TC-STAR: Specifications of Language Resources and Evaluation for Speech Synthesis

A. Bonafonte[1], H. Höge[2], I. Kiss[3], A. Moreno[1], U. Ziegenhain[2], H. van den Heuvel[4], H.-U. Hain[2], X. S. Wang[3], M. N. Garcia[5]

1 TALP, Universitat Politècnica de Catalunya, Barcelona, Spain
2 Siemens AG Corporate Technology, Munich, Germany
3 Multimedia Technologies Laboratory, Nokia Research Center, Tampere, Finland
4 SPEX, Dep. of Language and Speech, Radboud University Nijmegen, Netherlands
5 ELDA, Evaluation and Language Resources Distribution Agency, Paris, France

antonio.bonafonte@upc.edu, harald.hoege@siemens.com, imre.kiss@nokia.com, asuncion@gps.tsc.upc.edu, ute.ziegenhain@siemens.com, H.vandenHeuvel@let.ru.nl, horst-udo.hain@siemens.com, Xia.S.wang@nokia.com, garcia@elda.org

Abstract

In the framework of the EU funded project TC-STAR (Technology and Corpora for Speech to Speech Translation), research on TTS aims on providing a synthesized voice sounding like the source speaker speaking the target language. To progress in this direction, research is focused on naturalness, intelligibility, expressivity and voice conversion both, in the TC-STAR framework. For this purpose, specifications on large, high quality TTS databases have been developed and the data have been recorded for UK English, Spanish and Mandarin. The development of speech technology in TC-STAR is evaluation driven. Assessment of speech synthesis is needed to determine how well a system or technique performs in comparison to previous versions as well as other approaches (systems & methods). Apart from testing the whole system, all components of the system are evaluated separately. This approach grants better assessment of each component as well as identification of the best techniques in the different speech synthesis processes. This paper describes the specifications of Language Resources for speech synthesis and the specifications for evaluation of speech synthesis activities.

1. Introduction

During the last years a big effort has been devoted to build Language Resources (LR) for Speech Recognition. In Europe many of these resources have been designed by research groups and their specifications became standards (SpeechDat, Speecon, Orientel) that were applied in other projects. The EU project LC-STAR[1] specified contents, production and formats for lexica for TTS in many languages including Arab, Turkish, Mandarin, and Russian among others. However, there is not a standard for defining a database for speech synthesis and many differences between TTS systems become from the kind of data and methodology to produce the data.

In the framework of the EU funded project TC-STAR[2] (Technology and Corpora for Speech to Speech Translation), research on TTS aims on providing a synthesized voice sounding like the source speaker speaking the target language. To progress in this direction, research is focused on naturalness, intelligibility, expressivity and voice conversion both, in the TC-STAR framework. For this purpose, specifications on large, high quality TTS databases have been developed and the data have been recorded for UK English, Spanish and Mandarin.

The development of speech technology in TC-STAR is evaluation driven. Assessment of speech synthesis is needed to determine how well a system or technique performs in comparison to previous versions as well as other approaches (systems & methods). Apart from testing the whole system, all components of the system are evaluated separately. This approach grants better assessment of each component as well as identification of the best techniques in the different speech synthesis processes.

This paper describes the specifications of Language Resources for speech synthesis, validation criteria and the specifications for evaluation of speech synthesis activities in the framework of TC-STAR project (Bonafonte et al, 2005).

This paper is organized as follows. Section 2 summarizes the specifications of the synthesis databases, Section 3 shows the validation criteria Section 4 shows the evaluation methodology, and Section 5 ends with conclusions.

2. Specification of Language Resources for Speech Synthesis

The aim of this work is to come up with specifications for language resources (LR) useful to produce LRs in a variety of languages. Within TC-STAR project LRs for TTS systems and selected research areas on speech synthesis were generated for UK-English, Spanish and Mandarin languages. Furthermore the specifications aim at serving as a basis for other projects like ECESS[3] which in long term intend to produce more languages.

