*Human Language Technologies – The Baltic Perspective*
*A. Tavast et al. (Eds.)*

*doi:10.3233/978-1-61499-133-5-1*

# Multilingual Speech Processing Activities in Quaero: Application to Multimedia Search in Unstructured Data

Lori LAMEL
*CNRS-LIMSI*

**Abstract.** Spoken language processing technologies are principle components in most of the applications being developed as part of the Quaero program. Quaero is a large research and industrial innovation program focusing on the development of technologies for automatic analysis and classification of multimedia and multilingual documents. Concerning speech processing, research aims to substantially improve the state-of-the-art in speech-to-text transcription, speaker diarization and recognition, language recognition, and speech translation.

**Keywords.** Quaero, speech processing, speech-to-text transcription, speaker diarization, speaker recognition, language recognition

## Introduction

This paper provides an overview of the research carried out as part of the Quaero program to improve speech technologies[1]. Quaero is a large research and industrial innovation program focusing on the development of technologies for automatic analysis and classification of multimedia and multilingual documents. The program has two projects devoted to research and common resources led by academic partners, and several (8) application projects led by industrial partners. The core technologies are developed within the Quaero Core Technology Cluster (CTC) project. The main research goal of the CTC is to improve the state-of-the-art in automatic multimedia document structuring for indexing by developing and evaluating the underlying techniques and models. The core technologies are: text processing, translation, audio and speech processing, image and video processing, data protection, cross-modal processing as well as search and navigation methods for multimedia and multilingual documents. Evaluation campaigns have been held annually since the start of Quaero covering more than 30 technologies, including speech-to-text (STT) transcription, speaker diarization, language recognition, spoken language translation, and the detection of specific entities in spoken data.

[1] The interested reader can find many publications in the Research Corner of the Quaero website `www.quaero.org`.

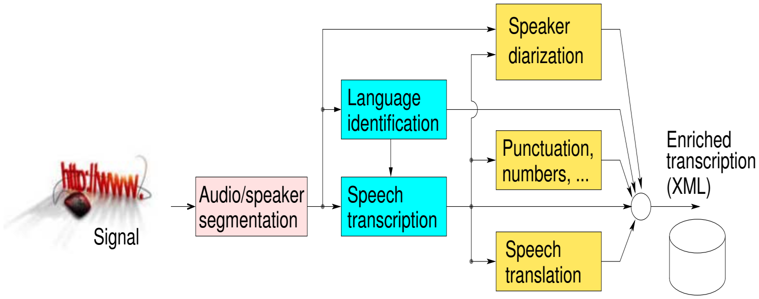


**Figure 1.** Speech technologies in Quaero

Spoken language processing technologies are key components for indexing and searching audio and audiovisual documents as speech is ubiquitous in multimedia data. Having the same underlying written representation as text data (which is not the case for image or video), applications developed for written data can be easily ported to audio data. The program aims to develop technologies that can deal with a wide variety of audiovisual data and to develop applications that can successfully use imperfect technologies.

Figure 1 shows the different speech-related technologies investigated in Quaero. The following sections focus on speech-to-text transcription and speaker diarization, with a brief mention of some other research activities.

## 1. Speech-to-text Transcription

The research in speech-to-text aims are: to significantly reduce the gap between machine and human performances; to develop technology usable for the targeted applications and targeted languages; and to reduce development and porting costs. For this last aspect two main directions have been investigated: finding and exploiting inexpensive data sources and utilization of unsupervised and lightly supervised training methods.

Four partners contribute to the STT task in Quaero: CNRS-LIMSI (www.limsi.fr), KIT (www.kit.edu), RWTH (www-i6.informatik.rwth-aachen.de) and Vocapia Research (www.vocapia.com). At the start of the program the performance of existing speech recognizers for the three primary Quaero languages English, French, German to determine baseline performance on varied multimedia data. The number of languages covered has grown, adding two languages in each phase. In 2011, STT was evaluated for 9 languages: English, French, German, Spanish, Russian, Greek, Polish, with Italian and Portuguese introduced in 2011. The STT evaluations are organized by LNE (Laboratoire National de Métrologie et d'Essais (www.lne.fr)) coordinated by DGA (Délégation Générale pour l'Armement (www.defense.gouv.fr/dga)) who also organizes the machine translation evaluations (from text and from speech). Concerning this latter evaluation, a Rover [1] combination of the best submission from each site is used to compare spoken language translation on automatic and manual transcripts [2].

