

SPEECH RECOGNITION FOR MACHINE TRANSLATION IN QUAERO

*Lori Lamel, Sandrine Courcinous, 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)

QUAERO VISION

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

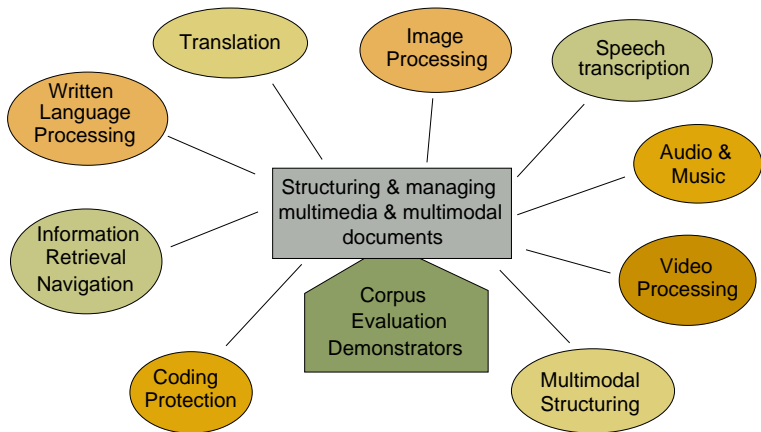
- Text, Speech: Question answering, speech recognition, language recognition, translation, semantic annotation
- Music: Music genre and mood identification, source separation, fingerprinting
- Image: Image identification (eg. face, object, adult content, ...), image clustering
- Video: Segmentation, 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



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)

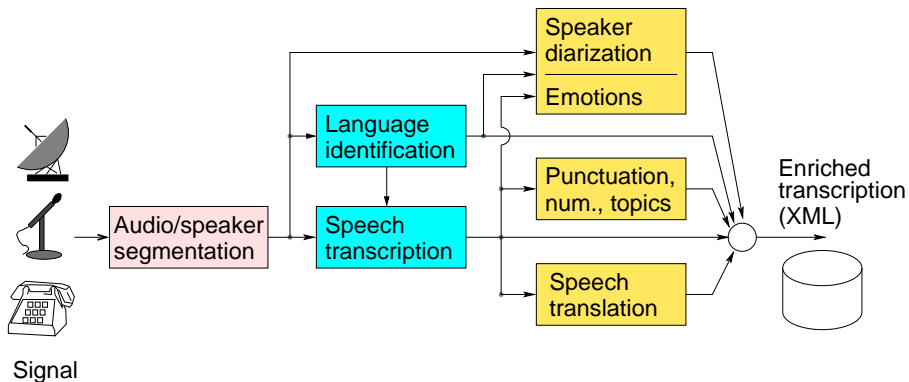
SPEECH PROCESSING TECHNOLOGIES

- 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

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



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

Language	Broadcast News	Broadcast Conversation	Total
French	0:53	2:07	3:00
German	1:20	2:14	3:34

2010 RESULTS FOR FRENCH AND GERMAN

Word Error Rate (%)

	Case		No Case	
	French	German	French	German
KIT	28.4	25.2	27.7	24.1
RWTH	23.7	21.2	22.5	16.9
VR	21.4	22.2	20.6	21.0
LIMSI+VR	19.9		19.0	

- 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

Site	French	German
KIT	382	352
RWTH	230	50 Q
VR/LIMSI	311	80

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

RECOGNITION LEXICONS

Site	Vocabulary Size		#Phones	
	French	German	French	German
KIT	170k	300k	38/35	52
RWTH	200k	300k	44	51
VR/LIMSI	65k/200k	300k	35	49

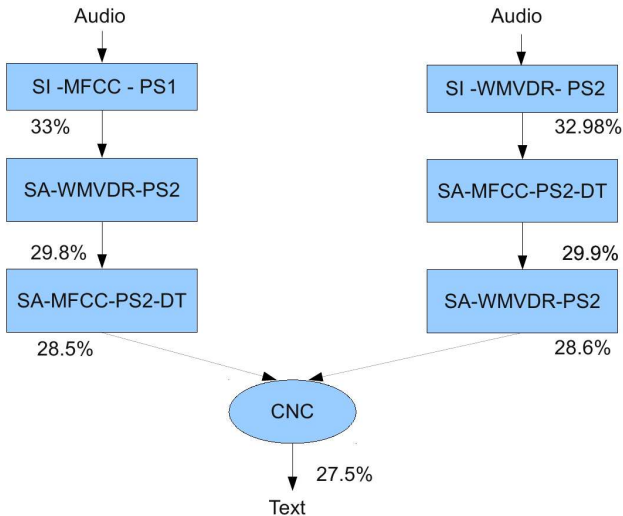
- Vocabulary selection: frequency cut-off, unigram interpolation
- OOV: 0.5-1.0% for French, 1.3-2.6% for German
- 1.5 - 2.0 pronunciations per word, pseudo phones for silence and non-speech
- Complete available internal or online dictionaries (Wiktionary)
- Different methods for pronunciation generation: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

LANGUAGE MODELS

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

- 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)

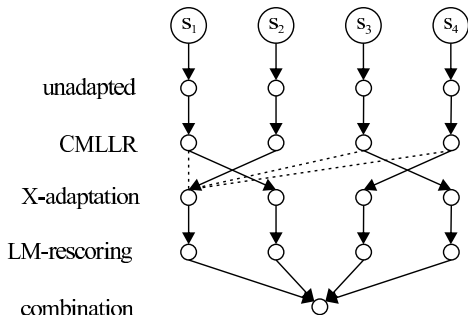
KIT FRENCH DECODING



RWTH DECODING

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 rescoreing 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 rescoreing in 2nd pass
- German system
 - plp+f0 AMs (SAT, ML trained)
 - Single pass recognition generating lattices
 - Consensus decoding using 4-gram rescoreing 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

