Speech Recognition for Machine Translation in Quaero

Lori Lamel, Sandrine Courcinois, Julien Despres, Jean-Luc Gauvain, Yvan Josse, Kevin Kilgour, Florian Kraft, Viet Bac Le, Hermann Ney, Markus Nußbaum-Thom, Ilya Oparin, Tim Schlippe, Ralf Schlüter, Tanja Schultz, Thiago Fraga da Silva, Sebastian Stüker, Martin Sundermeyer, Bianca Vieru, Ngoc Thang Vu, Alexander Waibel, Cécile Woehrling
Outline

• Brief presentation of the Quaero Program
• Speech processing in Quaero
• Quaero 2010 STT evaluation for French and German
• Results of Quaero 2011 STT evaluation
**Introduction**

- A program on multilingual and multimodal document processing
- Connecting research and innovation through data and evaluation
- *funded by OSEO, French innovation agency*
- Technologies to analyze and organize multimedia, multilingual documents
- Quaero in numbers:
  - 5 year program, starting May 2008
  - about 200 ME budget, 99 ME
  - Small and large companies and academics (French and German)
Quaero Partners

- 27 partners
- Application projects (Technicolor, France-Telecom, Jouve, Exalead, Yacast, ...)
- A shared research project
- A corpus project (data collection and annotation, evaluation data)
Information on the Web continues to grow rapidly but is mostly unstructured (ex. scanned books, audio, video)

We can store everything but don’t really know how to access it

- Need solutions to organize and search unstructured data
- Multilingual and multimedia context
  (text, speech, music, image, video)

- Quaero approach
  - Statistical methods for all modeling and decision problems
  - Application driven
  - All technologies evaluated annually
Main Research Objectives

Improve state-of-the art of technologies for automatic processing of multimedia/multilingual documents

• Text, Speech: Question answering, speech recognition, language recognition, translation, semantic annotation
• Music: Music genre and mood identification, source separation, fingerprinting
• Image: Image identification (eg. face, object, adult content, ...), image clustering
• Video: Segmentation, person and object tracking, event detection, motion recognition
• Search engine: Searching multimedia data, search by similarity (image, music, ...), content recommendation
Main Research Objectives

Improve state-of-the art of technologies for automatic processing of multimedia/multilingual documents

- Text, Speech: Question answering, speech recognition, language recognition, translation, semantic annotation
- Music: Music genre and mood identification, source separation, fingerprinting
- Image: Image identification (eg. face, object, adult content, ...), image clustering
- Video: Segmentation, person and object tracking, event detection, motion recognition
- Search engine: Searching multimedia data, search by similarity (image, music, ...), content recommendation
- Application and evaluation driven technology development
QUAERO TECHNOLOGIES

Structuring & managing multimedia & multimodal documents

- Translation
- Image Processing
- Speech transcription
- Audio & Music
- Video Processing
- Multimodal Structuring
- Coding Protection
- Information Retrieval Navigation
- Written Language Processing

Corpus Evaluation Demonstrators
Example Quaero Applications

- **Voxalead News** (news search engine) [voxaleadnews.labs.exalead.com]
  - keyword search in speech transcripts (content-based search)
  - named entity detection (people, organisations, locations)
- **Music Mashup** (music search engine) [muma.labs.exalead.com]
  - keyword search (artist name, song lyrics, ...)
  - search by sequence of chords, genre, mood
- **Media monitoring** [yacast.fr]
  [http://viphttpplayers.yacast.net/V4/fmpremium/fmpremium.html]
  - Audio and video fingerprinting to identify advertising and music
  - Automatic speaking time measure (for politicians), ASR, archive
- and more: real-time lecture translation, Audiobook and e-book synchronisation, France 24 HD Player, Presidency web site ...
Speech Processing Technologies

- Spoken language processing technologies are key components for indexing and searching audio and audiovisual documents
- Speech is ubiquitous in multimedia data
- Underlying written representation (lacking for image and video)
Spoken language processing technologies are key components for indexing and searching audio and audiovisual documents.