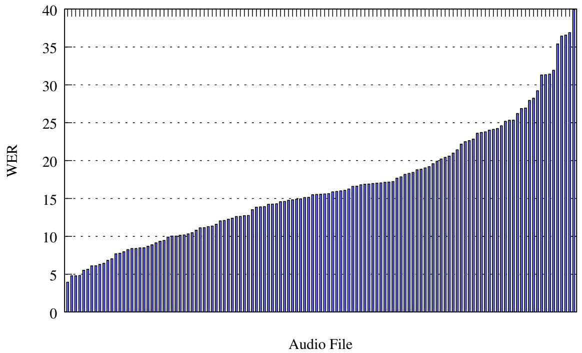
**Table 1.** Summary of results in case-insensitive WER using the 2010 LNE scoring for the 9 languages. The proportion of Broadcast News (BN) and Broadcast Conversation (BC) is specified, and the results on the subsets given in parentheses. †Since the first evaluation for the Italian and Portuguese languages was held in 2011, the 2010 eval column gives the results on the 2011 development data set.

| | 2010 Eval | | 2011 Eval | |
|---|---|---|---|---|
| Language | BN/BC | %WER (BN/BC) | BN/BC | %WER (BN/BC) |
| English | 50/50 | 17.3 (15.4/18.7) | 30/70 | 20.1 (16.3/21.9) |
| French | 50/50 | 19.0 (12.4/21.6) | 30/70 | 15.2 (9.7/17.9) |
| German | 50/50 | 16.9 (12.8/17.8) | 30/70 | 17.4 (13.9/17.8) |
| Russian | 50/50 | 19.2 (17.3/20.2) | 30/70 | 18.6 (20.7/18.2) |
| Spanish | 50/50 | 13.6 (10.1/17.2) | 30/70 | 16.1 (6.2/20.2) |
| Greek | 70/30 | 20.7 (20.7/21.6) | 30/70 | 17.0 (7.8/21.7) |
| Polish | 70/30 | 20.0 (18.0/24.7) | 30/70 | 12.7 (9.9/14.5) |
| Italian† | 50/50 | 22.8 (18.7/27.6) | 50/50 | 18.0 (14.4/21.7) |
| Portuguese† | 50/50 | 28.5 (24.7/32.0) | 50/50 | 22.7 (18.7/26.3) |

The test data are distributed via the LNE via web interface and participants upload their system outputs to the same site, and scoring was done by LNE.

Since the goal of Quaero is to push the forefront of speech recognition technology, few restrictions are imposed. This has the advantage of allowing participants to do whatever they think is best to develop the best technology, but has the disadvantage of confounding algorithmic advances with those due to data collection/selection. In order to better understand such factors, it was decided that participants can use any available data as long as they report what is used and that all training data predates the epoch of the test data, with the exception of any training data provided by Quaero. Audio partitioning (see Section 2) must also be automatic, and one or multiple systems can be used for different data types, but the data type is not side information and must be automatically determined. Although the primary metric is case insensitive WER, from an application point of view there is a preference for case-sensitive STT outputs, and a case-sensitive score is provided as a secondary metric. There were no constraints on processing time which was not specified for most submissions. Therefore, many of the submissions resulted from the combination of several component systems. An exception are the Vocapia submissions which are all close to real-time, with the goal of quick transfer for use by the application projects. All systems make use of state-of-the-art of techniques and comparable results are obtained across sites for mature systems.

