Evaluation of Modules and Tools for Speech Synthesis
- The ECESS Framework -

H. Höge[1], Z. Kacic[2], B. Kotnik[2], M. Rojc[2], N. Moreau[3], H.-U. Hain[1]

[1]Siemens AG, Corporate Technology, 81730 Munich, Germany [2]University of Maribor, [3]ELDA

E-mail: harald.hoege@siemens.com, kacic@uni-mb.si, bojan.kotnik@uni-mb.si, matej.rojc@uni-mb.si, moreau@elda.org, horst-udo.hain@siemens.com

Abstract

The consortium ECESS[1] (European Center of Excellence for Speech Synthesis) has set up a framework for evaluation of software modules and tools relevant for speech synthesis. Till now two lines of evaluation campaigns have been established:

- Evaluation of the ECESS TTS modules (text processing, prosody, acoustic synthesis)
- Evaluation of ECESS tools (pitch extraction, voice activity detection, phonetic segmentation).

The functionality and interfaces of the ECESS TTS have been developed by a joint effort between ECESS and the EC-funded project TC-STAR[2]. First evaluation campaigns were conducted within TC-STAR using the ECESS framework. As TC-STAR finished in March 2007, ECESS continued and extended the evaluation of ECESS TTS modules and tools by its own. Within the paper we describe a novel framework which allows performing **remote** evaluation for modules via the web. First experimental results are reported. Further the result of several evaluation campaigns for tools handling pitch extraction and voice activity detection are presented.

1. Introduction

The consortium ECESS (European Center of Excellence for Speech Synthesis) aims to speed up progress in speech synthesis technology by providing an appropriate framework. Evaluation - based on a common evaluation methodology - is one of the key elements of this framework. Till now two lines of evaluation campaigns have been established:

- Evaluation of the ECESS TTS modules (text processing, prosody, acoustic synthesis)
- Evaluation of ECESS tools (pitch extraction, voice activity detection, phonetic segmentation).

The functionality and interfaces of the ECESS TTS modules (Perez 2006) and the procedures for evaluation (Bonafonte 2006) have been developed by a joint effort between ECESS and the EC-funded project TC-STAR. Within TC-STAR three evaluation campaigns on speech synthesis have been conducted by ELDA. Because TC-STAR finished in March 2007, ECESS continued and extended the evaluation of ECESS TTS modules by its own. The next ECESS evaluation campaign of modules handling UK English as language is scheduled on Spring 08. This evaluation will be the first evaluation campaign on speech synthesis, totally based on a web-based **remote** evaluation procedure as described later.

Within ECESS, different tools mainly needed for generating language resources and to analyse speech signals have been investigated. Till now several evaluation campaigns have been organized for tools handling pitch extraction and voice activity detection. These campaigns were organized in the 'traditional' way: Training, test data, and an evaluation script have been distributed to researchers, which test their modules themselves and report the results achieved. In future, **remote** evaluation will be explored also for tools.

The paper is organized as follows. Chapter 2 describes the methodology for evaluating ECESS modules based on a remote evaluation scheme. Chapter 3 describes the evaluation campaigns recently performed on tools.

2. Evaluation of ECESS TTS Modules

2.1 The Remote Evaluation System RES

For the evaluation of the ECESS TTS Modules a new experimental platform – the Remote Evaluation System (RES) - has been developed. The RES is dedicated to distributed web-based online evaluation of the ECESS TTS modules, TTS systems and tools running at different institutions worldwide. The RES is based on web-based client- server architecture and is built up by three different RES components: the RES server, the RES module server and the RES client. Figure 1 shows the basic structure of the RES system. The core component is the RES server, which acts as a Managing Unit (MU) connecting RES clients with RES module servers. Each ECESS partner has to install one or more RES module servers locally by embedding one or more of his ECESS TTS modules into a server shell provided by the RES tool 'unforma' based on JavaCC (Tom Copeland 2007).

Figure 2 illustrates interconnections between modules of the RES system. The MU connects the RES module

[1] www.ecess.eu
[2] www.tc-star.org

servers from one or several partners to build partial or complete TTS systems. Using the RES client, the evaluating institution is able to perform different evaluation tasks by sending test data via the RES client and by receiving results from the RES server modules. Additionally, the RES can be used by the partners to use ECESS TTS modules of other partners needed to test and improve the performance of their own module(s) or tools. All TTS modules and tools included in the RES are accessible via TCP/IP. Between RES clients and the MU the MRCP protocol is used for data exchange (IETF standard 2006). Each of the RES components performs specific sequence of actions, which is defined by XML based description in the form of finite-state machine. The 'protocolgen' tool has been developed for generating the XML based protocol descriptions and their implementation as finite-state machines.

