ELRA HLT Workshop on Evaluation
Malta, 1-2/12/2005

Speech recognition evaluation:

Practices and issues

Edouard Geoffrois

DGA/DET/CEP (Centre d’expertise parisien)

`Edouard.Geoffrois@etca.fr`

Outline

- Domains and tasks (to date)
  - Metrics
  - Data types
  - Combinations with other domains
- Historical perspective
  - Early days
  - Recent and current campaigns
  - Perspectives
- Generic issues
  - Best practices
  - Communicating outside the community
  - Funding schemes

Domains and tasks

- Speech recognition (orthographic transcription)
- Speaker recognition
  - Identification
  - Segmentation
- Language recognition
  - Identification
- Speech understanding for dialog

Metrics

- Transcription
  - Word/Character error rate (WER/CER)
- Speaker and language recognition
  - False alarm / miss rate, ROC or DET curve
  - Minimum cost, Equal error rate (EER)
- Understanding
  - Concept (semantic attributes) error rate (CER)

Data types

- Talking to a computer
    - “Read” speech (command, dictation)
    - Human-machine dialog
- “Found” speech
    - Broadcast news
    - Lectures
    - Talk shows
    - Interviews
    - Telephone
    - Meeting

Combination with other modalities

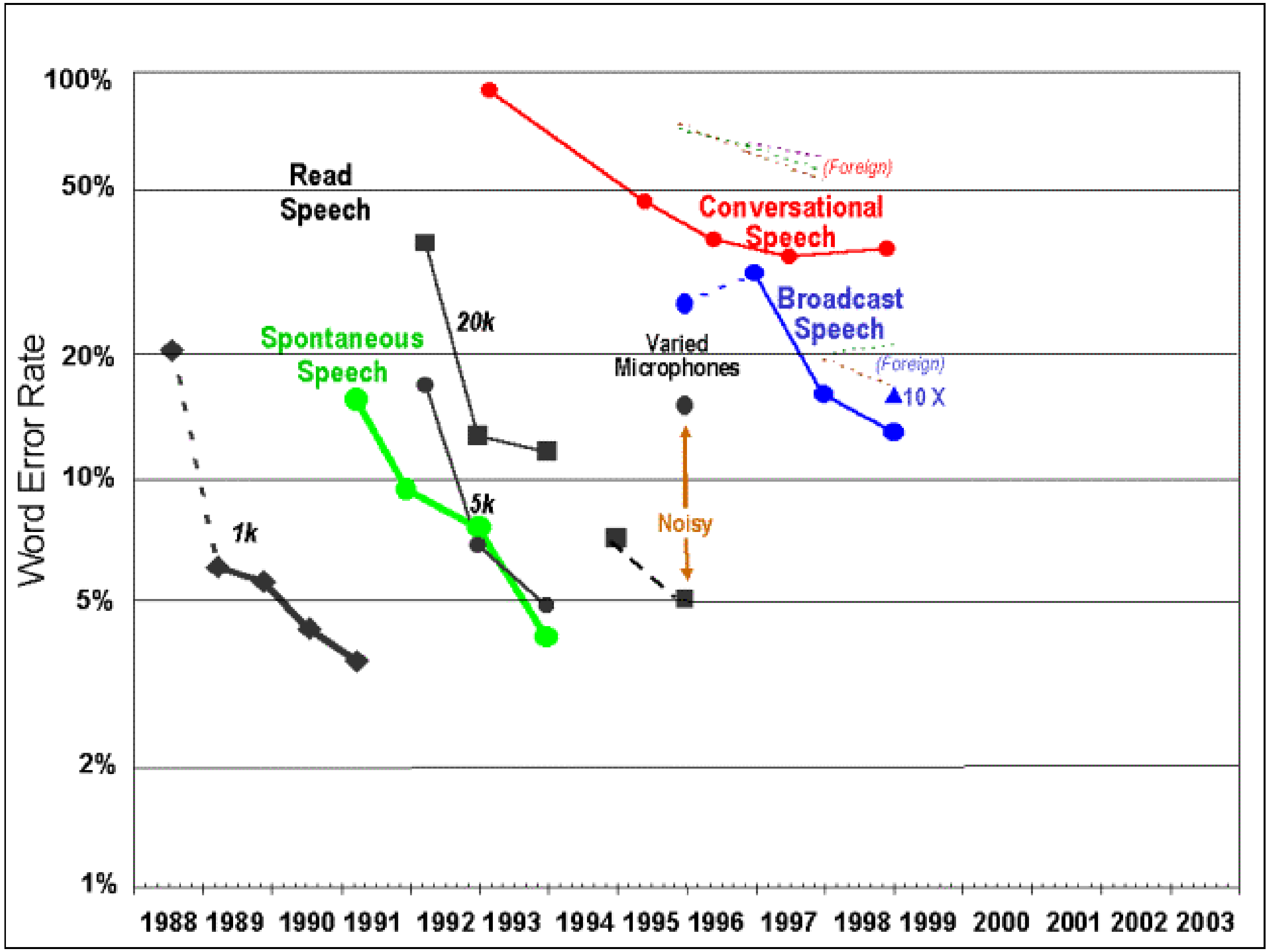
- With NLP
  - Spoken document retrieval (SDR, CL-SR)
  - Named entity detection on speech (Technolangue/ESTER)
  - Speech translation (TC-STAR, GALE)
  - …
- With image
  - Video document retrieval (TRECVID)

