Speech Processing Technologies in Quaero: application to multimedia search in unstructured data

Lori Lamel

lamel@limsi.fr

HLT 2012
October 5, Tartu

Overview

- Quaero vision and objectives
- Organizing unstructured data, multimedia & multilingual search (text, audio, music, image, video)
- Speech Processing Technologies
- Demos
- Conclusions and outstanding challenges

The Quaero Program

- 27 partners
- Technologies to organize multimedia and multlingual contents
- 6 application projects
  (Technicolor, France Telecom, Jouve, Exalead, Yacast, ...)
- A shared research project: core technology cluster
- A corpus project (data collection and annotation, evaluation data)

Why Quaero?

- Quaero initiative dates from 2004 (project started in 2008)
- What was the Web like in 2004?
  - no Youtube, no Facebook, no Twitter, no Google Books
  - multimedia content was much more limited
    (ex. heavily compressed video)
- Even today most of the information is unstructured (ex. scanned books, audio, video)
- We can store everything but don't really know how to access it

Quaero vision

- 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

Main Research Objectives

Improve state-of-the art in 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

Main Research Objectives (cont)

Video

Segmentation, person and object tracking, event detection, motion recognition

Search engine

Searching multimedia data, search by similarity (image, music, ...), content recommendation

Some Quaero Challenges

- Processing all styles of data (professional and amateur documents)
- Covering most European languages

Quiz: How many languages in EU?

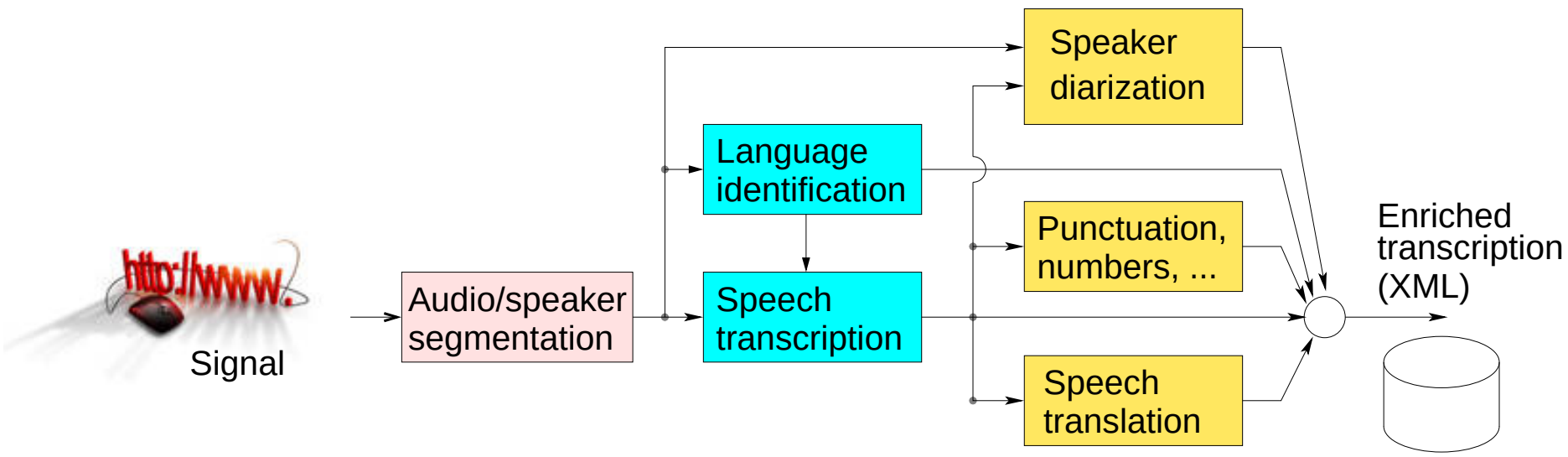
23 32 60

- Giving real answers to questions
- Reducing the gap between human and machine performance
- Creating successful applications using imperfect technologies

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
- Reduce gap between machine and human performances
- Develop technology usable for targeted applications and languages
- Reduce development and porting costs
- Applications: audiovisual media analysis, media monitoring (radio, TV), audiovisual archive indexing, captioning, speech analytics, ...

