Automatic Identification of Spoken Names and Addresses

– and why we should abolish account numbers!

Melvyn Hunt

Novauris UK

novauris

a james baker company

www.novauris.com

No doubt you know your own phone numbers and car registration number, BUT what about your:

- □ Credit card nos.
- □ Bank account nos.
- □ Mortgage no.
- □ Social Security no.
- □ Car insurance policy no.
- □ House ins. policy no.
- □ Medical ins. policy no.
- □ Passport no.
- □ Electricity account no.
- □ Gas account no.
- □ Water account no.
- □ Telephone account no.
- □ Warranty nos. on countless items

I don’t know any of these!

What I never forget is:

- My own name
- Where I live

- These items are unique identifiers for me – and for everyone else
- They are all we should ever need to identify ourselves

But there are a few of problems with names and addresses:

- Humans are bad at taking down names and addresses quickly and accurately on the phone
  - and in English, increasing use of offshore operators is making the problem worse
- Machine systems work more easily with index numbers
    - and the designers of such systems consequently impose index numbers on us humans

– Until recently machines were even worse than humans at taking down names and addresses

The good news is:

- Machines can now identify names and addresses spoken in a single utterance
  - Provided that these names and addresses are available in a database
- And they can do it:
  - much faster than a human
  - much more reliably than a human
  - and without needing to spell anything

This means that:

- We could abolish user-unfriendly account numbers
  - even if computers still represent individuals internally with numbers
- Call centres currently needing human operators to take down names and addresses can now be automated

The rest of this talk covers…

- ❑ Feasibility of name & address recognition
  - – To justify the claims I just made
- ❑ Advantages of single-utterance input
- ❑ Some immediate applications
- ❑ Technology for spoken database access
- ❑ Relationship to other ASR tasks
- ❑ Recognition of shorter inputs
- ❑ Importance of confidence measures
- ❑ Invitation to try the demos on our stand

Our Feasibility Tests

- To develop and test our capabilities, we needed:
  - A (US) name-and-address database
    - We made a semi-artificial but realistic database
  - A large set of test and training recordings
    - We first made direct-microphone office recordings
    - We later used telephone recordings from the publicly available *Macrophone* corpus
    - Now, we and our customers have begun conducting on-line performance tests

Some statistics on our task

- 244,947,552 addresses in the database
- 245,000 distinct words in the vocabulary
- 3,220 cities, with 2,809 distinct names
- ~1 million distinct street names
- 88,800 distinct last names
- 4,275 female & 1,219 male first names
- ~50 million distinct person names
- 49,384 “*James Smith*”s – 39,638 “*Mary Smith*”s

Generating artificial but realistic person names

- Frequencies of first and last names taken from the US 1990 census
- Equal numbers of male and female names generated by random combination of first and second names, reflecting the published frequencies

Name & Address Speech Corpus

- ❑ Corpus collected in the US
  - – speakers from all major US regions
- ❑ Each of the 181 speakers recorded between 100 and 200 names and addresses presented as if on an envelope
- ❑ 59 speakers held back for testing
  - – Only used once

Recognition Test Results
against 245 million items

- **99.8%** of the 7800 names and addresses were completely correct (with no rejection)
- Mean response time was **0.66 sec**
- Tests carried out on a standard PC:
  - 2.4GHz *P4* with 256MB of 266MHz RAM (only **40 MB** of RAM actually used)
- **So large-scale name and address recognition is more than feasible**