In the context of HLT these specifications can be seen as a starting point to specify a basic language resource kit, (BLARK)4 for speech synthesis. In the context of TC-STAR the LR should be suitable for:

- building the most advanced state-of-the-art TTS systems: the TTS system built will also serve as a

[1] http://www.lc-star.com
[2] http://www.tc-star.org
[3] www.ecess.org
[4] BLARK is an initiative of the HLT community to make available needed language resources for each language

backend for a speech-to-speech translation system developed in TC-STAR.
- performing research on intra-lingual and cross-language voice conversion,
- performing research on expressive speech in the context of speech-to-speech translation.

Table 1 summarizes the most relevant features of the specifications.

Corpora	
Size	Baselines voices: 90K words/voice (10 hours) Voice conversion voices: 9K words/voice/language (1 hour) Expressive voice: 9K words/voice/language (1 hour)
Domains	Parliamentary transcribed speech, novels, special sentences and words, etc.
Speaker selection (for each baseline voice)	
Preselection	Experts analyze 5 talent voices, select two candidates
Selection	Create TTS prototype using 1 hour of speech; select the best one
Recording platform	
Format	96 kHz, 24 bits
Channels	close talk microphone, membrane microphone, glottograph
Annotation	
Orthographic	100% supervised
Prosodic	Broad labeling, 100% supervised
Phonetic	SAMPA, 100% supervised
Segmentation	Phoneme segmentation, 20% baseline voices supervised, 5% voice conversion voices supervised
Pitch labelling	reference point defined; 20% baseline voices supervised, 5% voice conversion voices supervised
Interchange format	
Files format	Format defined for signal files, labeling files, lexicon
Other Information	Documentation, signal measures, speaker information, etc.
Validations	
A detailed validation protocol has been defined. The validation is done by an agency (SPEX) independent of the producer. The validation protocol includes checking corpus content and coverage, speaker selection, recording studio, signal quality, labeling and database formats.	

Table 1: Specification of Language resources

2.1. Corpora

The creation of voices for TTS systems and research on voice conversion are based on read speech.

The corpora design is divided in various sub-corpora:

1. Baseline corpora: Intended for the baseline system. Contains 10 hours of read recorded speech. Corpora are built from transcriptions of parliamentary speeches, novels and frequent phrases selected from some specific domains.
2. Voice conversion corpora: Recorded by bilingual speakers, contains one hour of read speech in each language (English/ Spanish, English/Mandarin). This corpus was designed by translating a set of sentences taken from the European parliament.
3. Mimic sentences: (same supra-segmental structure): Intended for intra-lingual voice conversion.
4. Expressive speech: Designed in the project for a speech-to-speech translation framework in a parliament translation application. Read data and recorded data (e.g. recordings from the Spanish or European parliament) is used.

2.2. Recording platform and Speaker selection

Speakers should be recorded in a silent room SNR_A>40dBA with a reverberation measure RT60< 0,3 sec at 96 KHz sampling rate and 24 bit/sample. Two microphones (a large membrane microphone and a close-talk microphone) plus the laryngograph signal are recorded simultaneously.

The procedure to select each baseline speaker consists of two steps. First, some experts analyze five professional speakers' voices and select two speakers.

Second, each selected candidate records one hour of speech. Signals are phonetically segmented and a speech synthesis voice is built. The final selection is carried out by a listening test that scores:

- Pleasantness of their voice
- Quality of the laryngograph signal
- Quality of speech manipulated using TD-PSOLA.
- Quality of synthesized signal.

2.3. Annotation

For each utterance (speech file) the database provides:

- the prompt text used to elicit the utterance,
- the orthographic annotation,
- the phonetic transcription,
- a rough annotation of symbolic prosody,
- the segmentation into the pitch marks, associated with the glottal closure.

2.3.1. Orthographic annotation

Orthographic annotation is a transliteration of what was actually said by the speaker without ambiguities at word level. Furthermore, if the signal of a given word is not suited for concatenative speech synthesis, the word is preceded by the symbol '*'.

2.3.2. Phonetic transcription

The recordings are fully phonetically transcribed. The transcription has to be 100% supervised to annotate what the speaker really said, including elision, reduction or assimilation present in continuous speech.

