TQE: Transcription Quality Evaluation

1 Project Title & Acronym and Abstract

Title: Transcription Quality Evaluation
Acronym: TQE
Abstract:

The current proposal is about a completely automatic Transcription Quality Evaluation (TQE) tool. Input is a corpus with audio files and phone transcriptions (PTs). Audio and PTs are aligned, phone boundaries are derived, and for each segment-phone combination it is determined how well they fit together, i.e. for each phone a TQE measure (a confidence measure) is determined, e.g. ranging from 0-100%, indicating how well the fit is, what the quality of the phone transcription is. The output of the TQE tool will consist of a TQE measure and the segment boundaries for each phone in the corpus. The tool will be useful for validating, obtaining, and selecting phone transcriptions, for detecting phone strings (e.g. words) with deviating pronunciation, and, in general, it can be usefully applied in all research - in various (sub-)fields of humanities and language and speech technology (L&ST) - in which audio and PTs are involved.

Target Start Date: 01-01-2010
Target End Date: 01-07-2010
Type: Demonstrator Project

2 Coordinator

Name: Dr. H. Strik
Function: Assistant professor
Organization: Centre for Language & Speech Technology (CLST) [6], Radboud University
Address: Erasmusplein 1, 6525 HT Nijmegen, The Netherlands
E-mail: H.Strik@let.ru.nl
Tel: +31 – 24 - 3616104
Fax: +31 – 24 - 3612907
Role(s): User and Technology Provider

3 Composition of the Project Team

Name: Drs. R. van Veenendaal
Function: Project manager
Organization: Institute for Dutch Lexicology (INL – Instituut voor Nederlandse Lexicologie), HLT Agency (TST-Centrale – Centrale voor Taal- en Spraaktechnologie) [7]
Address: Witte Singel/Doelencomplex, Matthias de Vrieshof 2-3, 2311 BZ Leiden, Netherlands
E-mail: remco.vanveenendaal@inl.nl
Tel: +31 - 71 - 5272495 / +32 32654601
Fax: +31 - 71 - 5272115
Role(s): Data Provider

Name: Drs. D. Broeder
Function: MPI CLARIN centre manager
Organization: Max Planck Institute for Psycholinguistics (MPI) [8]
Address: Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
E-mail: Daan.Broeder@mpi.nl
Tel: +31 – 24 - 3521103
Fax: +31 – 24 - 3521213
Role(s): Technology Provider

4 CLARIN centre

Organization: Max Planck Institute for Psycholinguistics (MPI) [8]
Name: Drs. D. Broeder
Function: MPI CLARIN centre manager
Address: Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
E-mail: Daan.Broeder@mpi.nl
Tel: +31 – 24 - 3521103
Fax: +31 – 24 - 3521213

5 Requested Budget: Just the requested amount in Euro

59.275 Euro

6 Description of the Proposed Project

6.1 Research Question(s)

In this section we only mention the questions, while in section 6.4 we explain how the TQE tool can be used to address these questions. The questions are:
(1) How to validate phone transcriptions (PTs)? (e.g. as a part of corpus validation)
(2) How to obtain PTs? (e.g. as a part of corpus construction or linguistic research)
(3) How to select utterances with reliable phone transcriptions for further research?
(4) How to detect words or phones that are realized differently?
These four (classes of) questions play an important role in different (sub-)fields of humanities and language and speech technology (L&ST). As corpora will play a more prominent role in the future, and their size will gradually increase, the importance of automatic tools for analyzing large amounts of data will grow accordingly.

6.2 Research Data

The research data will be the corpora CGN and JASMIN (for details see [9] & [10], resp.). For CGN IMDI metadata are available; JASMIN is an extension of CGN, compatible with CGN, and IMDI metadata will be created. All participants are very familiar with these corpora.

6.3 Technology

The core component is the calculation of the TQE measure. The input consists of audio and PTs. The alignment and segmentation are carried out by the well-known Viterbi algorithm. For each segment-phone combination a confidence measure (CM) is calculated indicating how well phone and segment fit together: the TQE measure.

We have ample experience with calculating such CMs for PTs [1, 2, 3, 4], in previous (e.g. Dutch-CAPT [14]) and current projects (e.g. the Stevin [13] projects DISCO [15] and ST-AAP [16]). An important aspect of our research on non-native speech (i.e. speech of language learners) is automatic pronunciation error detection (PED), which is carried out by calculating CMs for segment-phone combinations, just as it will be done in the TQE tool. The most relevant aspects are described here (for more details see [1, 2, 3, 4, 14, 15, 16]).