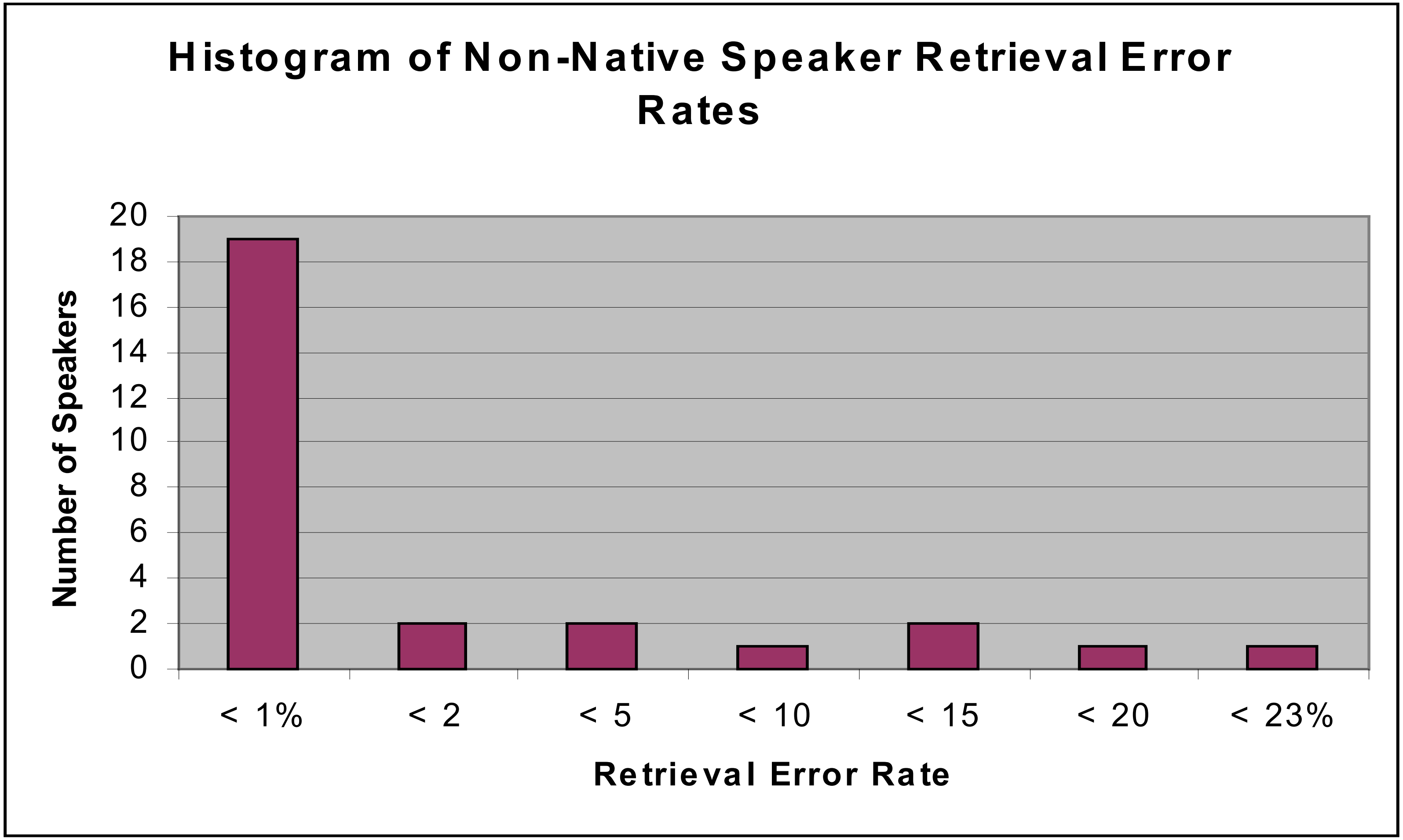
Non-native speaker tests

- Performance with most non-native speakers was as good as with native American English speakers
  - Our particular technical approach makes us more tolerant to variation in pronunciation

First Languages of the Non-Native U.S. English speakers

English (not USA)	6
Mandarin	3
Spanish (N. Am)	3
Arabic	2
German	2
Hindi	2
Korean	2
Bulgarian	1
Burmese	1
Czech	1
Hebrew	1
Japanese	1
Polish	1
Spanish (Spain)	1
Yoruba	1

Results with the 28 Non-Native Speakers



So most non-natives got good recognition accuracy;
67% had <1% errors; 82% had <5% errors;
those with high percentage error rates are barely comprehensible to human listeners.