Speech Processing Technologies (2)



Speech Processing Technologies (3)

- Speech-to-text transcription (STT)
  - KIT, LIMSI, RWTH, Vocapia Research
  - Main Quaero languages: English, French, German
  - Progessive increase in languages: assessed for 9 languages
  - Cover all European languages (plus Arabic and Mandarin)
- Speaker diarization
  - KIT, LIMSI, Vocapia Research
  - “Who spoke when”: speaker segmentation and clustering
  - Preprocessing for ASR and enriched transcription
  - Political Speaker Tracking task
  - Cross-show Speaker Diarization
- Language Identification
  - LIMSI

Extracting Information from speech


```
<?xml version="1.0" encoding="UTF-8"?>
<AudioDoc name="doc2" path="doc2.wav">


<SpeakerList>
<Speaker dur="33.36" gender="1" spkid="MS1" name="Tony Khan"/>
<Speaker dur="38.58" gender="2" spkid="FS2"/>
</SpeakerList>


<SegmentList>
<SpeechSegment stime="0.50" etime="2.09" spkid="MS1" lang="eng-usa">
<Word stime="0.80" dur="0.39" conf="0.971"> This </Word>
<Word stime="1.46" dur="0.13" conf="0.971"> is </Word>
<Word stime="1.59" dur="0.10" conf="0.971"> the </Word>
<Word stime="1.69" dur="0.36" conf="0.971"> world </Word>
<Word stime="2.05" dur="0.00" conf="0.594"> , </Word>
</SpeechSegment>


<SpeechSegment stime="4.12" etime="9.13" spkid="MS1" lang="eng-usa">
<Word stime="4.12" dur="0.12" conf="0.934"> a </Word>
<Word stime="4.27" dur="0.12" conf="0.934"> co </Word>
<Word stime="4.49" dur="0.38" conf="0.934"> production </Word>
<Word stime="4.87" dur="0.08" conf="0.934"> of </Word>
<Word stime="4.97" dur="0.08" conf="0.934"> the </Word>
<Word stime="5.15" dur="0.39" conf="0.934"> BBC </Word>
<Word stime="5.54" dur="0.26" conf="0.926"> world </Word>
<Word stime="5.80" dur="0.51" conf="0.444"> Service </Word>
<Word stime="6.47" dur="0.66" conf="0.568"> PRI </Word>
<Word stime="7.48" dur="0.14" conf="0.917"> and </Word>
</SpeechSegment>
...
</AudioDoc>
```

Extracting Information from speech


```
<?xml version="1.0" encoding="UTF-8"?>
<AudioDoc name="doc2" path="doc2.wav">


<SpeakerList>
<Speaker dur="33.36" gender="1" spkid="MS1" name="Tony Khan"/>
<Speaker dur="38.58" gender="2" spkid="FS2"/>
</SpeakerList>


<SegmentList>
<SpeechSegment stime="0.50" etime="2.09" spkid="MS1" lang="eng-usa">
<Word stime="0.80" dur="0.39" conf="0.971"> This </Word>
<Word stime="1.46" dur="0.13" conf="0.971"> is </Word>
<Word stime="1.59" dur="0.10" conf="0.971"> the </Word>
<Word stime="1.69" dur="0.36" conf="0.971"> world </Word>
<Word stime="2.05" dur="0.00" conf="0.594"> , </Word>
</SpeechSegment>


<SpeechSegment stime="4.12" etime="9.13" spkid="MS1" lang="eng-usa">
<Word stime="4.12" dur="0.12" conf="0.934"> a </Word>
<Word stime="4.27" dur="0.12" conf="0.934"> co </Word>
<Word stime="4.49" dur="0.38" conf="0.934"> production </Word>
<Word stime="4.87" dur="0.08" conf="0.934"> of </Word>
<Word stime="4.97" dur="0.08" conf="0.934"> the </Word>
<Word stime="5.15" dur="0.39" conf="0.934"> BBC </Word>
<Word stime="5.54" dur="0.26" conf="0.926"> world </Word>
<Word stime="5.80" dur="0.51" conf="0.444"> Service </Word>
<Word stime="6.47" dur="0.66" conf="0.568"> PRI </Word>
<Word stime="7.48" dur="0.14" conf="0.917"> and </Word>
</SpeechSegment>
...
</AudioDoc>
```

Why Is Speech Processing Difficult?