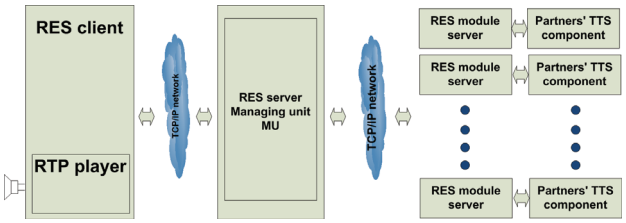

Figure 1: Basic structure of the RES system

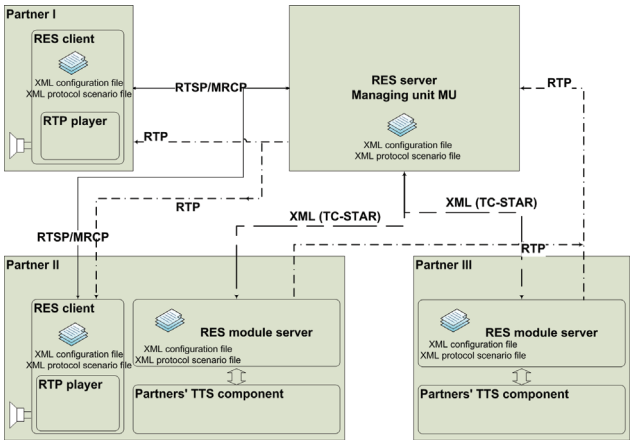

Figure 2: Interconnection between modules of the RES system.

First tests on using the RES framework have been already done. The integration of an ECESS TTS module into the RES module server is done by simply giving the name of a program or a script to be called in order to process the data. The input , which was sent by RES client through the MU, is stored by the RES module server in a predefined file. Then the ECESS tool or a script is started by the RES module server. The ECESS tool or script has to store the output results to a file which is then sent back through the MU to the RES client. In contrast to TC-STAR evaluation, no manual effort is needed for the developers of a TTS module to take part in the evaluation. The framework has been successfully tested under the Windows XP and the SuSE Linux 10.1.

2.2 Evaluation Campaigns on ECESS TTS Modules and Systems

In the framework of evaluation components and systems for speech to speech translation systems, three TTS evaluation campaigns took place within the TC-STAR project between 2005 and 2007. They addressed a large variety of evaluation tasks for TTS systems developed in three different languages: Chinese, English and Spanish. This resulted in the definition of evaluation protocols for different aspects of the TTS research field:

- global evaluation of TTS systems (subjective tests),
- separate evaluations of the three ECESS TTS modules (objective and subjective tests), and
- evaluation of TTS-related research tasks (voice conversion, expressive speech synthesis).

2.2.1 Evaluation Specifications for Text Processing

The first remote ECESS evaluation campaign deals with the text processing modules for the tasks already defined within the TC-STAR:

- Normalization of Non-Standard-Words (NSWs),
- End-of-Sentence Detection,
- POS Tagging, and
- Grapheme-to-Phoneme Conversion (G2P).

The normalisation task assesses the ability of the systems to disambiguate NSWs like abbreviations, acronyms, dates, times, phone numbers and so on. They have to be converted to the correct normalised form of the word, e. g. "Mr" to "Mister" or "29/02/08" to "29th of February 2008".

The end-of-sentence detection is evaluated by comparing the submitted end-of-sentence boundaries with a reference (manually segmented by an expert) resulting in a sentence error rate.

The submitted POS tag sequences are automatically aligned to a reference POS tag sequence (manually tagged by an expert), resulting in a POS tag error rate.

In the same way, the submitted phoneme sequences are automatically compared to a reference. In case there is a mismatch between the phoneme reference and the systems output, a human expert has to judge whether the generated transcription is a valid pronunciation variant or not. Two metrics are computed for the G2P evaluation: the Phoneme Error Rate (PER, percentage of erroneous phonemes) and the Word Error Rate (WER, percentage of words containing at least one erroneous phoneme).

2.2.2 Evaluation Procedure

The novelty of the ECESS TTS module evaluation is the usage of the RES framework. ELDA is responsible for running the evaluations using the RES client. At the moment, a text processing evaluation client has been implemented. The RES evaluation client will be extended to cover the other ECESS evaluation tasks which are envisaged in the future.

The evaluation procedure starts by activating the RES

evaluation client. The evaluation client sends the input text corpus (test set) to the MU, who disseminates the text to the RES module servers, embedding the different ECESS text processing modules to be evaluated. Once it gets the modules' output back (according to XML file format specified in LC-STAR[3]), the MU sends them to the RES evaluation client who performs the evaluation tasks mentioned above.
Next remote evaluation campaigns will focus on evaluating the other ECESS TTS modules for acoustic synthesis and prosody.