Address recognition:

without names or house numbers

~ 2.5 million addresses

needed for *e.g.* route planning applications

	Errors	**Rejections**
Top-choice error rate	2.34%	0%
Top-choice error rate with rejection	0.70%	6.9%
Top-3 error rate	1.09%	0%
Response time: 1.33 sec.		

Tests with Telephone Speech

- Indicate that accuracy remains high
  - With office recordings restricted to telephone bandwidth, accuracy = 99.66%
  - With telephone recordings from the public *Macrophone* corpus, with name and address concatenated from separate utterances plus a telephone number in place of ZIP code, accuracy = 99.34%
  - In live tests errors are rare

Single-utterance *vs.* Multi-utterance (dialogue-based) approaches

- **Usual interaction with dialogue:**
  - Which state do you live in?
    - Connecticut
  - Which city?
    - Greenwich
  - Speak the street & house no.
    - 143 Main Street
  - What is your last name?
    - Bawson
  - Please spell that name
    - B – A – W – S – O – N
  - Did you say “Dawson”?
  - *And so on…*

- **Interaction without dialogue:**
  - What is your name and address?
    - James Bawson
143 Main Street
Greenwich
Connecticut
06830
  - Thank you

❑ ***Much quicker!***

Single-Utterance Name & Address Input is User-Friendly

- □ Many automatic speech recognition systems are more convenient for the service provider
- □ Taking down a name and address automatically is also more convenient for the user because:
  - – It's faster
  - – More accurate
  - – No need to spell names

Single uttereance

Former Head of BT’s Speech Processing Research Department, Denis Johnston, said:

“This level of performance may permanently change how application designers approach dialogue design.

“Quite simply, it makes speech recognition systems far more attractive to users.”

Example of an Immediate Application

- A traveller has had her credit card stolen
- She needs to get it stopped immediately, but has no record of its number
- She identifies herself by her name and address
- Currently this has to be taken down (slowly) by a human operator
- Finance companies say there would be large savings if only a fraction of such calls could be handled automatically

Immediate Applications in General

- Obvious applications:
  - In call centres
  - In road-vehicle navigation systems
  - In parcel sorting
  - In financial info. and transaction processing
- But also some slightly less obvious ones not involving addresses…
  - In consumer entertainment for selecting music or video selections, artists, satellite TV channels and programmes, *etc*.

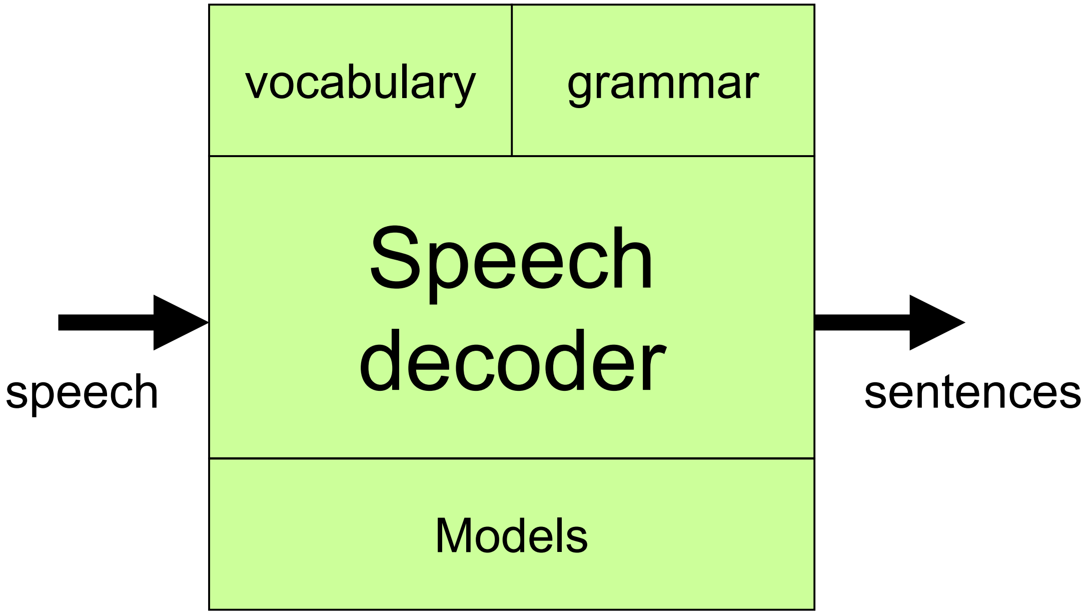
How have we achieved these capabilities?