In the 2011 evaluation, there was in total over 30 hours of evaluation data, with at least 3 hours per languages. The data are split between Broadcast News (BN) and more varied data including talk shows, debates, Web podcasts collectively called Broadcast Conversation (BC). The data in the 2011 evaluation contain of a larger proportion of broadcast conversation data (70% except for the Italian and Portuguese baselines that have 50%) than in 2010. Broadcast conversation data is more challenging to transcribe as is is much less prepared than news data, which varied acoustic conditions and highly interactive portions.



| Case sen. | Eng | Fre | Ger | Rus | Spa | Gre | Pol | Ita | Por |
|---|---|---|---|---|---|---|---|---|---|
| Best | 7.7 | 8.7 | 8.4 | 12.0 | 4.8 | 4.0 | 11.8 | 8.0 | 11.2 |
| Worst | 45.7 | 28.0 | 36.3 | 25.0 | 31.1 | 51.3 | 39.5 | 26.9 | 39.2 |

**Figure 2.** Distribution of word error rates across all 2011 eval audio documents

The widely adopted Word Error Rate (WER) is used as the evaluation metric[2]. Table 1 summarizes the 2010 and 2011 evaluation results in terms of case-insensitive WER using the LNE scoring tools for the best system for each language. Scoring taking into account case distinctions (not shown in the table) increases the WER by about 1%, This difference is lower for some systems than others, indicating that some models better account for case. The difference also varies by language, which may also reflect inherent difficulties in caseing or ambiguities in the writing conventions.

The overall WER is given along with the WER on the BN and BC subsets. It can be noted that there is a large difference in performance on the two data subsets, with BC data being more difficult with higher error rates. An exception is for Russian where one BN podcast which has a WER over 30%. This audio file has a long segment of reduced bandwidth from a telephone correspondent, recorded in a noisy environment. This highlights the problem in classifying complete audio files as opposed to classifying each segment. Figure 2 shows the distribution of the word error rates across individual audio files for the 2011 evaluation data. The letter under the bar specifies the language. The lowest word error rates for most languages is around or below 10%, and the highest usually over 30%. The range is large since the test data contain a large proportion of interactive and conversational data.

## 2. Speaker Diarization

Speaker diarization, also called speaker segmentation and clustering or 'who spoke when', is the process of partitioning an input audio stream into homogeneous seg-

[2] The word error rate counts the number of errors in an automatic transcript with respect to a reference one, taking into account three types of errors: word deletions (D), insertions (I) and substitutions (S). Is is defined as: $WER = \frac{S+D+I}{N}$ where N is the total number of words in the reference. It should be noted that the WER can be higher than 100%.

**Table 2.** Single-show and cross-show speaker diarization results for interactive podcasts in English, French and German (Quaero 2010 evaluation).

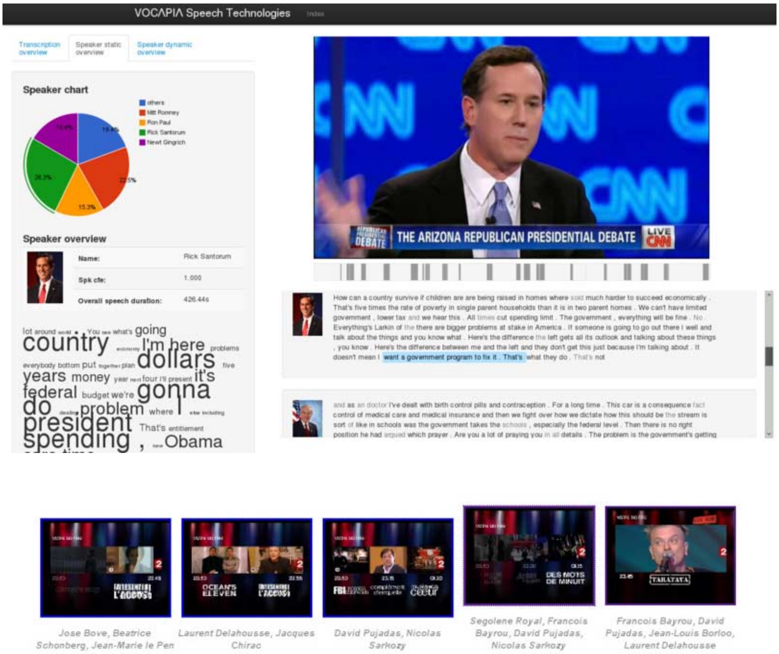
| Language | Single Show | | | | Cross-show | | | |
|---|---|---|---|---|---|---|---|---|
| | Miss | FA | Conf | DER | Miss | FA | Conf | DER |
| English | 0.4 | 0.7 | 10.5 | 11.6 | 0.3 | 0.7 | 21.3 | 22.5 |
| French | 2.0 | 0.4 | 14.0 | 16.4 | 4.1 | 0.8 | 21.5 | 26.7 |
| German | 2.5 | 1.9 | 13.8 | 18.3 | 2.9 | 2.5 | 20.6 | 26.1 |