We have compared and evaluated several algorithms for different combinations of features [1, 2, 3, 4]. The following features have been used as confidence predictors:

- MFCCs: Mel-frequency cepstral coefficients
- ASR-based features: posterior-probabilities, Goodness-of-Pronunciation
- Phonetic features: spectral (formants, etc.) and durations (raw and normalized)

These confidence predictors are combined to obtain a confidence measure. Our results show that substantial improvements can be obtained by combining different features, that MFCCs, ASR-based features (e.g. posterior-probabilities and Goodness-of-Pronunciation), and duration yield the best results, and that spectral features do not contribute significantly. Thus only an automatic speech recognition (ASR) system is sufficient to obtain the features needed (in the projects DISCO and ST-AAP we use SPRAAK [11]). We also experimented with different combination techniques. The best results were obtained with support vector machines (SVMs), for which we used the LIBSVM package [12]. For these auxiliary resources (SPRAAK and LIBSVM) IPR should not be an issue (see also section 8).

Note that in pronunciation error detection (PED) we have to make a binary decision, i.e. pronunciation error or not, while in the TQE tool the measure can vary more gradually, e.g. from 0 to 100%: the TQE measure reflects in a graded way how well the phone and the associated segment fit together.

6.4 Description

The TQE tool can be used for all (classes of) research questions mentioned in section 6.2. Here we will briefly explain how.

(1) How to validate PTs? The TQE measures obtained for a whole corpus can be useful as a basis for corpus validation (formerly often carried out by SPEX [17], now also by CLST [6]).

(2) How to obtain PTs? The TQE measure can be useful in generating PTs. For instance, PTs can first be derived automatically (e.g. canonical transcription, by lexicon lookup or grapheme-to-phoneme conversion), then the TQE tool is applied, for stretches containing many phones with low TQE carry out manual PT, use these manual PTs to improve the other automatic PTs (e.g. by means of decision trees, in [5] we showed that this procedure provided the best results).

(3) How to select utterances for research? For instance, utterances or words containing many phones with low TQE values could be discarded for humanities or L&ST research.

(4) How to detect phones or words that are realized differently? For instance, if the PTs of 'standard' (e.g. canonical) pronunciations are provided, the TQE measure will indicate which parts deviate from the provided 'standard' pronunciation.

The TQE tool can be usefully applied in various (sub-)fields of humanities and L&ST, in all research in which audio and PTs are involved.

6.5 Plan

The project will comprise five tasks: (1) user and (2) technological survey, development of (3) core component and (4) demonstrator, and (5) coordination (see sections 7 and 10). The project will start with a user and technology survey (see section 7). Depending on the outcomes of these surveys details regarding the plan (described below) will be finalized.

The input of the TQE tool will consist of audio (e.g. broadband, wav format) and PTs (e.g. SAMPA [19]). Information about conversion of formats, and pointers to conversion tools will be supplied. The software will be implemented in such a way that extra functionality (e.g. support for other formats) can easily be added later. The output will be the TQE measure (e.g. 0-100%) and the phone boundaries. The user interface (UI) will run completely in a browser (e.g. Internet Explorer (IE)) of the user. At the client side there will be a script for uploading a zip file containing the audio and PTs to the server side. At the server side a python application will then be started. Python APIs are already available for SPRAAK and LIBSVM. The zip file will be unzipped; the input for SPRAAK will be audio and PTs, the output the confidence predictors; input for LIBSVM will be the confidence predictors, output the confidence measures. When processing is finished, the user will be notified (e.g. by e-mail) that the results can be downloaded from the server side to the client side.

SPRAAK runs on Windows (XP or higher) and Linux/Unix. The required bandwidth and CPU time are mainly determined by the size of the audio files. However, since this is not a real-time application the processing does not have to be very fast. Bandwidth and CPU time can be delimited by specifying a maximum size for the uploaded zip file.

At the client side only a web-browser is needed. Users first have to register and agree on the conditions of use. If their registration is accepted, they receive an e-mail which contains information on how to activate their account, how to login, upload a zip file, etc.