Speech is ubiquitous in multimedia data.

Underlying written representation (lacking for image and video).

Developing core speech processing technologies:
  - Speech-to-text transcription (KIT, LIMSI, RWTH, Vocapia Research)
  - Speaker diarization (KIT, LIMSI, Vocapia Research)
  - Language identification (LIMSI)

Evaluations run by LNE/DGA.
Speech Processing Technologies

- Spoken language processing technologies are key components for indexing and searching audio and audiovisual documents
- Speech is ubiquitous in multimedia data
- Underlying written representation (lacking for image and video)
- Developing core speech processing technologies
  - Speech-to-text transcription (KIT, LIMSI, RWTH, Vocapia Research)
  - Speaker diarization (KIT, LIMSI, Vocapia Research)
  - Language identification (LIMSI)
- Evaluations run by LNE/DGA
- Develop technology usable for targeted applications and languages
- Applications: audiovisual media analysis, media monitoring (radio, TV), audiovisual archive indexing, captioning, speech analytics, ...
 Speech Technologies

Signal

Audio/speaker segmentation

Language identification

Speech transcription

Speaker diarization

Emotions

Punctuation, num., topics

Speech translation

Enriched transcription (XML)
Speech Processing Technologies

- Speech-to-text transcription
  - Main Quaero languages: English, French, German
  - Progressive increase in languages: 9 languages in 2011
  - Plan to cover all European languages: only dev/test data for some languages
Speech Processing Technologies

- Speech-to-text transcription
  - Main Quaero languages: English, French, German
  - Progressive increase in languages: 9 languages in 2011
  - Plan to cover all European languages: only dev/test data for some languages

- Speaker diarization
  - “Who spoke when”: speaker segmentation and clustering
  - Preprocessing for ASR and enriched transcription
  - Political Speaker Tracking task
  - Cross-show Speaker Diarization
QUAERO 2010 STT Evaluation

- STT assessed for 7 languages
  - Primary Quaero languages: English, French, German
  - Second evaluation for Russian & Spanish
  - Baseline evaluation for Greek & Polish
- Training data distributed in April for all languages
- 3 hours of development data per language
- 35 STT submissions (from 4 sites)
QUAERO 2010 STT Evaluation

- STT assessed for 7 languages
  - Primary Quaero languages: English, French, German
  - Second evaluation for Russian & Spanish
  - Baseline evaluation for Greek & Polish
- Training data distributed in April for all languages
- 3 hours of development data per language
- 35 STT submissions (from 4 sites)
- Evaluation guidelines
- Metrics: CI and CS word error rate, LNE scoring tools
- Total of 24 hours of test data in 2010
- Roughly 50% broadcast news, 50% broadcast conversation/podcasts
**Evaluation Conditions**

- Can use any available data, but must say what is used.
- Only restriction is that training data predates Mar 2010 with the exception of any Quaero training data.
- Automatic segmentation.
- Can use one or multiple systems for different data types, but type is not side information.
- Primary metric: case insensitive WER.
- Preference for case-sensitive output, score both case-sensitive and case-insensitive.
- No constraints on processing time (not specified for most submissions).
## Quaero 2010 evaluation data

<table>
<thead>
<tr>
<th>Language</th>
<th>Broadcast News</th>
<th>Broadcast Conversation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>0:53</td>
<td>2:07</td>
<td>3:00</td>
</tr>
<tr>
<td>German</td>
<td>1:20</td>
<td>2:14</td>
<td>3:34</td>
</tr>
</tbody>
</table>
### 2010 Results for French and German

<table>
<thead>
<tr>
<th></th>
<th>Case</th>
<th>No Case</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>French</td>
<td>German</td>
</tr>
<tr>
<td>KIT</td>
<td>28.4</td>
<td>25.2</td>
</tr>
<tr>
<td>RWTH</td>
<td>23.7</td>
<td>21.2</td>
</tr>
<tr>
<td>VR</td>
<td>21.4</td>
<td>22.2</td>
</tr>
<tr>
<td>LIMSI+VR</td>
<td>19.9</td>
<td>19.0</td>
</tr>
</tbody>
</table>