2.2.3 Language Resource Evaluation Packages for the First ECESS Evaluation Campaigns

The first ECESS evaluation campaigns will be performed for ECESS TTS modules designed for modules handling UK English as language. In order to evaluate the algorithmic performance of the modules of different partners the modules have to be trained using the same language resources.
The training LR package for UK English consists of a UK English phonetic lexicon (specifications according to LC-STAR, 50.000 common words), an UK English text of about 90.000 running words containing POS tags specified according to LC-STAR format, and annotated recordings of a female speaker (native speaker of UK English) of about 10 hours. In addition to the recordings a phonetic lexicon is delivered containing all pronunciation variants that were realised by this speaker.
The evaluation of the ECESS modules is based on an evaluation LR package. As it was done for the TC-STAR evaluation campaigns, ELDA will create that package for the ECESS campaigns. Such packages contain all the necessary data and information to reproduce the evaluation tasks and compare one module to all the modules participating on the campaign. It mainly includes the evaluation data sets, the scoring tools, results of the participants, and descriptions of the evaluation protocols.

2.2.4 Preliminary Evaluation

Preliminary evaluation runs will be conducted in March and April 2008 to test and tune the Remote Evaluation System. As a start, we will focus on the evaluation of the Grapheme-to-Phoneme Conversion (G2P) module in UK English. The evaluation data of the 2nd year of the TC-STAR evaluation are used. The G2P task is evaluated based on 1000 words for which phonetizations of reference (using the SAMPA alphabet for UK English) had been produced in the frame of the TC-STAR project. The word set consists of 3 sub-sets: common words (500 words), proper names (286 words) and geographic locations (214 words). The preliminary results will be presented at the conference.

3. Evaluation of ECESS Tools

Within ECESS different tools mainly needed for generating language resources and for analysing speech are developed and evaluated. In order to perform separate and independent evaluation of different tools, evaluation campaigns have been organized or are in the preparation phase in the following areas:

- Pitch determination and epoch marking (PDA/PMA evaluation campaign)
- VAD+ evaluation campaign (the acronym VAD+ stands for the voice activity and voicing detection),
- Phonetic segmentation campaign.

Till now, three PDA/PMA evaluation campaigns have been carried out in the scope of ECESS. The main objectives of these evaluation campaigns were to show the progress in development of various PDA and/or PMA algorithms by different ECESS partners, and to compare the performance of these tools against other well-known and broadly used PDA/PMA tools, like those integrated in the Praat toolkit. For each evaluation campaign, the reference database and the evaluation scripts were distributed to the contributing partners. The evaluation results were collected, compared, and presented on ECESS workshops.
The reference database constructed to evaluate PDA and PMA consists of parts of the SPEECON Spanish speech database (Iskra 2002). The three acoustical environments found in this database contain a large variety of distortions like additive noises, reverberations and channel distortions. From this database recordings of 60 male and female speakers were selected. Preparation of the evaluation database was carried out in several steps. In the first step, the 60 minutes of selected close-talking recordings were automatically pitch-marked (epoch marked). In the next step accurate manual rechecking and correcting of pitch marks is performed thus resulting in reference pitch-marked database. Furthermore, the database was manually segmented for non-speech, voiced, and unvoiced segments - the information used in the evaluation campaigns. The above described PDA/PMA/VAD+ reference database is already available by ELDA (ELRA Catalogue Reference S0218 2006). Additionally, the four smaller PDA/PMA evaluation datasets based on high-quality studio recordings (each of 10 minutes of speech material) are available.
The overall organization of the VAD+ evaluation campaign is very similar to the above presented PDA/PMA evaluation campaign. The training set consists of speech data from 50 speakers (2 manually VAD+ segmented sentences per speaker). The total amount of training files is 400, since there are 4 SPEECON channels. The test set contains different audio files from seen as well as unseen speakers.
The phonetic segmentation campaign (still under development) is slightly more complex due to the specific nature of this task. First, the set of phonemes has to be developed (i.e. SAMPA notation, phoneme set from SpeechDatCar US). Next, the training/development and evaluation datasets need to be selected, and manual phonetic segmentation must be performed. The HMMs used for the forced alignment will be trained on the

[3] www.lc-star.com

close-talking channel of the SpeechDatCar-US database. The result of segmentation on multiple speakers task will be evaluated on the test set of the TIMIT database (the training set of the TIMIT can be applied for the HMM adaptation procedure). The segmentation for the single speaker will be evaluated on the CMU ARCTIC database. The objective measures used in the evaluation procedure will be defined as the percentage of the segmentation errors smaller than certain tolerance interval(s), and the mean boundary distance between the reference and resulting phonetic segmentation data.