2.3.3. Symbolic prosody annotation

Phrase breaks were annotated using two levels: minor break (intermediate intonational phrase) and major break (full intonational phrase).

Pitch accent (intonational prominence) was annotated using two levels: 'normal' and 'emphatic'.

2.4. Segmentation

2.4.1. Phonetic segmentation

All the signals are segmented automatically and/or manually. Two hours of the baseline voices are manually supervised and a 5% of the conversion voices were

checked manually. The segmentation matches the manual phonetic transcription.

2.4.2. Word segmentation

All the expressive speeches are segmented either automatically and/or manually into words. For each word, the starting and ending time is provided.

2.4.3. Pitch Marking

Speech signals of all the baseline and conversion voices are labeled with pitch marks. The pitch marking points are defined with reference to the maximum of signal (maximum is defined in close neighborhood of the positive slope of laryngograph signal). Two hours of the baseline voices and one hour of the speech conversion voices have to be checked manually.

The LRs for UK English, Spanish and Mandarin are finished and will be accessible to third parties during 2007.

3. Validation of Language Resources

TC-STAR has validation protocols for the language resources (LRs) that are developed in the project. The validations are carried out by an independent validation centre, SPEX, that is not involved in the production of the resources. Validation criteria are formulated in close co-operation with the producers, though. Thus, the validation centre is involved from the start of the design of the resources to set up suitable quality measures. The validation protocols for LRs for Automatic Speech Recognition and Spoken Language Translation that are produced in TC-STAR are addressed elsewhere (Van den Heuvel et al, 2006). Here we limit ourselves to the validation criteria for the TTS training resources of TC-STAR.

A number of the validation specifications are quite similar to those of other resources. These concern:

- The documentation files of the resources. The information therein should be clear, complete and correct
- The formal structure of the resources. The directory structure and file names should obey the specifications, as should the formal content of the annotation and meta-files.
- The completeness of the design. Minimum counts are defined to ascertain that sufficient types and tokens of different text materials are collected and recorded.

On the other hand specific criteria for TTS LR are defined. These typically concern:

Speakers:

- The presence of the minimum amount of speakers for each voice type will be checked. Fewer speakers will not be accepted.
- The baseline voices will be judged on a 1-5 point scale by auditory inspection along the dimensions:
    - nativeness
    - age (22-50)
    - proficiency / experience as professional speaker
    - fluency
    - intelligibility

Scores below 3 will be reported as insufficient.

Signal quality:

- SNRA > 40 dBA must be achieved for 90% of the speech; SNRA measured on labeled data; (automatic inspection of SAM labels)
- Clipping less than 0.1%; (automatic inspection of SAM labels)
- Digitizing: 24bit A/D accuracy (16 bits optional), 96KHz sampling) ; (automatic inspection of SAM labels)
- Frequency range 40 - 20 000Hz; 0.5dB deviation (channel after the microphone has flat frequency response in this range). The frequency range and frequency response of the acquisition system should be documented.
- Reverberation RT60 < 0.3s; has to be documented for a typical session; (automatic inspection of SAM labels)
- A maximum of 5% of the label files may have REV > 0.3; (automatic inspection of SAM labels)
- A maximum of 2% of the label files may have REV > 0.4; (automatic inspection of SAM labels)
- It is checked if there is a laryngograph file for each speech file and vice versa (unless specified otherwise). (automatic check on file presence)

Phonemic transcriptions and segmentations:

- A maximum of 5% PER in the phonemic transcriptions is allowed.
- The segmentations of speech segments are checked, equally distributed between the manual and automatic segmentation. A max. of 5% wrong segmentations is allowed for the manual part and 10% for the automatic part. An error in segmentation is defined as in terms of deviations in ms. Deviations of more than 25 ms are considered an error.

20 minutes of speech will be validated. A native speaker of the language performs the check on phonetic transcriptions and segmentations. The transcriptions in the label files are checked by listening to the corresponding speech files and by correcting the transcriptions and/or segment boundaries if necessary. As a general rule, the delivered transcription and segmentation should always have the benefit of the doubt; only overt errors should be corrected.

Prosodic transcriptions:

- A max. of 20% deviation on prosodic annotations is allowed (WER at prosodic level).