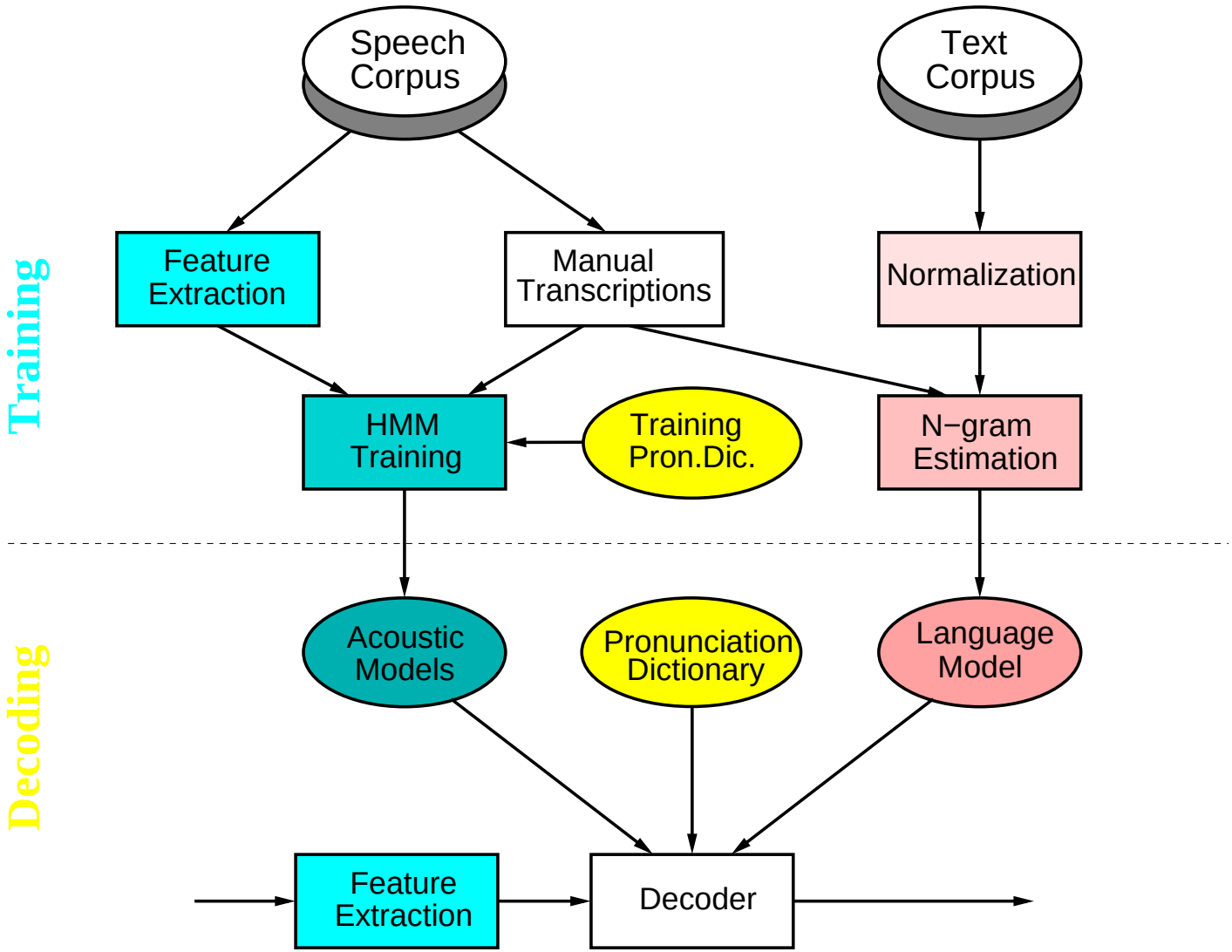
Text:	I do not know why speech recognition is so difficult
Continuous:	Idonotknowwhyspeechrecognitionissodifficult
Spontaneous:	Idunnowhyspeechrecnitionsodifficult
Pronunciation:	YdonatnowYspiCrEkxgnISxnIzsodIfIk∧lt
	YdonowYspiCrEknISNsodIfxk∧I
	YdontnowYspiCrEkxnISNsodIfIk∧lt
	YdxnowYspiCrEknISNsodIfxk∧lt

(after lecture notes from T. Schultz)

Important variability factors:

Speaker	*Acoustic environment*
physical characteristics (gender, age, ...), accent, emotional state, situation (lecture, conversation, meeting, ...)	background noise (cocktail party, ...) room acoustic, signal capture (microphone, channel, ...)

STT System Development



STT Progress

Goal

Development of generic technology: speaker/task independent, robust (noise, microphone...)

