator did both evaluations (first that of fluency, then that of adequacy) for a certain number of segments. For the evaluation of fluency, evaluators had to answer the question: "Is the text written in good Spanish?". For the evaluation of adequacy, evaluators had to answer the question: "How much of the meaning expressed in the reference translation is also expressed in the target translation?".

For both evaluations, a five-point scale is proposed to the evaluators, where only extreme values are explicitly defined. Three evaluations are carried out per segment, done by three different evaluators, and segments are divided randomly, because evaluators must not recreate a "story" and thus be influenced by the context. The total number of judges was 10, with around 100 segments per judge. Furthermore, the same number of judges was recruited for both categories: experts, from the domain with a knowledge of the technology, and non-experts, without that knowledge.

End-to-End evaluation. The End-to-End evaluation consists in comparing the speech in the source language to the output speech in the target language. Two important aspects should be taken into account when assessing the quality of a speech-to-speech system.

First, the information preservation is measured by using "comprehension questionnaires". Questions are created from the source texts (the English excerpts), then questions and answers are translated into Spanish by professional translators. These questions are asked to human judges after they have listened to the output speech in the target language (Spanish). At a second stage, the answers are analysed: for each answer a Spanish validator gives a score according to a binary scale (the information is either correct or incorrect). This allows us to measure the *information preservation*. Three types of questions are used in order to diversify the difficulty of the questions and test the system at different levels: simple Factual (70%), yes/no (20%) and list (10%) questions. For instance, questions were: *What is the larynx responsible for?*, *Have all sites participating in CHIL built a CHIL room?*, *Which types of knowledge sources are used by the decoder?*, respectively.

The second important aspect of a speech-to-speech system is the quality of the speech output (hereafter *quality evaluation*). For assessing the quality of the speech output one question is asked to the judges at the end of each comprehension questionnaire: "Rate the overall quality of this audio sample", and values go from 1 ("1: Very bad, unusable") to 5 ("It is very useful"). Both automatic system and interpreter outputs were evaluated with the same methodology.

Human judges are real users and native Spanish speakers, experts and non-experts, but different from those of the SLT evaluation. Twenty judges were involved (12 excerpts, 10 evaluations per excerpt and 6 evaluations per judge) and each judge evaluated both automatic and human excerpts on a 50/50 percent basis.

5 Components Results

5.1 Automatic Speech Recognition

The ASR output has been evaluated using the manual transcriptions of the excerpts. The overall Word Error Rate (WER) is 11.9%. Table 1 shows the WER level for each excerpt.

Excerpts	WER [%]
L043-1	14.5
L043-2	14.5
L043-3	9.6
T036-1	11.3
T036-2	11.7
T036-3	9.2
Overall	11.9

Table 1: Evaluation results for ASR.

T036 excerpts seem to be easier to recognize automatically than *L043* ones, probably due to the more general language of the former.

5.2 Machine Translation

5.2.1 Human Evaluation

Each segment within the human evaluation is evaluated 4 times, each by a different judge. This aims at having a significant number of judgments and measuring the consistency of the human evaluations. The consistency is measured by computing the Cohen's Kappa coefficient (Cohen, 1960).

Results show a substantial agreement for fluency (kappa of 0.64) and a moderate agreement for adequacy (0.52).The overall results of the human evaluation are presented in Table 2. Regarding both experts' and non-experts' details, agreement is very similar (0.30 and 0.28, respectively).

	All judges	Experts	Non experts
Fluency	3.13	2.84	3.42
Adequacy	3.26	3.21	3.31

Table 2: Average rating of human evaluations [1<5].

Both fluency and adequacy results are over the mean. They are lower for experts than for non-experts. This may be due to the fact that experts are more familiar with the domain and therefore more demanding than non experts. Regarding the detailed evaluation per judge, scores are generally lower for non-experts than for experts.

5.2.2 Automatic Evaluation

Scores are computed using case-sensitive metrics. Table 3 shows the detailed results per excerpt.

Excerpts	BLEU [%]	mWER [%]
L043-1	25.62	58.46
L043-2	22.60	62.47
L043-3	28.73	62.64
T036-1	34.46	55.13
T036-2	29.41	59.91
T036-3	35.17	50.77
Overall	28.94	58.66

Table 3: Automatic Evaluation results for SLT.