This check will be performed on the same sample as for the orthographic transcriptions, if possible by the same person, and at least by a native speaker of the language. The transcriptions in the label files are checked by listening to the corresponding speech files and by correcting the prosodic transcriptions if necessary. As a general rule, the delivered transcription should always have the benefit of the doubt; only overt errors should be corrected.

Pitch marks:

- Maximum deviation from reference pitch mark: 5% of pitch period but not bigger than 0.5 ms, 3% of voiced/unvoiced errors; 3% voiced/unvoiced errors (automatic pitch tragger.).

Since the TTS training LR were still under production at the time this contribution was written, there are no validation results available at present. Our account serves to illustrate the importance of defining validation criteria before LR are finalised, and to give an idea of the broad

range of criteria involved in the validation of this type of LR.

4. Specifications of Evaluation of Speech Synthesis

A number of tests have been defined in the project to evaluate speech synthesis systems as a whole and the different speech synthesis activities. Finally we are interested on the quality of the overall system. However, the evaluation of the whole (black box evaluation) does not allow pinpointing which part of the system causes the most relevant problem. Furthermore, this method does not allow participating on the evaluation to small teams of researchers whose speciality of research is in one specific topic. The specifications define tests for evaluating the whole systems but also for evaluating different tasks to drive more valid conclusions about the results of different algorithms. Three broad modules have been agreed between partners: symbolic preprocessing, prosody generation and acoustic synthesis. Defining modules, with well defined input and output allows evaluating the performance of different approach under the same conditions (glass evaluation). For instance, the prosody is evaluated using correct input (pronunciation, normalization, etc.) and using the same acoustic back-end which generates the speech or acoustic output from the prosody description. Some objective metrics have been proposed to evaluate some modules of the speech synthesis system, but in most of the cases the evaluation relies on human judges. Specific tests have been defined to evaluate voice conversion and expressive speech in the speech-to-speech translation scenario and are outlined in Table 2.

Module 1: Text analysis	
Test M1.1	Evaluation of text normalization and end of sentence detection
Test M1.2	Evaluation of word segmentation (Mandarin)
Test M1.3	Evaluation of POS tagger
Test M1.4	Evaluation of Pronuntiation
Module 2: Prosody	
Test M2.1	Evaluation of prosody (using segmental information, resynthesis)
Test M2.2	Judgment test using delexicalized utterances
Test M2.3	Functional test using delexicalized utterances (identify written sentences which the produced delexicalized prosody)
Module 3: Acoustic generation	
Test M3.1	Intelligibility (functional test)
Test M3.2	Naturalness
System evalution	
Test S	System evaluation (based on ITU P.85), MOS Evaluation in end-to-end system, including ASR and translation
Voice conversion	
Test VC.1	Voice conversion *removing* prosody effect
Test VC.2	Voice conversion *including* prosody
Expressive speech	
Test E	Judgement test about speech expresivity

Table 2: Evaluation Test for Module Evaluation

This evaluation procedure was used in the first TC-STAR evaluation campaign and will be used in the second open evaluation campaign[5] .

5. Conclusions

This paper described the TTS Language Resources specifications and evaluation criteria generated in the TC-STAR project. The databases have been produced in three languages: UK English, Spanish and Mandarin and the first evaluation campaign finished with the collaboration of several TTS groups.

6. Acknowledgement

The TC-STAR project is supported by the EC in the sixth Framework Programme under contract FP6-506738.

7. References

A. Bonafonte, H. Höge , H. S. Tropf , A. Moreno, H. van der Heuvel, D. Sündermann , U. Ziegenhain, J. Pérez, I. Kiss (2005). TTS Baselines and Specifications. Deliverable D8 of the EU project TC-STAR "Technology and corpora for Speech to Speech Translation" (FP6-506738)

Van den Heuvel, H., Choukri, K., Gollan, Chr., Moreno, A., Mostefa, D. (2006). TC-STAR: New language resources for ASR and SLT purposes. In Proceedings of the LREC 2006, Genoa, Italy.

[5]The call can be found in www.tc-star.org