- About 1% absolute difference for most languages when case is scored
- Case is important for SMT
- Large differences across shows (French 5.7-40.3, German 9.9-22.8)
- All sites obtained large absolute WER reductions for all languages
**System Components**

- Acoustic models: HMMs (10-20M parameters), allophones (triphones), discriminative features, discriminative training
- Pronunciation dictionary, pronunciation probabilities
- Language models: statistical N-gram models (10-20M N-grams), model interpolation, connectionist LMs, text normalization
- Decoder: multipass decoding, unsupervised acoustic model adaptation, system combination (Rover, cross-system adaptation)
## Audio Data used by sites

<table>
<thead>
<tr>
<th>Site</th>
<th>French</th>
<th>German</th>
</tr>
</thead>
<tbody>
<tr>
<td>KIT</td>
<td>382</td>
<td>352</td>
</tr>
<tr>
<td>RWTH</td>
<td>230</td>
<td>50 Q</td>
</tr>
<tr>
<td>VR/LIMSI</td>
<td>311</td>
<td>80</td>
</tr>
</tbody>
</table>

- Quaero (50-100 hours per language)
- French: BREF, ESTER, EPPS, + misc
- German: BN, GlobPhon, Verbmobil, LBW, WDR, Mainz, Zeit
Recognition Lexicons

<table>
<thead>
<tr>
<th>Site</th>
<th>Vocabulary Size</th>
<th>#Phones</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>French</td>
<td>German</td>
</tr>
<tr>
<td>KIT</td>
<td>170k</td>
<td>300k</td>
</tr>
<tr>
<td>RWTH</td>
<td>200k</td>
<td>300k</td>
</tr>
<tr>
<td>VR/LIMSI</td>
<td>65k/200k</td>
<td>300k</td>
</tr>
</tbody>
</table>

- Vocabulary selection: frequency cut-off, unigram interpolation
- OOV: 0.5-1.0% or French, 1.3-2.6% for German
- 1.5 - 2.0 pronunciations per word, pseudo phones for silence and non-speech
- Complete available internal or online dictionaries (Wiktionary)
- Different methods for pronunciation generation: Rule-based g2p, statistical methods
- Pronunciation variants for liaison handling
## Language models

<table>
<thead>
<tr>
<th>Site</th>
<th>Training texts</th>
<th>Dev10 ppx/oov(%)</th>
<th>Site</th>
<th>Training texts</th>
<th>Dev10 ppx/oov(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>French</td>
<td>German</td>
<td>French</td>
<td>German</td>
<td></td>
</tr>
<tr>
<td>KIT</td>
<td>980M</td>
<td>1.6B</td>
<td>181 / 0.5</td>
<td>260 / 2.4</td>
<td></td>
</tr>
<tr>
<td>RWTH</td>
<td>1.3B</td>
<td>580M</td>
<td>131 / 0.5</td>
<td>269 / 1.1</td>
<td></td>
</tr>
<tr>
<td>VR/LIMSI</td>
<td>2.2B</td>
<td>1.3B</td>
<td>153 / 1.1</td>
<td>261 / -</td>
<td></td>
</tr>
</tbody>
</table>

- Assorted newspaper and newswires, commercial transcripts, EPPS, Webtexts
- Several million to billions of words
- Interpolation of component LMs
- Large models: typically over 400M 4-grams
- Audio transcripts: several 100k to several million words
- Perplexity: 130-150 (French), 260 (German)
Multi-pass recognition setup

- **Pass1**: VTLN trained ML model
- **Pass2**: SAT/CMLLR + MLLR adaptation
- **Pass3**: Cross-adaptation (SAT/MLLR)
- **Pass4**: Full language model rescoring
- **Pass5**: System combination (CNC)
LIMSI/Vocapia Decoding