7 Deliverables and Milestones

Table 1. List of deliverables and milestones

Id.	deliv.	Month	effort	involved	Responsible
<user>	docum.	3	1	INL	R. van Veenendaal
<core>	SW & metadata	5	5,5	CLST	H. Strik
<dem>	SW & docum.	6	3	CLST	H. Strik
<tech>	docum.	6	2	MPI	D. Broeder

In Table 1 the deliverables and milestones are listed. The columns contain information on the identifier, deliverables, deadline of the deliverable (month relative to the start), effort (number of months), participants involved, and responsible person. Below a description is provided of the tasks involved (each of these tasks results in the deliverables mentioned in Table 1).

+ <tech> Technology survey & support for hosting the demonstrator

This task consists of a technical survey and the technical support for the (initial) hosting of the demonstrator. It includes studying how the demonstrator can be made available as webservice. This task will be carried out in close cooperation with an infrastructural specialist that will be assigned by CLARIN-NL. This should also make clear how and to what extent these aspects will be supported by CLARIN-NL, CLARIN, or other means.

The technology survey will be carried out in different stages. At the start of the project, all details regarding the required resources (the packages SPRAAK and LIBSVM, and the corpora JASMIN and CGN) will be checked, details regarding technical issues, IPR, relation with CLARIN infrastructure and standards, etc. These results should also provide the final details regarding the options that are presented in the user survey.

We will look for existing similar applications, such as WebASR [18]. Given the similarities with WebASR (in which audio files are uploaded and the recognition results are returned), we will carefully consider how things are handled in WebASR, regarding the UI (e.g. uploading the data), software used, how to register, login, conditions of use, etc. We have good contacts with our colleagues in Sheffield, with Thomas Hain in particular, and we will inform about their experiences and motivations for the choices made.

At the end of the project a document will be delivered in which experiences, findings, and information on how to comply with current CLARIN-NL and CLARIN standards and desiderata for the CLARIN infrastructure are clearly presented. The outcomes of this survey, and the resulting document, will thus contain valuable information for other CLARIN-NL and CLARIN (related) projects.

+ <user> User survey

For the user survey we will contact researchers from diverse sub-fields of humanities and L&ST. Not only will this provide us with valuable information on common practices, wishes and user preferences, it will also increase the visibility of our project and CLARIN-NL plus CLARIN in general. Users will be asked questions about their research topics and the use of audio and phone transcriptions, the formats and requirements for audio and computer phonetic alphabets, and their opinions about the TQE tool. The resulting deliverable will be a document describing the results of the user survey and the implications for (the development of) the TQE tool.

+ *<core> Developing and evaluating the core component*
The technology underlying the core component has been described above. The deliverables are the software (SW) and the metadata (i.e. TQE measures plus phone boundaries) for selected subsets of JASMIN and CGN. These metadata will first be made publicly available on the participants' websites (i.e. in isolation). Later it will be decided whether they will become part of CGN and JASMIN, in which case the IMDI metadata of these corpora should be adjusted.

+ *<dem> Demonstrator: integration, evaluation, and documentation*
The deliverables are the demonstrator (software), which is made available at the CLARIN center, and the following two documents: (1) user documentation, and (2) API and developer documentation. The user documentation will include a demonstration scenario: a series of screenshots on how to upload the audio and phone transcriptions (PTs), and how to download the resulting TQE measures and a screen-captured movie of an interaction with the UI. The second document will also contain information on how the TQE tool can be adjusted for other tasks (e.g. other computer phonetic alphabets, audio formats, and other languages), which will mainly consist of training acoustic models (for the ASR) for all the symbols present in the computer phonetic alphabet. Most likely, for characterization of he demonstrator we will use the CLARIN Component Metadata Framework (CMDI). Regarding the Persistence Identifier, we will probably make a request using the Persistence Identifier Service provided by CLARIN-NL later this year. All this will also be done in close cooperation with an infrastructural specialist that will be assigned by CLARIN-NL.

8 IPR and Ethical Issues: Risks

The data needed for this project are the corpora CGN [9] and JASMIN [10]). Both corpora are distributed by the HLT agency [7]. The software used are an ASR toolkit, probably SPRAAK [11], and software for SVMs, probably LIBSVM [12]. For both SPRAAK and LIBSVM IPR should not be an issue (see, e.g., http://www.spraak.org/obtaining-spraak and http://www.csie.ntu.edu.tw/~cjlin/libsvm/COPYRIGHT). Still this will be thoroughly checked.