Scores are rather low, with a mWER of 58.66%, meaning that more than half of the translation is correct. According to the scoring, the *T036* excerpts seem to be easier to translate than the *L043* ones, the latter being of a more technical nature.

6 End-to-End Results

6.1 Evaluators Agreement

In this study, ten judges carried out the evaluation for each excerpt. In order to observe the inter-judges agreement, the global Fleiss's Kappa coefficient was computed, which allows to measure the agreement between m judges with r criteria of judgment. This coefficient shows a global agreement between all the judges, which goes beyond Cohen's Kappa coefficient. However, a low coefficient requires a more detailed analysis, for instance, by using Kappa for each pair of judges. Indeed, this allows to see how deviant judges are from the typical judge behaviour. For m judges, n evaluations and r criteria, the global Kappa is defined as follows:

$$\kappa = 1 - \frac{nm^2 - \sum_{i=1}^{n}\sum_{j=1}^{r} X_{ij}^2}{nm(m-1)\sum_{j=1}^{r} P_j(1-P_j)}$$

where:

$$P_j = \frac{\sum_{i=1}^{n} X_{ij}}{nm}$$

and: X_{ij} is the number of judgments for the i^{th} evaluation and the j^{th} criteria.

Regarding quality evaluation ($n = 6$, $m = 10$, $r = 5$), Kappa values are low for both human interpreters ($\kappa = 0.07$) and the automatic system ($\kappa = 0.01$), meaning that judges agree poorly (Landis and Koch, 1977). This is explained by the extreme subjectivity of the evaluation and the small number of evaluated excerpts. Looking at each pair of judges and the Kappa coefficients themselves, there is no real agreement, since most of the Kappa values are around zero. However, some judge pairs show fair agreement, and some others show moderate or substantial agreement. It is observed, though, that some criteria are not frequently selected by the judges, which limits the statistical significance of the Kappa coefficient.

The limitations are not the same for the comprehension evaluation ($n = 60$, $m = 10$, $r = 2$), since the criteria are binary (i.e. *true* or *false*). Regarding the evaluated excerpts, Kappa values are 0.28 for the automatic system and 0.30 for the interpreter. According to Landis and Koch (1977), those values mean that judges agree fairly. In order to go further, the Kappa coefficients were computed for each pair of judges. Results were slightly better for the interpreter than for the automatic system. Most of them were between 0.20 and 0.40, implying a fair agreement. Some judges agreed moderately.

Furthermore, it was also observed that for the 120 available questions, 20 had been answered correctly by all the judges (16 for the interpreter evaluation and 4 for the automatic system one) and 6 had been answered wrongly by all judges (1 for the former and 5 for the latter). That shows a trend where the interpreter comprehension would be easier than that of the automatic system, or at least where the judgements are less questionable.

6.2 Quality Evaluation

Table 4 compares the quality evaluation results of the interpreter to those of the automatic system.

Samples	Interpreter	Automatic system
L043-1	3.1	1.6
L043-2	2.9	2.3
L043-3	2.4	2.1
T036-1	3.6	3.1
T036-2	2.7	2.5
T036-3	3.5	2.5
Mean	3.03	2.35

Table 4: Quality evaluation results for the interpreter and the automatic system [1<5].

As can be seen, with a mean score of 3.03 even for **the interpreter**, the excerpts were difficult to interpret and translate. This is particularly so for

L043, which is more technical than *T036*. The *L043-3* excerpt is particularly technical, with formulae and algorithm descriptions, and even a complex description of the human articulatory system. In fact, *L043* provides a typical presentation with an introduction, followed by a deeper description of the topic. This increasing complexity is reflected on the quality scores of the three excerpts, going from 3.1 to 2.4.

T036 is more fluent due to the less technical nature of the speech and the more general vocabulary used. However, the *T036-2* and *T036-3* excerpts get a lower quality score, due to the description of data collections or institutions, and thus the use of named entities. The interpreter does not seem to be at ease with them and is mispronouncing some of them, such as “Grenoble” pronounced like in English instead of in Spanish. The interpreter seems to be influenced by the speaker, as can also be seen in his use of the neologism “el cenario” (“the scenario”) instead of “el escenario”. Likewise, “Karlsruhe” is pronounced three times differently, showing some inconsistency of the interpreter.