- Partitioner: speech/non-speech segmentation and speaker clustering
- French system: Two pass decoding with 2 chains
  - Lattice generation and consensus decoding using 4-gram rescoring and pronunciation probabilities
  - Unsupervised
  - Pass1: plp+f0 AMs (SAT, MMI trained)
  - Pass2a: mlp+plp+f0 AMs (MMI trained), 64k LM
  - Pass2b: mlp+plp+f0 AMs (MMI trained), 200k LM, NN LM rescoring in 2nd pass
- German system
  - plp+f0 AMs (SAT, ML trained)
  - Single pass recognition generating lattices
  - Consensus decoding using 4-gram rescoring and pronunciation probabilities
Some System Characteristics

- Data partitioning (segmentation and clustering)
- Cepstral features, mean/variance normalization, (H)LDA
- Pitch/voicing features
- Vocal tract normalization
- Multi-layer perceptron based features
- Warped minimum variance distortionless response
- Speaker adaptive training (SAT/CMLLR), discriminative training (MMIE, MPE)
- Contextual dependent phone models
- N-gram language models, Neural net language model
- Single-pass or multi-pass decoding with unsupervised adaptation
- System combination (cross adaptation, Rover, CNC)
System Combination for SMT

- Done by KIT
- Single-best output from the three sites were combined using Recognizer Output Voting Error Reduction (ROVER)
- Alignment using word timing information in CTM files
- Parameter settings determined empirically on 2010 dev data
- For German best results obtained with a majority vote
- For French the best results obtained by voting based on maximum confidence
### Rover Results

<table>
<thead>
<tr>
<th>Language</th>
<th>dev2010 [%WER]</th>
<th>eval2010 [%WER]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ci</td>
<td>cs</td>
</tr>
<tr>
<td>French</td>
<td>18.1</td>
<td>19.2</td>
</tr>
<tr>
<td>German</td>
<td>17.0</td>
<td>18.0</td>
</tr>
</tbody>
</table>

- For German, case-insensitive ROVER no gain over best system, case-sensitive gain of 14.6% relative
- For French 5.8% relative improvement for case-insensitive, case-sensitive gain 17.4%
## Quaero 2011 STT Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BN/BC</td>
<td>WER (%)</td>
</tr>
<tr>
<td>English</td>
<td>50/50</td>
<td>17.3</td>
</tr>
<tr>
<td>French</td>
<td>50/50</td>
<td>19.0</td>
</tr>
<tr>
<td>German</td>
<td>50/50</td>
<td>16.9</td>
</tr>
<tr>
<td>Russian</td>
<td>50/50</td>
<td>19.2</td>
</tr>
<tr>
<td>Spanish</td>
<td>50/50</td>
<td>13.6</td>
</tr>
<tr>
<td>Greek</td>
<td>70/30</td>
<td>20.7</td>
</tr>
<tr>
<td>Polish</td>
<td>70/30</td>
<td>20.0</td>
</tr>
<tr>
<td>Italian</td>
<td>50/50</td>
<td>22.8*</td>
</tr>
<tr>
<td>Portuguese</td>
<td>50/50</td>
<td>28.5*</td>
</tr>
</tbody>
</table>
Summary

• Many research directions for ASR
• Evaluations of speech technologies (Quaero and external)
• Comparable results across sites for mature systems
• Manual and automatic punctuation annotation
• Reduce gap between machine and human performances
• Study of ASR errors & human perceptual experiments
• Cross-show Speaker Diarization and Political Speaker Tracking
• Annotation of metadata (speaker, language, topic, emotion, style ...)
• Model adaptation: keeping models up-to-date
• Alternative combination strategies?
• Contrast SMT with best system?
Music Fingerprinting (Yacast)
FACE TRACKING (KIT)
Image Search by Similarity (LTU)
Event Detection (Technicolor)

Violent scene detection

Soccer events
Question Answering (LIMSI)

Documents database

Preprocessing and indexing

Document index

Extracted documents

Snippets

Documents extraction

Question database

Question analysis

Question traits

Documents processing

Answer extraction

Answers
Lecture Translator (KIT)