

TC-STAR: Proposal for end-to-end evaluation

Antonio Bonafonte and Jordi Adell

19th December 2005

1 Introduction

In the second evaluation campaign of TC-STAR, an end-to-end evaluation is planned. This evaluation includes speech recognition, spoken translation and speech synthesis. In translation, the two basic concepts to take into account are *adequacy* and *fluency*. However, we think that in *speech-to-speech* translation, rather than asking for this questions to translation experts it is preferable to use *adequacy* and *fluency* questionnaires, to be filled by human judges acting as potential users. In particular, we believe it is very difficult for an expert to make a *judgement* about the adequacy, based on the listening of the synthetic speech in the target language and the source speech. Instead, we propose to use a *functional test* were the understanding is rated.

- **Adequacy:** comprehension test on potential users allows measuring the intelligibility rate.
- **Fluency:** judgement test with several questions related to fluency and also usability of the system

In this document we make a proposal to evaluate the system and discuss some preliminary result.

2 Proposal for end-to-end evaluation

2.1 Production of evaluation samples

1. The evaluation agency (*EA*), select some *semantically interesting* excerpts, taken from English (source) politicians (not interprets). Approximately 20 speeches \times 3 minutes.
2. ASR \rightarrow ST \rightarrow TTS. We will get \sim 20 *evaluation samples* for each end-to-end system.
3. Proposal: one TC-STAR systems is evaluated in the 2nd campaign.
4. Furthermore, the speech from the interpret (in Spanish, as target language) is collected. These are the \sim 20 *reference samples*, which is the *top-line*.

2.2 Protocol for the judges

The evaluation is done by human judges without any specific experience on speech technology. They are explained the tc-star system and the evaluation procedure. Furthermore, they listen to one minute of synthetic speech to become familiar with the voice.

They are instructed to:

- Read the *adequacy* questionnaire.

- Listen to all the excerpt.
- Listen a second time. They are allowed to stop the playback to write down the answers to the *adequacy* questionnaire.

At the end of the evaluation session, they are asked to fill the the *fluency* questionnaire.

2.3 Adequacy questionnaire

For each excerpt, based on the English speech (and in the English and Spanish final text edition), the *EA* prepares a comprehension questionnaire (~10 question for each evaluation sample).

2.4 Fluency questionnaire

This questions are related to the whole evaluation and it is done at the end of the evaluation (after listening to several evaluation samples).

Please, rate from one to five the following questions:

- Q1 Do you think that you have understood the message?
1: not at all ,5: yes, absolutely
- Q2 The system is fluent? (The system makes good use of the language: syntax, pronunciation, etc.)
1: No, it is very bad! 5: Yes, it is perfect Spanish.
- Q3 Rate the listening effort
1: very high 5: low, as natural speech
- Q4 Rate the overall quality of this translation system
1: Very bad; unusable; 5: It is very useful

Evaluation subjects

3 Preliminary test and results

We have prepared one test for one *evaluation sample*. This test has been give to two judges (in this case, *experts* on the technology). The results for the *adequacy* test is:

3.1 Results about adequacy

- **TC-STAR system:** Subject 1: 8.5 over 16; Subject 2: 10 over 16
- **Spanish interpret** (from the EPPS): Subject 2: 13 over 16. (this test was done after evaluating the TC-STAR system, so this is not a fair condition, but I think the results would have been the same).

With the interpret, the results are not 100% because the interpret decided to skip some information. One reason maybe that they have to operate in real time.

3.2 Results about fluency

The results for the *fluency* test will be

- **TC-STAR system:** Q1: 3; Q2: 1.5; Q3: 1; Q4: 2
- **Interpret:** Q1: 5; Q2: 4; Q3: 4; Q4: 5

4 Questionnaire

To end this document, here we include questions and answers that we have defined for this evaluation sample. It is only a draft. This were originally formulated in English (i.e., the source language) and translated into Spanish. We have included different kind of questions (yes/no, very particular, some other a little more general).

Furthermore, you can listen to the speech files (original from the politician, interpret and tc-star output, in the wp3 workpackage area, (item *end2end proposal*, <http://gps-tsc.upc.es/wp3@tc-star/end2end/>

- 1.- On which river is the House?
- 2.- What has been tested and prevailed?
- 3.- Did they asked for more than their rights?
- 4.- What did they asked for?
- 5.- At which edge was it tested?
- 6.- What is a sign of maturity of institutions?
- 7.- What the speakers thinks the president do not believe?
- 8.- Who looses today?
- 9.- How much the voice of Democracy risen up?
- 10.- Where Democracy has made itself heard?
- 11.- What would happen if the governments of Europe do not respect the parliamentarians?
- 12.- What chance has the president?
- 13.- What did the speaker spoke about yesterday and it is back again?
- 14.- What do Liberals and Democrats call on the council?
- 15.- Where did they find themselves?
- 16.- When does the speaker want the Commission to be approved?

Figure 1: Questions (in English)

- A1. - Rhine
- A2. - The will of the House
- A3. - No
- A4. - Their considered judgment to be treated with respect at every stage.
- A5. - Until the unwalked ground of political crisis
- A6. - The ability to turn tension into mutual benefit
- A7. - That is was anti-European to vote against the president's Commission
- A8. - Euroscepticism
- A9. - One octave
- A10.- In every national capital and beyond
- A11.- They wouldn't respect the president neither.
- A12.- To rebuild a Commission that can win the support of the House.
- A13.- An invisible elephant.
- A14.- To repect the prerrogatives of the Chamber and the independence of the Commission
- A15.- At an impass
- A16.- On November

Figure 2: Answers (in English)

- 1.- ¿Al lado de qué río está la Casa?
- 2.- ¿Qué se ha puesto a prueba y ha prevalecido?
- 3.- ¿Pidieron algo que estuviera fuera de sus derechos?
- 4.- ¿Qué pidieron?
- 5.- ¿Hasta qué punto fue probada la Cámara?
- 6.- ¿El qué es una señal de la madurez de las instituciones?
- 7.- ¿Qué dice el ponente, que el presidente no cree?
- 8.- ¿Quién pierde hoy?
- 9.- ¿Cuánto ha subido la voz de la Democracia?
- 10.- ¿Dónde se ha hecho escuchar la Democracia a sí misma?
- 11.- ¿Qué pasaría si los gobernantes de Europa no respetaran los parlamentarios?
- 12.- ¿Qué oportunidad tiene ahora el presidente?
- 13.- ¿De qué habló el ponente ayer y hoy repite?
- 14.- ¿A qué instan los Liberales y los Demócratas al Consejo?
- 15.- ¿En qué situación se encontraron?
- 16.- ¿Cuándo quiere el locutor que se apruebe la Comisión?

Figure 3: Questions (in English)