3.1 PDA/PMA Evaluation Criteria

In order to be able to evaluate the performance of the PDA/PMA algorithms, a set of evaluation criteria has been determined. The most important PDA evaluation criteria are voiced/unvoiced errors (VE/UE), and gross errors high/low (GEH/GEL) (B. Kotnik et al., 2006). The full set of definitions of evaluation criteria used are:

- **Voiced error (VE) and unvoiced error (UE)**

The voiced error (VE) presents the percentage of voiced speech segments which are misclassified as unvoiced. The unvoiced error (UE) presents the percentage of unvoiced speech segments which are misclassified as voiced. Both, the VE and UE are used to evaluate the performance of the voiced/unvoiced detection stage of the PDA algorithm:

$$VE[\%] = \frac{\sum((\text{Est\_Seg} = \text{U})\ AND\ (\text{Ref\_Seg=V}))}{\sum \text{Ref\_V}} \cdot 100\% \quad (3.1)$$

$$UE[\%] = \frac{\sum((\text{Est\_Seg} = \text{V})\ AND\ (\text{Ref\_Seg=U}))}{\sum \text{Ref\_U}} \cdot 100\% \quad (3.2)$$

- **Gross error high (GEH) and gross error low (GEL)**

The gross error high (GEH) presents the percentage of voiced speech segments for which the detected pitch is more than 20% higher than the reference pitch:

$$GEH[\%] = \frac{\sum(\text{Est\_Seg\_F0} > 1.2 \cdot \text{Ref\_Seg\_F0})}{\sum \text{Ref\_V}} \cdot 100\% \quad (3.3)$$

The gross error low (GEL) presents the percentage of voiced speech segments for which the detected pitch is more than 20% lower than the reference pitch:

$$GEL[\%] = \frac{\sum(\text{Est\_Seg\_F0} < 0.8 \cdot \text{Ref\_Seg\_F0})}{\sum \text{Ref\_V}} \cdot 100\% \quad (3.4)$$

- **Success Rate (SR) and Accuracy (ACC)**

The performance of the PMA is evaluated by the success rate (SR) and the accuracy (ACC) of settled pitch-marks (epochs) (H. Höge et al., 2006). The success rate ($SR_{\%}$) of the PMA is computed as follows:

$$SR_{\%} = \frac{|Corr|}{|Ref|} \cdot 100\% \ , \quad where \ \ |Corr| = |\{x | (x \in Test) \wedge (x \in Ref)\}| \quad (3.5)$$

In equation (3.5) *Ref* represents the set of all reference epochs, and *Test* represents the set of estimated (test-set) epochs. The number of correct epochs is the sum of PMA generated epochs which are in the tolerance interval of the reference pitch-marks. Duplicated (or replicated) pitch-marks are not considered as correct pitch-marks. In counting the correct epochs, a maximal tolerance deviation of 20% of the period time T at maximal presumable pitch frequency F (T=1/F) is allowed. Thus, this definition is consistent with the definition of the gross error in the case of pitch detection algorithms evaluation. The accuracy ($ACC_{\%}$) of the PMA estimation is defined as follows:

$$ACC_{\%} = \frac{|Corr| - |Ins|}{|Ref|} \cdot 100\% \quad (3.6)$$

In this case all the inserted pitch-marks (*Ins*) are subtracted from the number of correctly estimated pitch-marks.

3.2 VAD+ Evaluation Criteria

The VAD+ evaluation criteria are nonspeech detection accuracy (N_{ACC}), voiced speech detection accuracy (V_{ACC}), and unvoiced speech detection accuracy (U_{ACC}). Furthermore, the confusion matrix is computed to provide better insight into the classification error analysis. The accuracies N_{ACC}, V_{ACC}, and U_{ACC}, can be defined using the following equations:

$$N_{ACC} = \frac{\sum \text{Est\_Correct\_N\_Seg}}{\sum \text{All\_Ref\_N\_Seg}} \cdot 100\%$$
$$V_{ACC} = \frac{\sum \text{Est\_Correct\_V\_Seg}}{\sum \text{All\_Ref\_V\_Seg}} \cdot 100\% \quad (3.7)$$
$$U_{ACC} = \frac{\sum \text{Est\_Correct\_U\_Seg}}{\sum \text{All\_Ref\_U\_Seg}} \cdot 100\%$$

3.3 PDA/PMA and VAD+ Evaluation Results

This section presents a brief summary of the 3rd ECESS PDA/PMA evaluation campaign. Four institutions were involved in the evaluation: UBC (University of the Basque Country), UMB (University of Maribor), TUD (Technical University of Dresden), and METU (Middle East Technical University). Additionally, the evaluation performance of well known and broadly used PRAAT toolkit is also given. Table 1 present the PDA results of the SPEECON channel 0.