ments according to speaker identity, without any prior information on the number of speakers or on their voices. Speaker partitioning is a useful preprocessing step for automatic speech recognition since by clustering segments from the same speaker, the amount of data available for unsupervised speaker adaptation is increased. Speaker diarization can also improve the readability of an automatic transcription by structuring the audio stream into speaker turns, in some cases by providing the true speaker identity. For example, in broadcast news programs, the linguistic content often provides the true identities of those taking part in the show. Acoustic and linguistic methods are being explored in Quaero to extract meta-data from the speech signal which are used both to improve transcription performance, and to provide an enriched text output for downstream processing. Multiple sources of information are available to associate true speaker names with speech segments via speaker recognition for a known set of speakers or linguistic information extracted from the transcription.

The first Quaero evaluations assessed acoustic speaker diarization on a per audio file basis, making no a priori assumption on the number of speakers or their voices similar to the diarization tasks. This type of evaluation is similar to that used in the DARPA EARS program and the Technolangue Ester benchmark tests [3,4]. In order to support the application projects, in 2010 the diarization task was extended to the task of cross-show speaker diarization. This consists of assigning the same speaker label (which may or may not be the true speaker name) to the same speaker in multiple audio files. A second extension is the task of famous speaker tracking (or VIP) task, where a data from a known set of speakers has to be retrieved in the test data. This latter task is closely related to that of political speaker timing. During the Presidential election period in France there is close control to ensure equal media access for all candidates. This task is currently carried out primarily by humans, and there is interest in facilitating the work of the operators.

Three partners contribute to the speaker diarization task: KIT, LIMSI and Vocapia. In 2010 and 2011 different approaches for cross-show diarization were compared with a baseline, single-show diarization system: concatenation of all the shows in a batch approach, or an incremental architecture for an online approach. The incremental approach was found to preserve the single-show performance with a limited degradation compared to the batch approach for cross-show diarization. However, cross-show speaker diarization was also found to be sensitive to the show order and the factors of this variability were explored in [5,6].

The primary metric used to measure Speaker Diarization performance is the overall speaker diarization error rate (DER) [7]. This is basically the sum of


VOCAPIA Speech Technologies
Index
Transcription overview
Speaker static overview
Speaker dynamic overview
Speaker chart
others
Mitt Romney
Ron Paul
Rick Santorum
Newt Gingrich
19.4%
22.5%
15.3%
26.3%
16.5%
THE ARIZONA REPUBLICAN PRESIDENTIAL DEBATE
LIVE CNN
lot around world . You see what's going
country economy I'm here problems
everybody bottom put together plan dollars five
years money year next four I'll present it's
federal budget we're gonna
do problem where I else including
president That's entitlement
spending , Obama

Speaker overview

| | |
|---|---|
| Name: | Rick Santorum |
| Spk cfe: | 1.000 |
| Overall speech duration: | 426.44s |