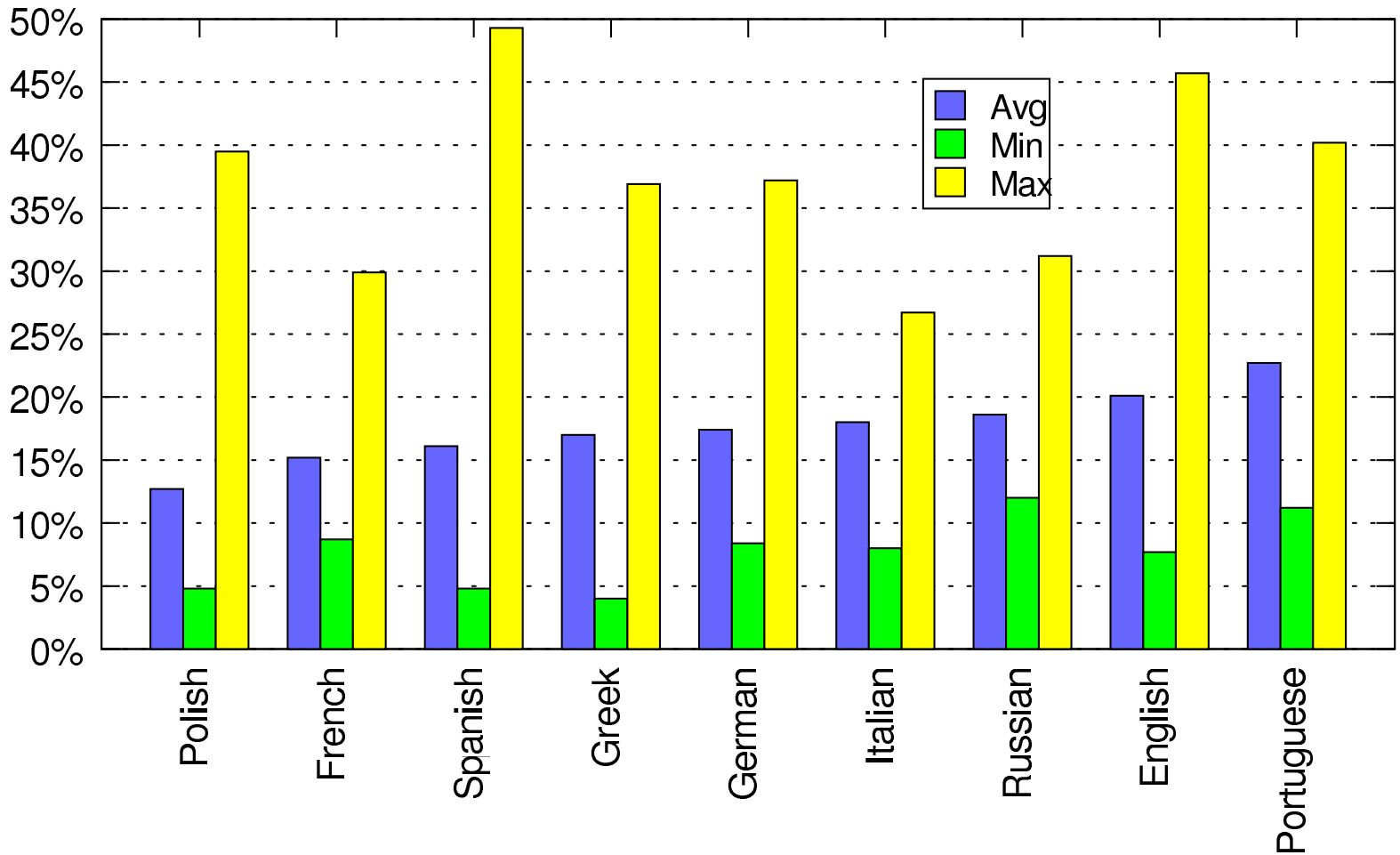
- Lower error rates
- More varied found data with varied speaking styles, uncontrolled conditions: BN, BC, CTS, lectures, ...
- More languages covered
- Enriched STT output (case, punctuated output, accurate confidence scores, multiple hypotheses, topic tags, ...)

QUAERO 2011 STT Evaluation

- STT assessed for 9 languages
  - Primary Quaero languages: English, French, German
  - Third evaluation for Russian & Spanish
  - Second evaluation for Greek & Polish
  - Baseline evaluation for Italian & Portuguese
- 3 hours of development data per language
- Evaluation guidelines
- Metrics: CI and CS word error rate, LNE scoring tools
- Total of 30 hours of test data in 2011
- 70/30, 50/50, 30/70 split broadcast news/conversation
- 91 STT submissions (from 4 sites)