9 Expertise of the applicant(s)

All participants are very familiar with CGN & JASMIN.

The first participant (CLST [6] and the department of Linguistics of the Radboud University Nijmegen) have played a leading role in the development of both CGN and JASMIN (coordinators were N. Oostdijk and C. Cucchiarini, resp.) and these corpora have already been used extensively in our research (e.g., in the Stevin project DISCO, coordinated by H. Strik). CLST also has ample experience with calculating confidence measure (CMs) for phone-segment combinations and the software involved in the current project, as described above [1, 2, 3, 4, 14, 15, 16].

The HLT Agency (TST-Centrale – Centrale voor Taal- en Spraaktechnologie [7]) is the Dutch-Flemish agency for management, maintenance and distribution of Dutch digital language resources, including the corpora CGN and JASMIN. The HLT Agency is located at the Institute for Dutch Lexicology (INL – Instituut voor Nederlandse Lexicologie).

The Max Planck Institute for Psycholinguistics (MPI, [8]) is an institute of the German Max Planck Society, whose mission is to undertake basic research in the psychological and biological foundations of language. MPI offers access to a large amount of linguistic data through the Internet, and is actively involved in CLARIN, CLARIN-NL, and many related initiatives.

10 Project budget details

Table 2. Project budget details

Id.	Participant / Organization	Effort (PM)	Salary Costs/PM (Euro)	Salary Costs (Euro)	Travel & subsistence (Euro)	Total (Euro)
<user>	INL	1	5400	5400	250	**5650**
<core>	CLST	5,5	3400	18700	1375	**20075**
<dem>	CLST	3	3400	10200	750	**10950**
<tech>	MPI	2	5400	10800	500	**11300**
<coord>	CLST	2	5400	10800	500	**11300**
Total		**13,5**		**55900**	**3375**	**59275**

Listed in Table 2 are the project budget details. The first four tasks have already been described in section 7, the last task (<coord>) concerns coordination and management of the project.

11 Literature

Publications

[1] C. Cucchiarini, A. Neri, H. Strik (2009) “Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback”. Speech Communication 51(10) 853-863.
[2] H. Strik, K. Truong, F. de Wet, C. Cucchiarini (2009) “Comparing different approaches for automatic pronunciation error detection”. Speech Communication. 51(10) 845-852.
[3] J. van Doremalen, Catia Cucchiarini, Helmer Strik (submitted) Automatic Detection of Vowel Pronunciation Errors Using Multiple Information Sources. Submitted to the IEEE ASRU2009 Workshop (http://lands.let.ru.nl/~doremalen/publications/Doremalen-EtAl-ASRU2009-PED.pdf)
[4] M. Gubian, B. Schuppler, J. van Doremalen, E. Sanders, L. Boves (2009) Novelty Detection as a Tool for Automatic Detection of Orthographic Transcription Errors. Proc. of 13-th International Conference on Speech and Computer (SPECOM'2009).
[5] C. Van Bael, L. Boves, H. van den Heuvel, H. Strik (2007) Automatic phonetic transcription of large speech corpora. Computer Speech & Language, Volume 21, Issue 4, pp. 652-668.

URLs

[6] CLST, http://www.ru.nl/clst/
[7] HLT Agency, http://www.inl.nl/index.php?option=com_content&task=view&id=448&Itemid=552&lang=en
[8] MPI, http://www.mpi.nl/
[9] CGN: Corpus Gesproken Nederlands (Spoken Dutch Corpus), http://www.inl.nl/index.php?option=com_content&task=view&id=347&Itemid=654
[10] JASMIN: Jongeren, Anderstaligen, Senioren & Mens-machine Interactie in het Nederlands, http://www.inl.nl/index.php?option=com_content&task=view&id=601&Itemid=660
[11] http://www.spraak.org/
[12] http://www.csie.ntu.edu.tw/~cjlin/libsvm
[13] http://taalunieversum.org/taal/technologie/stevin/
[14] http://lands.let.ru.nl/~strik/research/Dutch-CAPT/
[15] http://lands.let.ru.nl/~strik/research/DISCO/
[16] http://lands.let.ru.nl/~strik/research/ST-AAP.html
[17] http://www.spex.nl/
[18] http://www.webasr.com/; http://www.webasr.org/
[19] http://www.phon.ucl.ac.uk/home/sampa/