How can a country survive if children are are being raised in homes where sold much harder to succeed economically . That's five times the rate of poverty in single parent households than it is in two parent homes . We can't have limited government , lower tax and we hear this . All times cut spending limit . The government , everything will be fine . No . Everything's Larkin of the there are bigger problems at stake in America . If someone is going to go out there I well and talk about the things and you know what . Here's the difference the left gets all its outlook and talking about these things , you know . Here's the difference between me and the left and they don't get this just because I'm talking about . It doesn't mean I want a government program to fix it . That's what they do . That's not

and as an doctor I've dealt with birth control pills and contraception . For a long time . This car is a consequence fact control of medical care and medical insurance and then we fight over how we dictate how this should be the stream is sort of like in schools was the government takes the schools , especially the federal level . Then there is no right position he had argued which prayer . Are you a lot of praying you in all details . The problem is the government's getting

Jose Bove, Beatrice Schonberg, Jean-Marie le Pen

Laurent Delahousse, Jacques Chirac

David Pujadas, Nicolas Sarkozy

Segolene Royal, Francois Bayrou, David Pujadas, Nicolas Sarkozy

Francois Bayrou, David Pujadas, Jean-Louis Borloo, Laurent Delahousse

**Figure 3.** Speaker Indexing in Video demonstration (top) and Gallery of indexed video (bottom)

the missed time (M), false alarm time (FA) and speaker confusion time (SC), normalized by the total amount of scored speech: $DER = \frac{M+FS+SC}{scored\_speaker\_time}$

Table 2 gives results of single- and cross-show speaker diarization for the 2010 evaluation for interactive podcasts. The test data for English come from *Naked Scientist Show* (47 shows, 20 hours), the French from the talk show *Ce soir ou jamais* (44 shows, 16 hours) and for German from a variety of TV shows and podcasts (23 shows, 7 hours). It can be seen that the cross-show DER is 1.5 to 2 times that of the single show DER.

The 2011 Speaker Diarization evaluation was based on 34 files (10 hours) in English from the *Naked Scientist Show*. The single- and cross-show DERs were about 3% and 7% respectively. This lower error can in part be attributed to system improvements targeting closely matched development data and to the homogeneity of the test data. Future Speaker Diarization evaluations for Quaero will be carried out in coordination with the Repere challenge [8] which aims to support research on person recognition in multimodal data.

A demonstration was developed to illustrate "Speaker Indexing in Video" as illustrated by the interface shown in upper part of Figure 3. The contents of a video corpus was processed using Speech-To-Text and Speaker Identification modules to automatically transcribe the spoken words and identify the speakers in the video. All speakers present in the video are listed with their photo if one of the known speakers. The transcriptions are converted into subtitles overlaid on the the video. Other information can be optionally displayed (time-codes, confidence measures, speaker gender, ...). A gallery of selected videos is shown in the lower part of the figure, with a thumbnail and a list of known speakers appearing in each video.

## 3. Other Speech Technology Related Activities

There are a number of other activities being pursued, aimed at improving speech technology or supporting studies to provide general knowledge about spoken language in large corpora.

Language identification (that is identifying the language and/or dialect of an audio document) is also being studied under Quaero. The language identification module identifies the language(s) spoken in the audio document, relying on models estimated on representative data from each language known to the system. Different modeling approaches are being explored, including acoustic, phonotactic and lexical-based models as well as combinations of these. Phonotactic models have the advantage of being easy to build, do not require transcribed audio data for training and obtain reasonable performance. Lexical models are more accurate, but require that an STT system exist for the concerned languages. In order to be able to deal with multilingual documents, in future work the language labels will be provided for segments located by the audio partitioner.

The work in Quaero has supported perceptual and linguistic studies. Concerning the first, it is widely acknowledged that humans listeners significantly outperform machines when it comes to transcribing speech. Bridging the gap between humans and machines by taking advantage of the perceptual strategies is gaining more attention. An underlying hypothesis motivating these studies [9] at LIMSI is that ambiguities may result from simplified models (model bias), or be due to intrinsic acoustic or contextual confusability of speech regions. The role of *context length* has been in particular investigated through perceptual recovery of small homophones involved in frequent automatic transcription errors in both French and English. Concerning the second, there has been growing interest in the use of speech technologies as an aid for acoustic-phonetic, linguistic and sociolinguistic studies. Semi-automated approaches enable researchers to study, propose and validate phenomena on very large corpora [10,11].