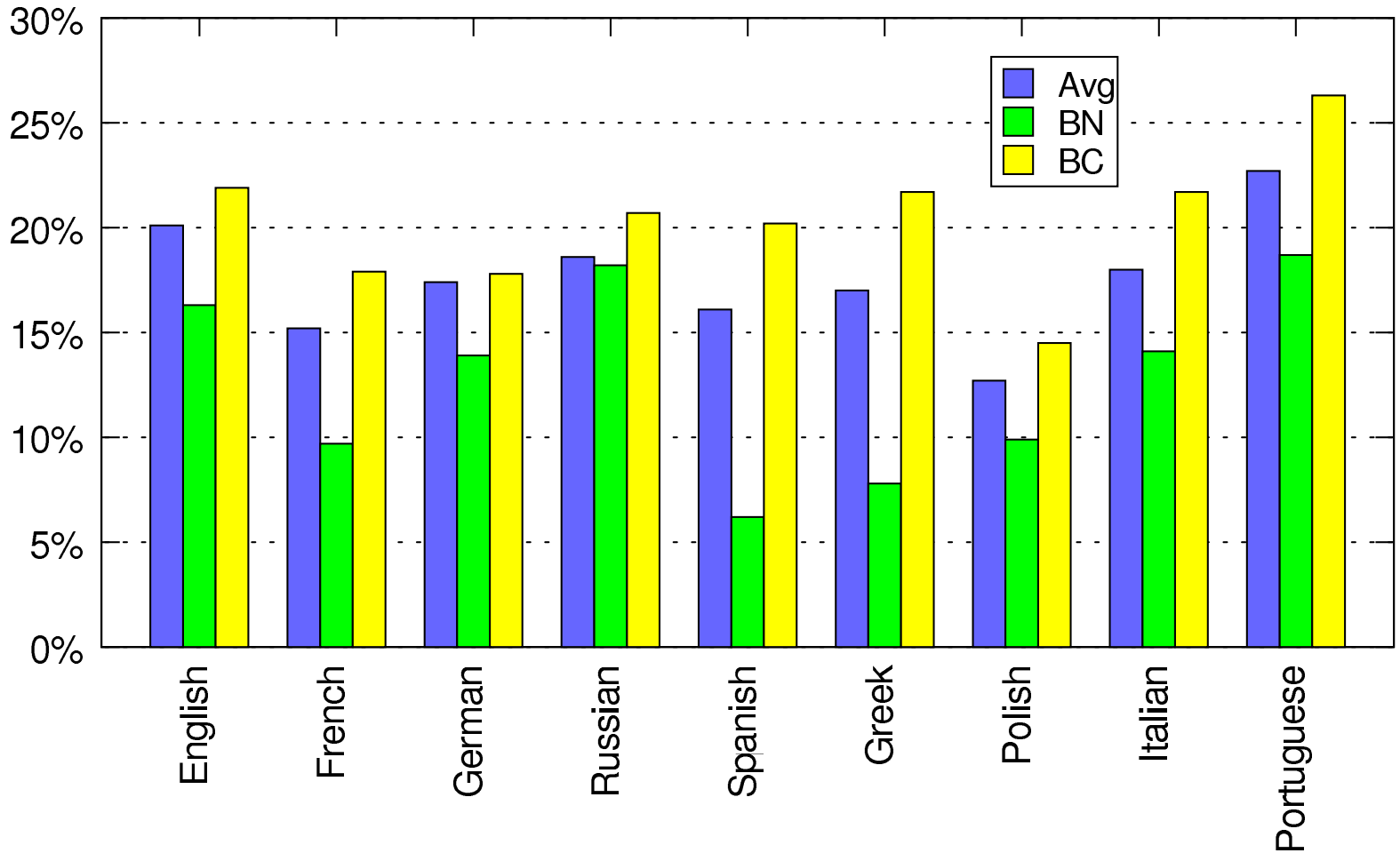
CI WER versus Language

Mix of broadcast news and broadcast conversations
Average, lowest and highest document WER



CI WER versus Type

Mix of broadcast news and broadcast conversations

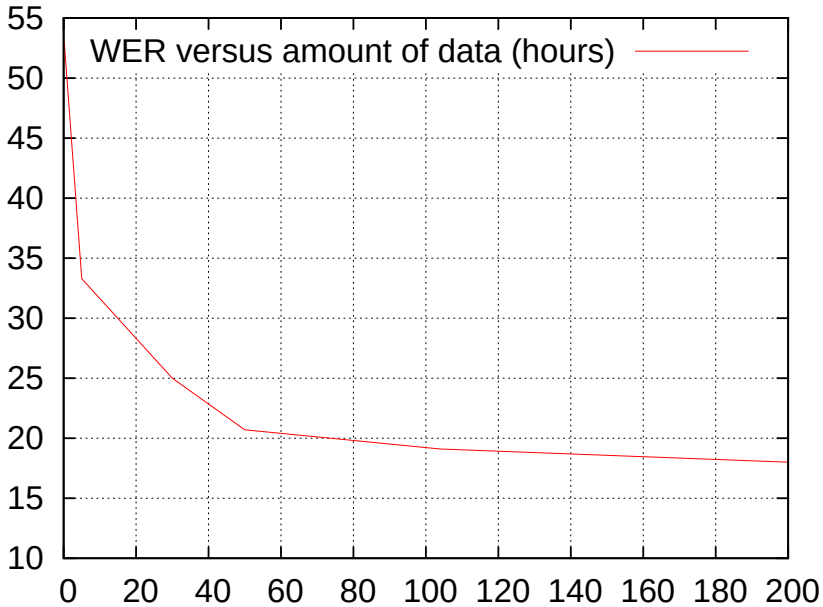


STT System Components

- Acoustic models: HMMs (10-20M parameters), allophones (triphones), discriminative features (MLP), discriminative training
- Pronunciation dictionary with probabilities (30-50 phones), g2p, statistical
- 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)
- Adaptation (unsupervised & supervised)

Data for Model Training

- Data collection and transcription is costly
- How much does data bring?