- By exploiting redundancy in the grammar, but also:
    - With special, novel database search techniques
    - With special, novel speech recognition techniques, employing more speech knowledge than conventional systems

- Patent applications submitted on both sets of techniques

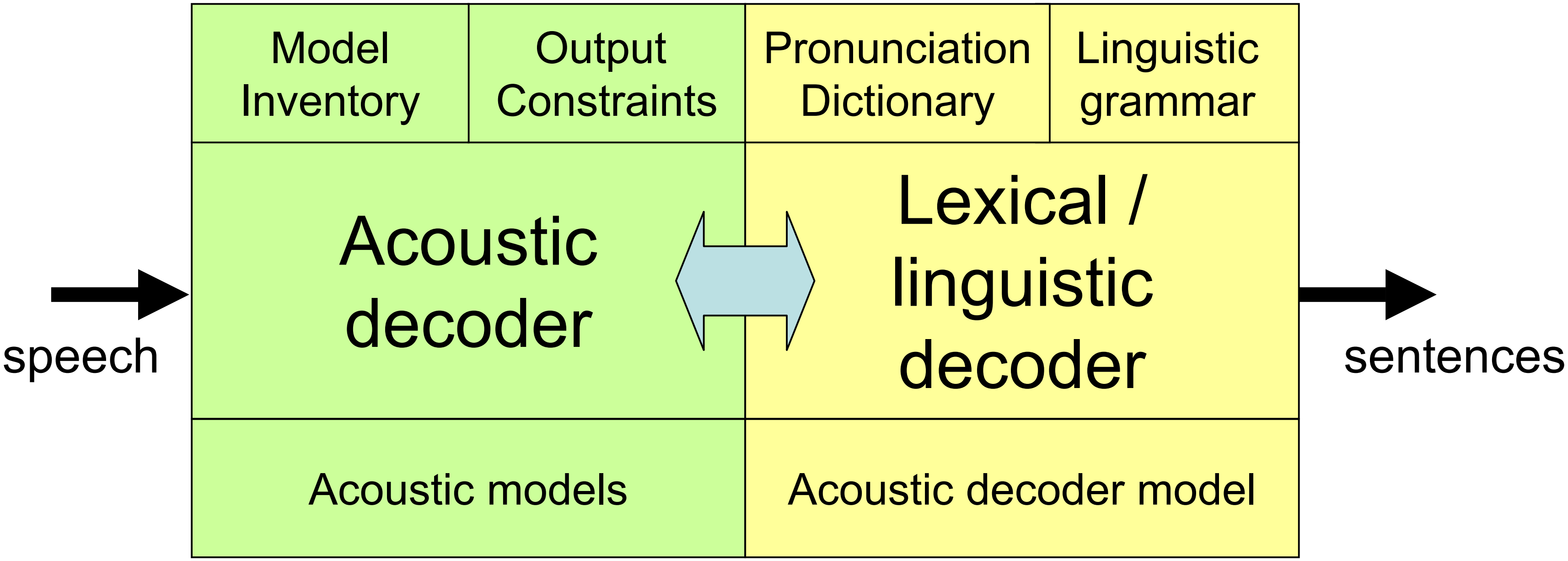
Our basic architecture is also unconventional…