Language	dev2010 [%WER]		eval2010 [%WER]	
	ci	cs	ci	cs
French	18.1	19.2	19.6	20.8
German	17.0	18.0	17.0	18.1

- 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

<i>Language</i>	<i>P3 Eval (2010)</i>		<i>P4 Eval (2011)</i>	
	<i>BN/BC</i>	<i>WER (%)</i>	<i>BN/BC</i>	<i>WER (%)</i>
English	50/50	17.3	30/70	19.8
French	50/50	19.0	30/70	14.9
German	50/50	16.9	30/70	17.4
Russian	50/50	19.2	30/70	18.3
Spanish	50/50	13.6	30/70	15.9
Greek	70/30	20.7	30/70	16.9
Polish	70/30	20.0	30/70	12.5
Italian	50/50	22.8*	50/50	18.0
Portuguese	50/50	28.5*	50/50	22.7

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)

The screenshot displays the Yacast web interface for FM Premium. At the top left, the "FM Premium" logo is visible. On the right, it says "Powered by yacast". The main interface features a control bar at the top with a play button, volume sliders, and a search box labeled "Recherche". Below this is a grid of radio station cards, each with a microphone icon and the text "Animateur". The stations shown include:

- Inter**: Animateur
- BFM** (LA RADIO DE L'ÉCOLE): Animateur
- Europe 1**: Animateur
- RFM**: Francis Cabrel, La cabane du pêcheur
- Info**: Animateur
- RTL**: Jean-Jacques Goldman, Bonne idée
- Chérie**: Whitney Houston, I wanna dance with so...
- RTL**: Animateur
- RMC** (INFO TALKS SPORT): Animateur
- NO 7 MUSIC**: Mungo Jerry, In the summertime
- fun radio**: Animateur

At the bottom, a playback control bar shows the current time as 00:00 and the total duration as 01:12 01:00. A "Retour à direct" button is located on the right side of the bar.

FACE TRACKING (KIT)

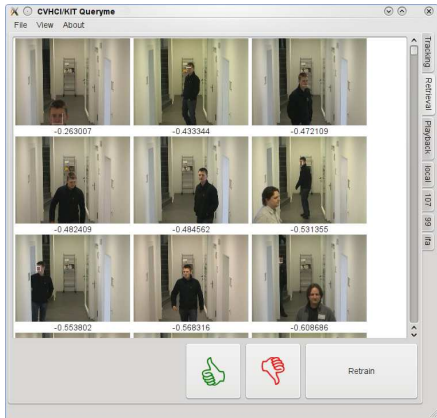
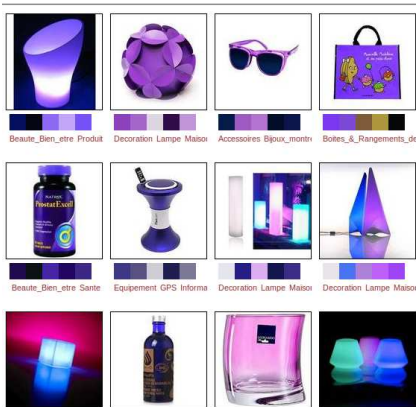


IMAGE SEARCH BY SIMILARITY (LTU)

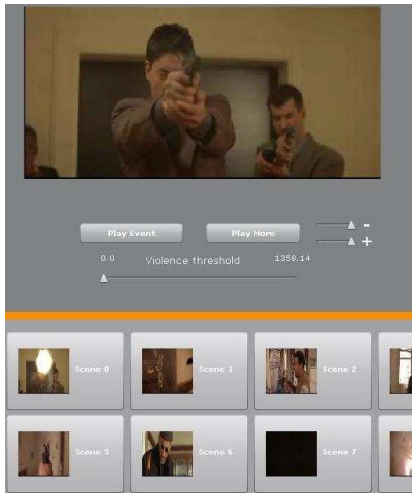


Dynamically Computed Palette



EVENT DETECTION (TECHNICOLOR)

Violent scene detection



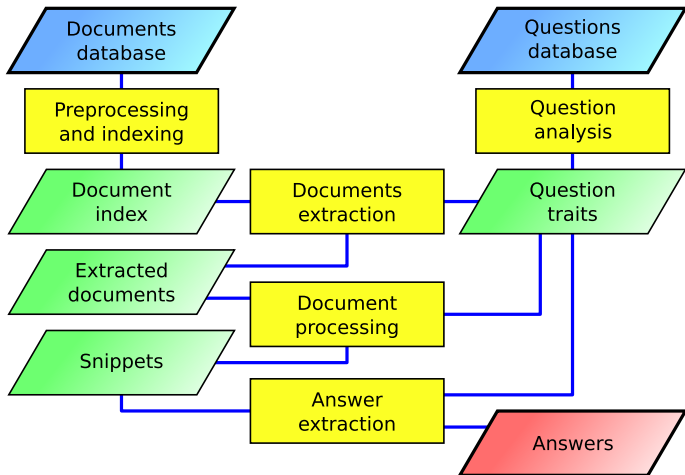
The interface for violent scene detection features a large video player at the top showing a man pointing a gun. Below the player are two buttons: "Play Event" and "Play Here". A "Violence threshold" slider is positioned below these buttons, with a value of 0.0 on the left and 1358.14 on the right. Below the slider is a grid of scene thumbnails labeled "Scene 0" through "Scene 7".

Soccer events



The interface for soccer events displays a grid of 23 scene thumbnails, each labeled from "Scene 0" to "Scene 22". Each thumbnail shows a different view of a soccer game in progress.

QUESTION ANSWERING (LIMSI)



LECTURE TRANSLATOR (KIT)