The general trend in quality errors is similar to those of previous evaluations: lengthening words (“seeeeñales”), hesitations, pauses between syllables and catching breath (“caracterís...ticas”), careless mistakes (“probibilidad” instead of “probabilidad”), self-correction of wrong interpreting (“reconocien-/reconocimiento”), etc.

An important issue concerns gender and number agreement. Those errors are explained by the presence of morphological gender in Spanish, like in “estos señales” instead of “estas señales” (“these signals”) together with the speaker’s speed of speech. The speaker seems to start by default with a masculine determiner (which has no gender in English), adjusting the gender afterward depending on the noun following. A quick translation may also be the cause for this kind of errors, like “del señal acustico” (“of the acoustic signal”) with a masculine determiner, a feminine substantive and ending in a masculine adjective. Some translation errors are also present, for instance “computerizar” instead of “calcular” (“compute”).

The errors made by the interpreter help to understand how difficult oral translation is. This should be taken into account for the evaluation of the automatic system.

The **automatic system results**, like those of the interpreter, are higher for *T036* than for *L043*. However, scores are lower, especially for the *L043-1* excerpt. This seems to be due to the type of lexicon used by the speaker for this excerpt, more medical, since the speaker describes the articulatory system. Moreover, his description is sometimes metaphorical and uses a rather colloquial register. Therefore, while the interpreter finds it easier to deal with these excerpts (known vocabulary among others) and *L043-3* seems to be more complicated (domain-specific, technical aspect), the automatic system finds it more complicated with the former and less with the latter. In other words, the interpreter has to “understand” what is said in *L043-3*, contrary to the automatic system, in order to translate.

Scores are higher for the *T036* excerpts. Indeed, there is a high lexical repetition, a large number of named entities, and the quality of the excerpt is very training-dependant. However, the system runs into trouble to process foreign names, which are very often not understandable. Differences between *T036-1* and the other *T036* excerpts are mainly due to the change in topic. While the former deals with a general vocabulary (i.e. description of projects), the other two excerpts describe the data collection, the evaluation metrics, etc., thus increasing the complexity of translation.

Generally speaking, quality scores of the automatic system are mainly due to the translation component, and to a lesser extent to the recognition component. Many English words are not translated (“bush”, “keyboards”, “squeaking”, etc.), and word ordering is not always correct. This is the case for the sentence “how we solve it”, translated into “cómo nos resolvers lo” instead of “cómo lo resolvemos”. Funnily enough, the problems of gender (“maravillosos aplicaciones” - masc. vs fem.) and number (“pueden realmente ser aplicado” - plu. vs sing.) the interpreter has, are also found for the automatic system. Moreover, the translation of compound nouns often shows wrong word ordering, in particular when they are long, i.e. up to three words (e.g. “reconocimiento de habla sistemas” for “speech recognition system” instead of “sistemas de reconocimiento de habla”).

Finally, some error combinations result in fully non-understandable sentences, such as:

> “usted tramo se en emacs es squeaking ruido y dries todos demencial”

where the following errors take place:

- *tramo*: this translation of “stretch” results from the choice of a substantive instead of a verb, giving rise to two choices due to the lexical ambiguity: “estiramiento” and “tramo”, which is more a *linear distance* than a *stretch* in that context;
- *se*: the pronoun “it” becomes the reflexive “se” instead of the personal pronoun “lo”;
- *emacs*: the recognition module transcribed the couple of words “it makes” into “emacs”, not translated by the translation module;
- *squeaking*: the word is not translated by the translation module;
- *dries*: again, two successive errors are made: the word “drives” is transcribed into “dries” by the recognition module, which is then left untranslated.

The TTS component also contributes to decreasing the output quality. The prosody module finds it hard to make the sentences sound natural. Pauses between words are not very frequent, but they do not sound natural (i.e. like catching breath) and they are not placed at specific points, as it would be done by a human. For instance, the prosody module does not link the noun and its determiner (e.g. “otros aplicaciones”). Finally, a not user-friendly aspect of the TTS component is the repetition of the same words always pronounced in the same manner, what is quite disturbing for the listener.