Conventional Speech Recognition



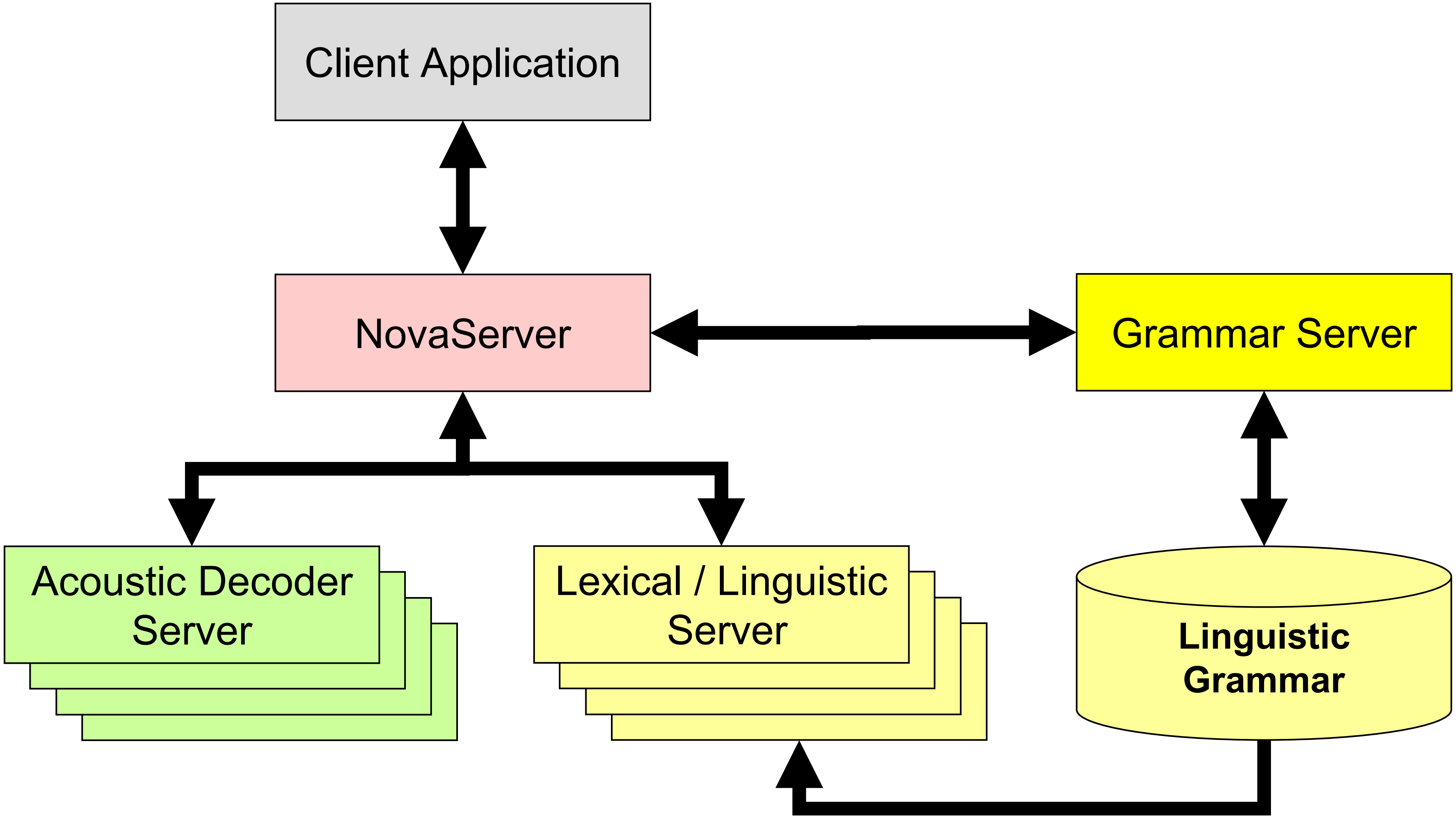
- The acoustic, lexical and linguistic modelling is done entirely within the decoder

Novauris' System



- Different modules are used for the acoustic modelling and the lexical / linguistic modelling
  - A more memory efficient lexical vocabulary is possible
  - A long-range linguistic grammar is possible