Historical perspective: early days

- The 70's : ARPA SUR
  - Performances of systems for the same task were measured, but on different databases
- Early 80's: NATO/RSG10 evaluation database
  - Common database, but no strong incentive to use it
- Mid-80's: First DARPA/NIST evaluation campaign

These steps paved the way toward an organized community using objective and reproducible measurements to share results and make progress

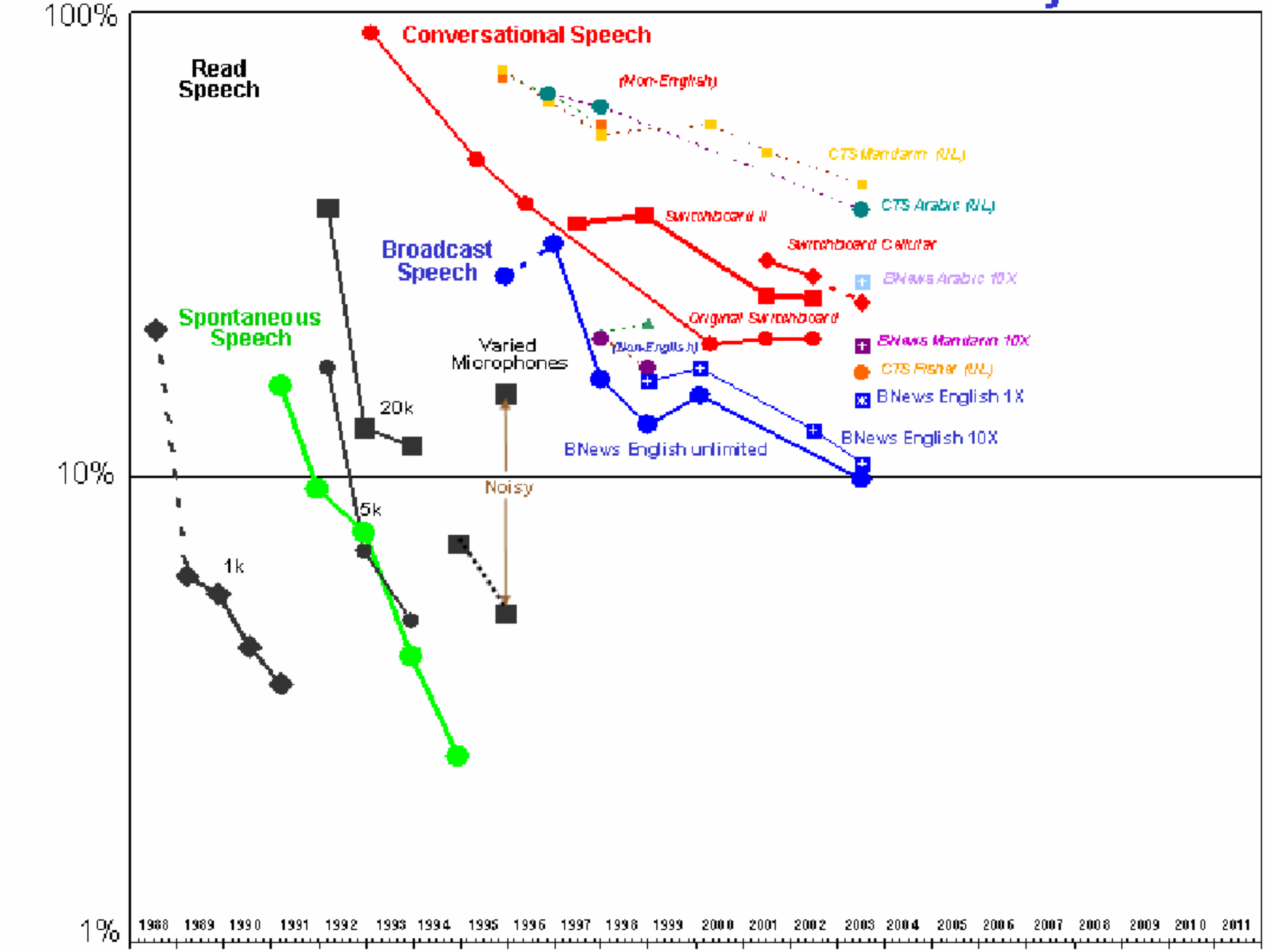
NIST evaluations as of 1999



Courtesy NIST

4 years later…

2003 - NIST Benchmark Test History



Courtesy NIST

Since then…

- Very challenging targets in DARPA EARS / NIST RT Fall 04
  - targets were met…
  - but program was stopped anyway!
  - followed by new GALE program
- News European programs systematically include evaluation
- Evaluation initiative launched in France as part of the Technolangue program (2003-2006)

Main current campaigns

- DARPA/NIST
  - RT: rich transcription
  - SRE: speaker Identification
  - LRE: language identification
  - GALE: transcription and translation of broadcast news, talk shows and meetings
- European projects
  - TC-STAR: transcription of lectures and broadcast news
  - CHIL: transcription of seminars
  - AMI: transcription of meetings
- Technolangue
  - ESTER: rich transcription of broadcast news
  - MEDIA: spoken dialog (out of / in context)

Technolangue/ESTER impacts

- More, better technology
  - 8 automatic transcriptions systems submitted, whereas only 1 existed previously
  - Significant performance improvement between dry run and official evaluation
- More, better data
  - Production of 60h of data in addition to 40 existing ones
  - Data validated and soon distributed
- Better communication among the community
  - All national research centers involved, adopted methodology
  - Corpus starts to be used by linguists

Perspectives

- Other types of material
  - General broadcast, teleconferences, VoIP, …
- Multiple types of material, multilingual data
  - To encourage genericity and coverage
- Recognition of other types of information
  - Emotions, noises, acoustic scene analysis
- Machine reaches the level of a human by 2030?

  … if the pace of error reduction is kept steady…

Generic issues

- What is the reference?
- How to publish results?
- “Technology” vs. “usage” evaluation?