BN data, ASR2000

- Asymptotic behavior of the error rate
  - rapid progress on new problems (i.e. new data)
  - but slow progress on old problems (on average 6% per year)
- Addl data should cost less (need to learn to better use lowcost data)
- Need more varied data

Speech Transcription Demo

STT Results on 2012 dev

LIMSI/Vocapia	*System*	
Language	2011	2012
German	18.0	16.4
Greek	17.0	17.0
Italian	17.9	13.5
Spanish	16.1	15.2
Russian	19.4	18.9
Polish	12.7	12.0
Portuguese	23.7	17.4

	System
Language	2012
Czech	18.7
Bulgarian	29.0
Hungarian	27.7
Latvian	18.8
Luxembourgish	-
Romanian	15.1
Slovak	20.1

- 9 languages from 2011
- 7 additional languages (dev data only) - no training data provided
- Case-insensitive WER on 2012 development data

Speaker Diarization

- Speaker Diarization (Who spoke when?)
  - Speaker segmentation
  - Speaker segment clustering
  - Speaker identification using speaker model and speech transcription
  - Within and cross show
- Person identification in video: speaker diarization, OCR in video, face recognition, fusion
- Some specific uses:
  - precise timing of political interventions (debats televises).
  - searching a specific declaration of a politician
  - meeting transcription (very complex)

Cross-Show Speaker Diarization

- Speaker diarization on interactive data
- Index/search a set of shows from the same source
- Contrast with single-show processing

General architecture

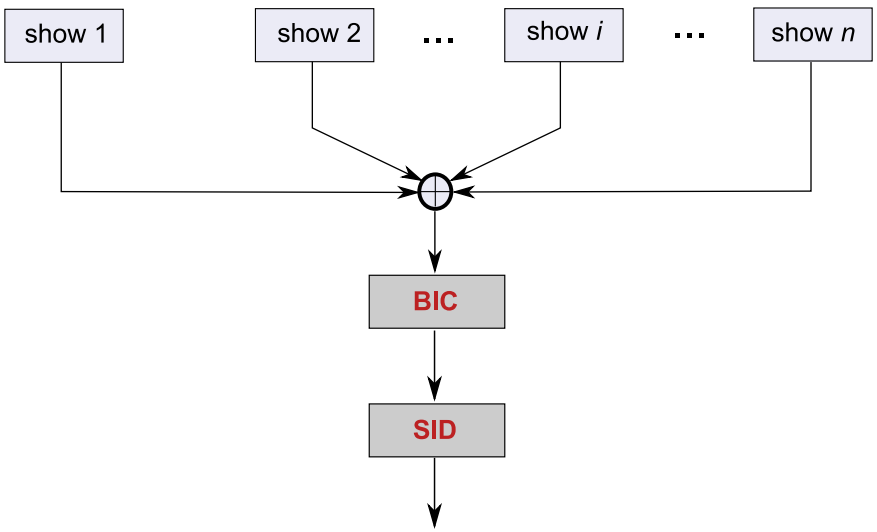
- Audio segmentation with GMM models
- Speaker change detection
- Agglomerative hierarchical clustering with BIC and/or SID
- Schemes: concatentation, show-based clustering followed by global clustering, incremental diarization

Improvements

- Selective clustering (minimal length constraint)
- Alternative acoustic features and fusion [$\Rightarrow$] 10% relative to last year's baseline
- Large variation across shows
- Main sources of errors in more interactive shows: short turns, laugh and applause, overlapping speech