SPEECON Channel C0	VE+UE	GEH+GEL
Institution	(%)	(%)
UBC	17.22	0.78
TUD	14.18	8.09
UMB	11.23	2.54
METU	12.09	4.40
PRAAT	12.34	1.84

Table 1: ECESS PDA evaluation results

SPEECON Channel C0	SR	ACC
Institution	(%)	(%)
UBC	83.45	70.06
TUD	87.26	68.11
UMB	86.65	74.78
METU	86.36	70.45
PRAAT	87.16	70.89

Table 2: ECESS PMA evaluation results

SPEECON Channel C0	N_{ACC}	V_{ACC}	U_{ACC}
Institution	(%)	(%)	(%)
METU	85.52	86.12	82.40
UVIGO	91.35	94.63	85.41
UMB	93.70	96.81	79.15

Table 3: ECESS VAD+ evaluation results

It is evident from Table 1 that the lowest VE+UE error is achieved by the algorithm from UMB. Similarly, the lowest GEH+GEL error is produced by the algorithm from UBC. Table 2 shows the PMA performance evaluation results for the algorithms from the same institutions. In this case the highest value presents the better result. It can be concluded that the algorithm from TUD achieves the best SR performance among all the five tested procedures. The best ACC is achieved with the algorithm from the UMB.

Table 3 presents the results of the 1st ECESS VAD+ evaluation campaign. The three institutions were involved in this evaluation: METU, UVIGO (University of Vigo), and UMB. The algorithm from UMB achieved the best nonspeech (N) and voiced speech (V) accuracies. The best unvoiced speech classification accuracy is achieved by the algorithm of UVIGO.

The presented evaluation frameworks were found to be very efficient way to compare different algorithms in various evaluation scenarios. The overall performance of the selected tools outperformed the performance of the broadly used PDA/PMA tools (like Praat) substantially. The performances presented in evaluation campaigns can be implicated also in real use of PDA, PMA, or VAD+ algorithms.

4. Conclusion

The evaluation framework set by the ECESS partners provides a common evaluation methodology. It concerns the evaluation of TTS systems, TTS modules, as well as tools needed for generating the necessary language resources. Multilinguality is one of the key issues. Several evaluation campaigns of different tools were already performed with reports about substantial improvements of evaluated tools per campaign. A new perspective in evaluation is given by the development of the remote evaluation system (RES) that enables web-based on-line evaluation of TTS systems and TTS modules. Using the RES system, partners can make on-line evaluation of components running at different institutions worldwide. They can evaluate TTS modules or even the whole TTS system by including modules, if needed, from other partners in their own configuration or they could even construct and evaluate the whole TTS system without any of their own modules. To take part in the evaluation of particular module, no manual effort is needed by the developers of a TTS module, which is used by the RES. The defined evaluation framework represents an efficient and effective way of evaluation of TTS technology.

5. Acknowledgements

We want to thank all ECESS partners contributing to the evaluation campaigns.

6. References

J. Perez, A. Bonafonte , H.-U. Hain , E. Keller, S. Breuer, J. Tian 2006, ECESS Inter-Module Interface Specification for Speech Synthesis, Proc. LREC 2006

A. Bonafonte, H. Höge, I. Kiss, A. Moreno, U. Ziegenhain, H. van den Heuvel, H.-U. Hain, X. S. Wang, M. N. Garcia 2006, TC-STAR: Specifications of Language Resources and Evaluation for Speech Synthesis, Proc. LREC 2006

D. J. Iskra et al., 2002. SPEECON - Speech Databases for Consumer Devices: Database Specification and Validation. Proc. LREC'2002. ELRA Catalogue Reference S0218, http://www.elda.org

B. Kotnik et al., 2006, Evaluation of Pitch Detection Algorithms in Adverse Conditions, Proc. Speech Prosody 2006

H. Höge et. All., 2006, Evaluation of Pitch Marking Algorithms, Proc. ITG Fachtagung Sprachkommunikation 2006.

Tom Copeland 2007, Generating Parsers with JavaCC, Centennial Books, Alexandria, 2007.

IETF standard »A Media Resource Control Protocol (MRCP)«, RFC 4463, April 2006.