- What exactly is “evaluation” about?
- What are the appropriate funding schemes?

What is the reference?

- Multiple gold standards
  - e.g. orthographic variants
  - more variants can be added in adjudication phase (cf. pooling method of TREC and edit distance of GALE)
  - no such thing as a single gold standard (“silver standard”?)
  - metric is distance from system output to a set rather than to a point
- Validity of reference
  - measurable by degree of consensus among annotators
  - inter-annotator disagreement of a few percent is common
  - defines target for “human-like” machine performance

How to publish results?

- Nominative or anonymous?
  - nominative is the only scientifically acceptable option, but commercial stakes, and risk of misunderstanding out of context
  - anonymous results can generally be reconstructed anyway!
- Creating a catalog or summary of evaluation campaign results?
  - would be a nice tool to give an objective view of the state of the art in a broad domain
  - is it possible without distorting reality?

Technology vs. usage evaluation?

- Evaluation is a bridge between research and industry

Technology	*Usage*
Fully automatic, reproducible	Human in the loop, not perfectly reproducible
Human creates reference, user is modeled	Human executes the metric, real users
Drives progress	Measures acceptability

- Automatic metrics might involve approximations, but a metric monotonically related or at least correlated to the application is better than no metric at all

What exactly is “evaluation” about?

- Is HLT evaluation special, or just another case of benchmarking?
  - evaluating learning-based technology needs new test set for each evaluation to avoid overtraining
  - implies organizing regular evaluation campaigns
  - HLT evaluation is closer to evaluating students (new test for each exam) than to benchmarking products like cars
- Is the word “evaluation” appropriate?
  - means many different things to different persons
  - is it about imposing standard or providing infrastructure?
  - is it about metrology? specification? simulation?

Appropriate funding schemes?

- Can HLT evaluation become profitable?
- Can HLT evaluation deliver “labels”?

Imagine a world where students exams
are expected to be organized
with only partial public funding…

Thank you for your attention!

Any question?