The Scaleable Architecture



Client Application

NovaServer

Grammar Server

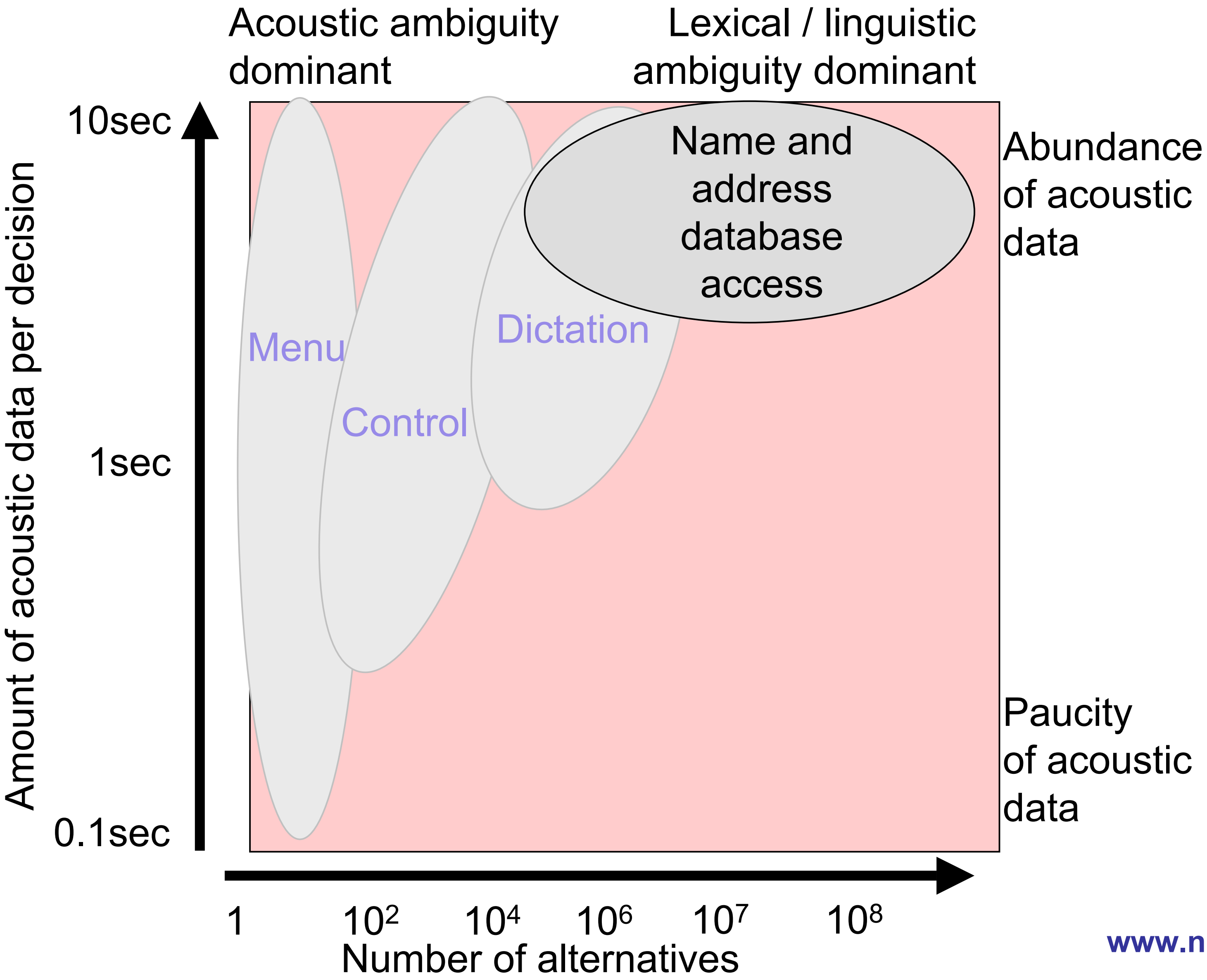
Acoustic Decoder Server

Lexical / Linguistic Server

Linguistic Grammar

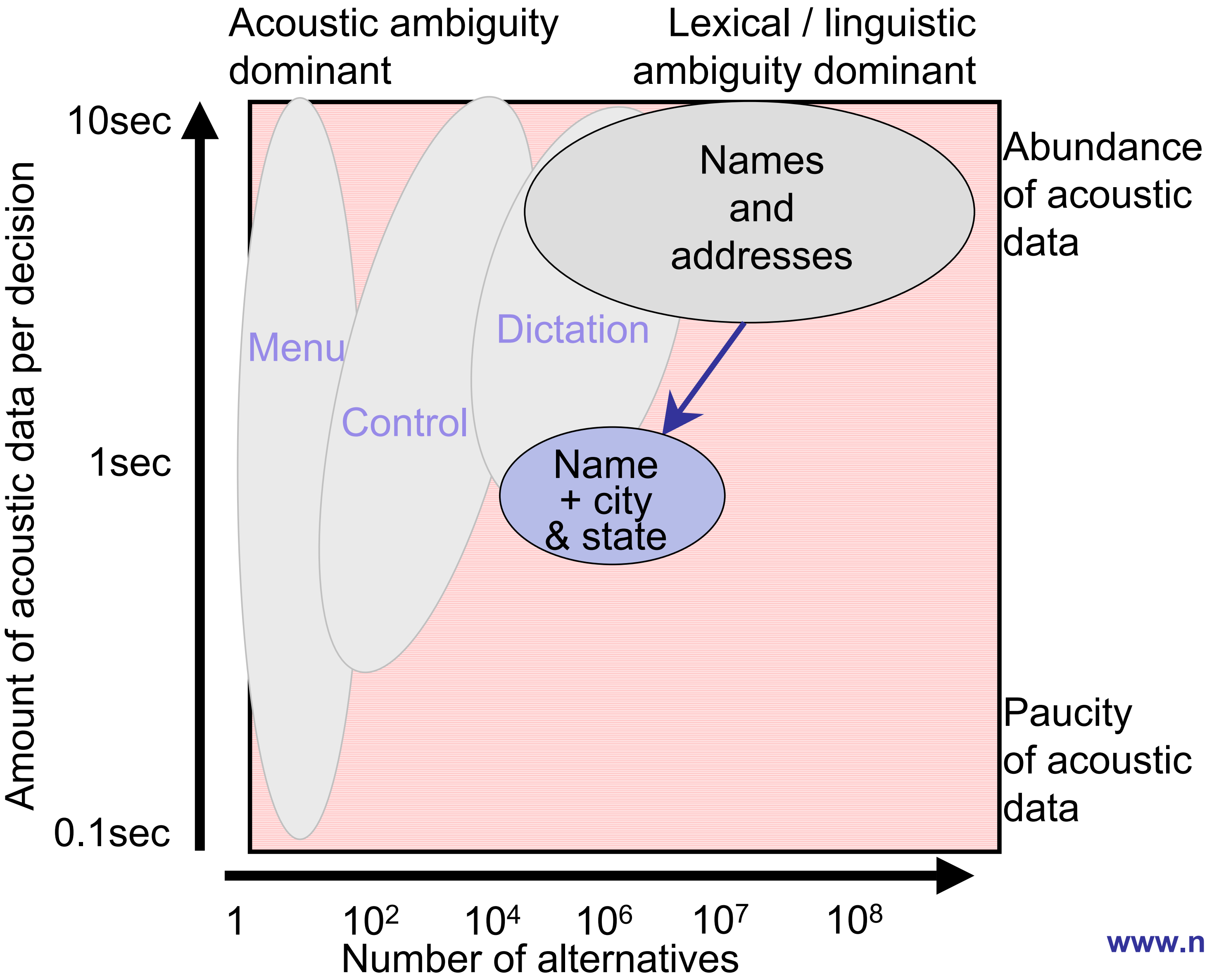
www.novauris.com

Speech Task Problem Space


Acoustic ambiguity dominant
Lexical / linguistic ambiguity dominant
10sec
1sec
0.1sec
Amount of acoustic data per decision
Menu
Control
Dictation
Name and address database access
Abundance of acoustic data
Paucity of acoustic data
1
10²
10⁴
10⁶
10⁷
10⁸
Number of alternatives