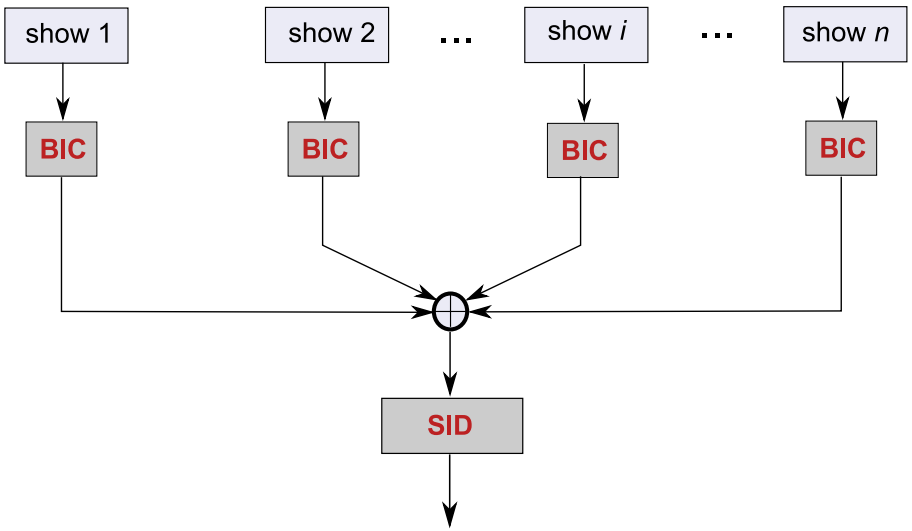
Cross-Show architectures (1)

Global BIC + Global CLR



Obvious, but costly

Local BIC + Global CLR



Reduces computation cost

from V.A. Tran et al, Interspeech 2011

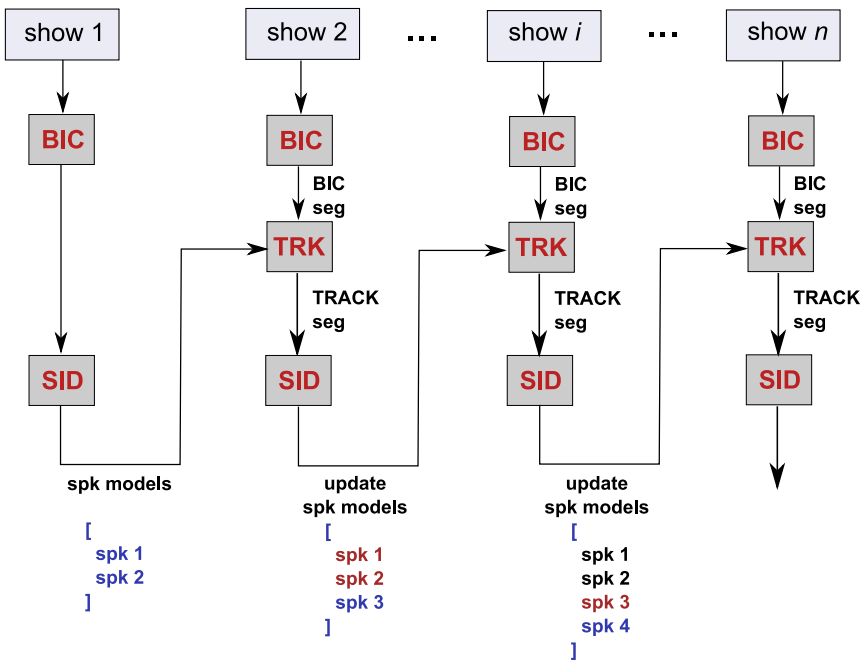
Cross-Show architectures (2)

Limitations of glocal schemes:

- Process all shows simultaneously
- In real application, shows are presented to the system over time

Alternative incremental approach: local BIC, incremental CLR

- Only information from previous shows available (no prior information for first show)



from V.A. Tran et al, Interspeech 2011

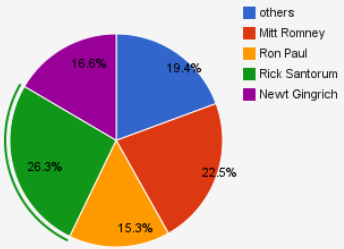
Single- and cross-show speaker diarization

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

- Cross-show diarization error is about twice single show
- Most errors are confusions

Transcription overview | Speaker static overview | Speaker dynamic overview

Speaker chart



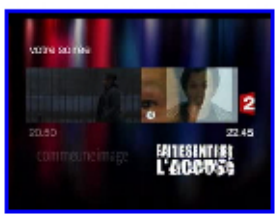
Speaker overview

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

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 dealing problem where I else including president That's entitlement spending Obama

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

Language Recognition

- Phonotactic language recognition systems
- Context-dependent phone recognizers better performance than context indepeenent
- But the computational requirements are high (several times real time)
- Extensive experimental work to assess the effect of several parameters and schemes on both system performance and processing speed (acoustic scale factor, phone insertion penalty, LM prunning, beam search width, ...).
- Language model smoothing techniques
- Discriminative features (MLP)
- Updated Quaero LID system used to identify languages in TrecVID 2012 and MediaEval 2012 data sets