6.3 Comprehension Evaluation

Tables 5 and 6 present the results of the comprehension evaluation, for the interpreter and for the automatic system, respectively. They provide the following information:

identifiers of the excerpt: Source data are the same for the interpreter and the automatic system, namely the English speech;

subj. E2E: The subjective results of the end-to-end evaluation are done by the same assessors who did the quality evaluation. This shows the percentage of good answers;

fair E2E: The objective verification of the answers. The audio files are validated to check whether they contain the answers to the questions or not (as the questions were created from the English source). This shows the maximum percentage of answers an evaluator managed to find from either the interpreter (speaker audio) or the automatic system output (TTS) in Spanish. For instance, information in English could have been missed by the interpreter because he/she felt that this information was meaningless and could be discarded. We consider those results as an objective evaluation.

SLT, ASR: Verification of the answers in each component of the end-to-end process. In order to determine where the information for the automatic system is lost, files from each component (recognised files for ASR, translated files for SLT, and synthesised files for TTS in the “fair E2E” column) are checked.

Excerpts	subj. E2E	fair E2E
L043-1	69	90
L043-2	75	80
L043-3	72	60
T036-1	80	100
T036-2	73	80
T036-3	76	100
Mean	74	85

Table 5: Comprehension evaluation results for the interpreter [%].

Regarding Table 5, **the interpreter** loses 15% of the information (i.e. 15% of the answers were incorrect or not present in the interpreter’s translation) and judges correctly answered 74% of the questions. Five documents get above 80% of correct results, while judges find almost above 70% of the answers for the six documents.

Regarding **the automatic system** results (Table 6), the information rate found by judges is just above 50% since, by extension, more than half the questions were correctly answered. The lowest excerpt, *L043-1*, gets a rate of 25%, the highest, *T036-1*, a rate of 76%, which is in agreement with the observation for the quality evaluation. Information loss can be found in each component, especially for the SLT module (35% of the information is lost here). It should be noticed that the TTS module made also errors which prevented judges

Excerpts	subj. E2E	fair E2E	SLT	ASR
L043-1	25	30	30	70
L043-2	62	70	80	70
L043-3	43	40	60	100
T036-1	76	80	90	100
T036-2	61	70	60	80
T036-3	47	60	70	80
Mean	52	58	65	83

Table 6: Comprehension evaluation results for the automatic system [%].

from answering related questions. Moreover, the ASR module loses 17% of the information. Those results are certainly due to the specific vocabulary used in this experiment.

So as to *objectively compare* the interpreter with the automatic system, we selected the questions for which the answers were included in the interpreter files (i.e. those in the “fair E2E” column of Table 5). The goal was to compare the overall quality of the speech-to-speech translation to interpreters’ quality, without the noise factor of the information missing. The assumption is that the interpreter translates the “important information” and skips the useless parts of the original speech. This experiment is to measure the level of this information that is preserved by the automatic system. So a new subset of results was obtained, on the information kept by the interpreter. The same study was repeated for the three components and the results are shown in Tables 7 and 8.

Excerpts	subj. E2E	fair E2E	SLT	ASR
L043-1	27	33	33	78
L043-2	65	75	88	75
L043-3	37	67	83	100
T036-1	76	80	90	100
T036-2	69	88	75	100
T036-3	47	60	70	80
Mean	53	60	70	80

Table 7: Evaluation results for the automatic system restricted to the questions for which answers can be found in the interpreter speech [%].

Comparing the automatic system to the interpreter, the automatic system keeps 40% of the information where the interpreter translates the documents correctly. Those results confirm that ASR loses a lot of information (20%), while SLT loses 10% further, and so does the TTS. Judges are quite close to the objective validation and found most of the answers they could possibly do.

Excerpts	subj. E2E
L043-1	66
L043-2	90
L043-3	88
T036-1	80
T036-2	81
T036-3	76
Mean	80

Table 8: Evaluation results for interpreter, restricted to the questions for which answers can be found in the interpreter speech [%].

Subjective results for the restricted evaluation are similar to the previous results, on the full data (80% vs 74% of the information found by the judges). Performance is good for the interpreter: 98% of the information correctly translated by the automatic system is also correctly interpreted by the human. Although we can not compare the performance of the restricted automatic system to that of the restricted interpreter (since data sets of questions are different), it seems that of the interpreter is better. However, the loss due to subjective evaluation seems to be higher for the interpreter than for the automatic system.