Since one aim is to produce speech processing results which are both easily searchable by a machine and can be easily read by a human. For the latter, number conversion (amounts, dates, measures) and reliable punctuation are needed. Ongoing studies aim to develop algorithms to identify punctuation (periods, question marks, commas, etc.) and disfluency markers, using a combination of language and acoustic/prosodic models (features such as pitch contours, duration, energy, pause lengths, etc) [12]. Keeping the recognition vocabulary up-to-date is also very important for search in breaking news, therefore experiments assess the use of web-texts for daily language model updates.

## 4. Conclusions and Perspectives

The processing of so-called ‘found data’, that is data produced for other purposes than to be transcribed by a machine, was a pivotal change in speech recognition research. Research, initially focused on broadcast news data [13], has expanded to more diverse sources of broadcast audio (talk shows, debates, radio call-in shows) as well as personal data posted on the Internet (Pod-casts) for which are much more challenging for todays technology.

Speech technologies are central components for automatic processing of audio and audiovisual documents, and is one of the central research topics in the Quaero program, serving for several application projects: Multimedia search, Media Monitoring and social impact, Video and music access on web portals, Lecture translation. Links to some demos can be found on the Quaero web site.

In the context of the Quaero program, speech technology research has thus far addressed 9 European languages, with plans to cover all 23 official European languages by the end the program. The upcoming 2012 speech transcription evaluation will include 8 additional languages (Bulgarian, Czech, Estonian, Hungarian, Latvian, Luxembourgish, Romanian, Slovak).

## Acknowledgments

The author acknowledges the contribution of colleagues at LIMSI and the partners contributing to the development of speech technologies and their evaluation to the work reported in this paper. This work has been partly financed by OSEO, the French State Agency for Innovation, under the Quaero program.

## References

[1] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," *IEEE ASRU*, Santa Barbara,1997.

[2] L. Lamel et al, " Speech Recognition for Machine Translation in Quaero," *IWSLT 2011*, San Francisco, 2011.

[3] C. Barras, X. Zhu et al., "Multistage speaker diarization of broadcast news," *IEEE Transactions on Audio, Speech & Language Processing*, **14**:150–1512, 2006.

[4] S. Galliano, G. Gravier, L. Chaubard, "The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts," *ISCA InterSpeech*, Brighton, 2009.

[5] V.A. Tran, V.B. Le, et al,."Comparing Multi-Stage Approaches for Cross-Show Speaker Diarization," *ISCA InterSpeech*, Florence 2011.

[6] Q. Yang, Q. Jin, T. Schultz, "Investigation of Cross-show Speaker Diarization", *ISCA InterSpeech*, Florence, 2011.

[7] http://www.itl.nist.gov/iad/mig//tests/rt/2003-spring/docs/rt03-spring-eval-plan-v4.pdf

[8] A. Giraudel, M. Carr, et al., "The REPERE Corpus: a multimodal corpus for person recognition," *LREC*, Istanbul, 2012.

[9] I. Vasilescu, D. Yahia, et al., "Cross-lingual study of ASR errors: on the role of the context in human perception of near homophones," *InterSpeech*, Florence, 2011.

[10] M. Adda-Decker, E. Delais-Roussarie, et al., "La liaison dans la parole spontane familire: explorations semi-automatiques de grands corpus," *JEP*, Grenoble, 2012.

[11] M. Adda-Decker, M. Candea, L. Lamel, "Recent evolution of some non standard variants in French broadcast news", *Sociolinguistics Symposium 19*, Berlin, 2012, to appear.

[12] J. Kolář, L. Lamel, "Development and evaluation of automatic punctuation for French and English speech-to-text," *ISCA InterSpeech*, Portland, 2012, to appear.

[13] Speech Communication Special issue on automatic transcription of broadcast news data, **37**(1-2), May 2002