Human benchmarks

- Human listeners significantly outperform machines on speech transcription tasks (5 to 6 times better than machines) [Greenberg, 1996; Lipmann, 1997; Pools, 1999]
- Variation handling: machines have trouble with rare events that are poorly modeled (pronunciation variants, disfluencies, ungrammatical sentences, noise, native and non-native accents etc.)
- Information sources
  - Humans use “higher-level” knowledge
  - Human listeners and ASR systems likely use different acoustic cues
- Speech Communication (2007) special issue on Bridging the Gap: HSR vs ASR

Comparing Human and System Errors

- Analysis of speech regions involving ASR errors
  - To increase knowledge of speech variation
  - To identify potential shortcomings in speech models of ASR systems
  - Focus on frequent short, acoustically poor, function words subjects to contextual homophony in French and English
- Sources of ASR errors
  - Intrinsic spoken language ambiguities (language bias)
  - Simplified speech models (model bias)
  - Role of context (Shinozaki & Furui, 2003, Vasilescu et al, 2009, 2011)

Perceptual Paradigm on Role of Context

Objective: Assess the role of increasing context in disambiguating problematic targets

- et, est, des, les, à, a (French)
- and, in, the, a, is, was (English)

Experimental protocol:

- English Quaero 2009 (24% WER) and 2010 (17% WER) data
- 4 distinct sets of 200 (French) and 200 (English) stimuli
- Each stimulus is presented in n=3, 5, 7 and 9-grams
- One stimulus per context length in each test
- 90% of stimuli with ASR error on central target word
- Different types of ASR errors (deletions, substitutions, insertions)
- 40 native French subjects, 76 native English subjects

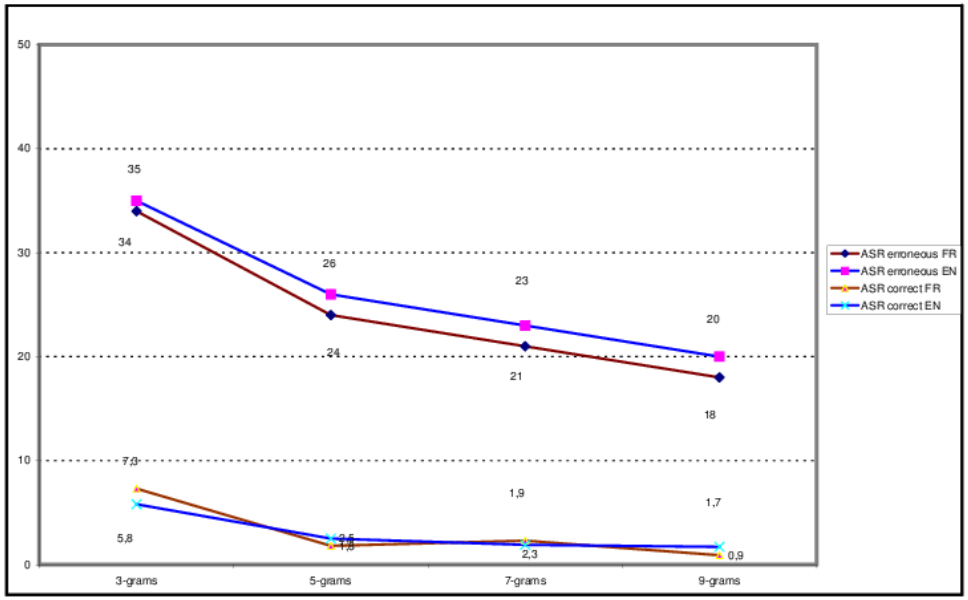
Examples

ASR error		English
sub	hyp	so the review panel WAS headed by David Davis the review panel WAS headed by David review panel WAS headed by panel WAS headed
	ref	so the review panel IS headed by David Davis

ASR error		French
sub	hyp	comme la région Auvergne EST légitime pour communiquer auprès la région Auvergne EST légitime pour communiquer région Auvergne EST lgitime pour Auvergne EST légitime
	ref	la région Auvergne ET légitime pour communiquer

Perceptual results

- Target words pose problems for humans:
  Human WER 21.5% French, 22.5% English
- Higher human error rate on stimuli with ASR errors



- Strong reduction of human WER with increasing context
- Increasing from 3-gram to 5-gram gives largest gain
- Humans more errors for ASR deletions (poor acoustic information), least for ASR insertions

Conclusions and Outstanding Challenges

- Still a large gap between human and machine performances
- Incorporating semantic and world knowledge in models
- E.g. the punctuation task: “woman without her man is nothing”
  - Woman, without her man, is nothing.
  - Woman, without her, man is nothing.
- Automatic learning from data
- Building successful applications with imperfect technology
  - User in the loop
  - Using dialog as humans do
- Speech and language technologies will continue to play a major role

Example Quaero Applications

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

Lecture Translator (KIT)