7 Conclusions

Regarding the SLT evaluation, the results achieved with the simultaneous translation system are still rather low compared to the results achieved with offline systems for translating European parliament speeches in TC-STAR. However, the offline systems had almost no latency constraints, and parliament speeches are much easier to recognize and translate when compared to the more spontaneous talks and lectures focused in this paper. This clearly shows the difficulty of the whole task. However, the human end-to-end evaluation of the system in which the system is compared with human interpretation shows that the current translation quality allows for understanding of at least half of the content, and therefore, may be already quite helpful for people not understanding the language of the lecturer at all.

References

Rajai Al-Khanji, Said El-Shiyab, and Riyadh Hussein. 2000. On the Use of Compensatory Strategies in Simultaneous Interpretation. *Meta : Journal des traducteurs*, 45(3):544–557.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. In *Educational and Psychological Measurement*, volume 20, pages 37–46.

Christian Fügen and Muntsin Kolss. 2007. The influence of utterance chunking on machine translation performance. In *Proc. of the European Conference on Speech Communication and Technology (INTERSPEECH)*, Antwerp, Belgium, August. ISCA.

Christian Fügen, Martin Westphal, Mike Schneider, Tanja Schultz, and Alex Waibel. 2001. LingWear: A Mobile Tourist Information System. In *Proc. of the Human Language Technology Conf. (HLT)*, San Diego, California, March. NIST.

Christian Fügen, Shajith Ikbal, Florian Kraft, Kenichi Kumatani, Kornel Laskowski, John W. McDonough, Mari Ostendorf, Sebastian Stüker, and Matthias Wölfel. 2006a. The isl rt-06s speech-to-text system. In Steve Renals, Samy Bengio, and Jonathan Fiskus, editors, *Machine Learning for Multimodal Interaction: Third International Workshop, MLMI 2006, Bethesda, MD, USA*, volume 4299 of *Lecture Notes in Computer Science*, pages 407–418. Springer Verlag Berlin/ Heidelberg.

Christian Fügen, Muntsin Kolss, Matthias Paulik, and Alex Waibel. 2006b. Open Domain Speech Translation: From Seminars and Speeches to Lectures. In *TC-Star Speech to Speech Translation Workshop*, Barcelona, Spain, June.

Donna Gates, Alon Lavie, Lori Levin, Alex. Waibel, Marsal Gavalda, Laura Mayfield, and Monika Woszcyna. 1996. End-to-end evaluation in janus: A speech-to-speech translation system. In *Proceedings of the 6th ECAI*, Budapest.

Olivier Hamon, Djamel Mostefa, and Khalid Choukri. 2007. End-to-end evaluation of a speech-to-speech translation system in tc-star. In *Proceedings of the MT Summit XI*, Copenhagen, Denmark, September.

Muntsin Kolss, Bing Zhao, Stephan Vogel, Ashish Venugopal, and Ying Zhang. 2006. The ISL Statistical Machine Translation System for the TC-STAR Spring 2006 Evaluations. In *TC-Star Workshop on Speech-to-Speech Translation*, Barcelona, Spain, December.

Andrzej Kopczynski, 1994. *Bridging the Gap: Empirical Research in Simultaneous Interpretation*, chapter Quality in Conference Interpreting: Some Pragmatic Problems, pages 87–100. John Benjamins, Amsterdam/ Philadelphia.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. In Biometrics, *Vol. 33, No. 1 (Mar., 1977), pp. 159-174.*

Barbara Moser-Mercer, Alexander Kunzli, and Marina Korac. 1998. Prolonged turns in interpreting: Effects on quality, physiological and psychological stress (pilot study). *Interpreting: International journal of research and practice in interpreting*, 3(1):47–64.

Sonja Niessen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An evaluation tool for machine translation: Fast evaluation for mt research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens, Greece.

Rita Nübel. 1997. End-to-end Evaluation in Verbmobil I. In *Proceedings of the MT Summit VI*, San Diego.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), Research Report, Computer Science IBM Research Division, T.J.Watson Research Center.

Accipio Consulting Volker Steinbiss. 2006. Sprachtechnologien für Europa. `www.tc-star.org/pubblicazioni/D17_HLT_DE.pdf`.

John S. White and Theresa A. O'Connell. 1994. Evaluation in the arpa machine translation program: 1993 methodology. In *HLT '94: Proceedings of the workshop on Human Language Technology*, pages 135–140, Morristown, NJ, USA. Association for Computational Linguistics.

Sane M. Yagi. 2000. Studying Style in Simultaneous Interpretation. *Meta : Journal des traducteurs*, 45(3):520–547.