www.novauris.com

Moving to shorter utterances



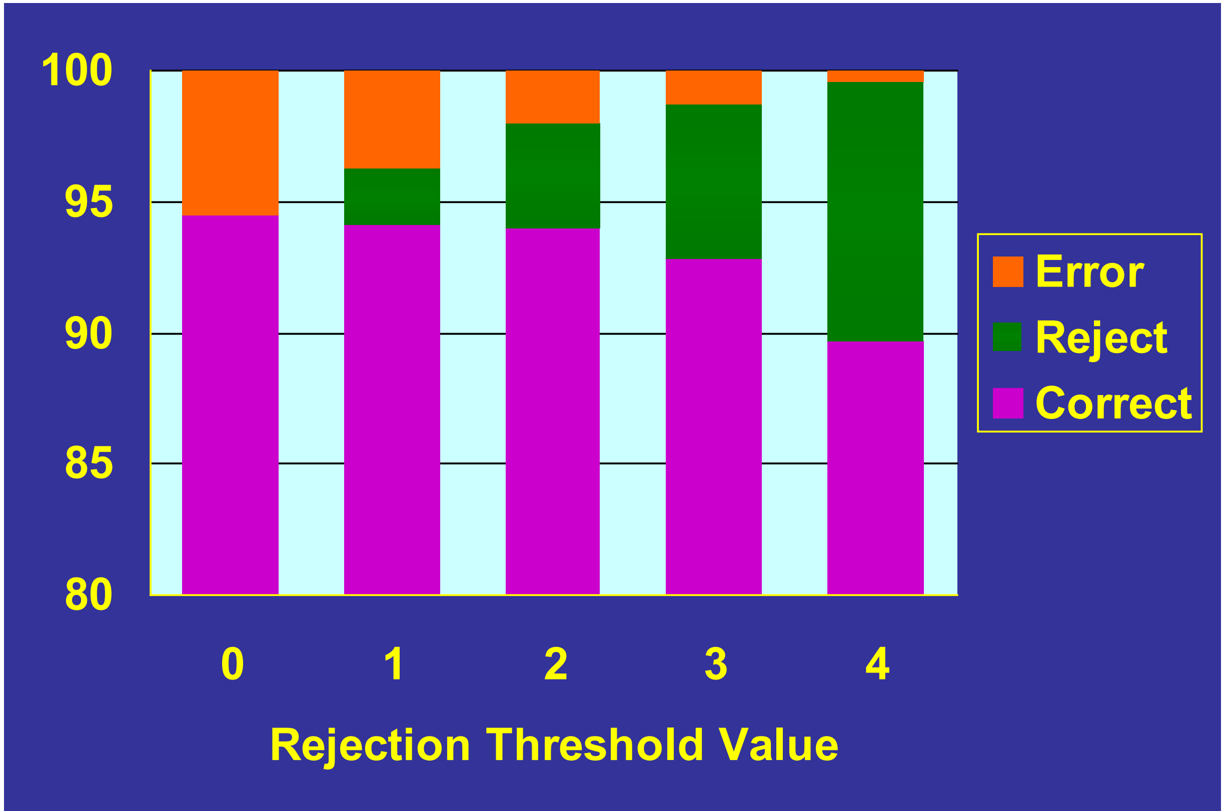
www.novauris.com

Telephone Tests on name + city & state

100,000 items

Confidence measures for rejection are needed to:

- Detect inputs that are not in the database
- Detect possible errors in database matches
  - – Up to threshold value 2 or 3, almost all rejected items would otherwise have been errors



In Summary

- The technology needed to abolish user-unfriendly index numbers is already available
  - Though it won’t happen right away
- In the meantime, there are many uses for automatic identification of (names and) addresses
- The technology providing this capability has potential in other applications

Please try the demonstrations on the Novauris stand (16)

- 245 million names & addresses
  - US English, direct input
- 24,000 names + city & state
  - US English, direct input
- Also, possibility of telephone demo
  - US English demos as 1 & 2
  - Plus name & address demo for British English

novauris
a james baker company

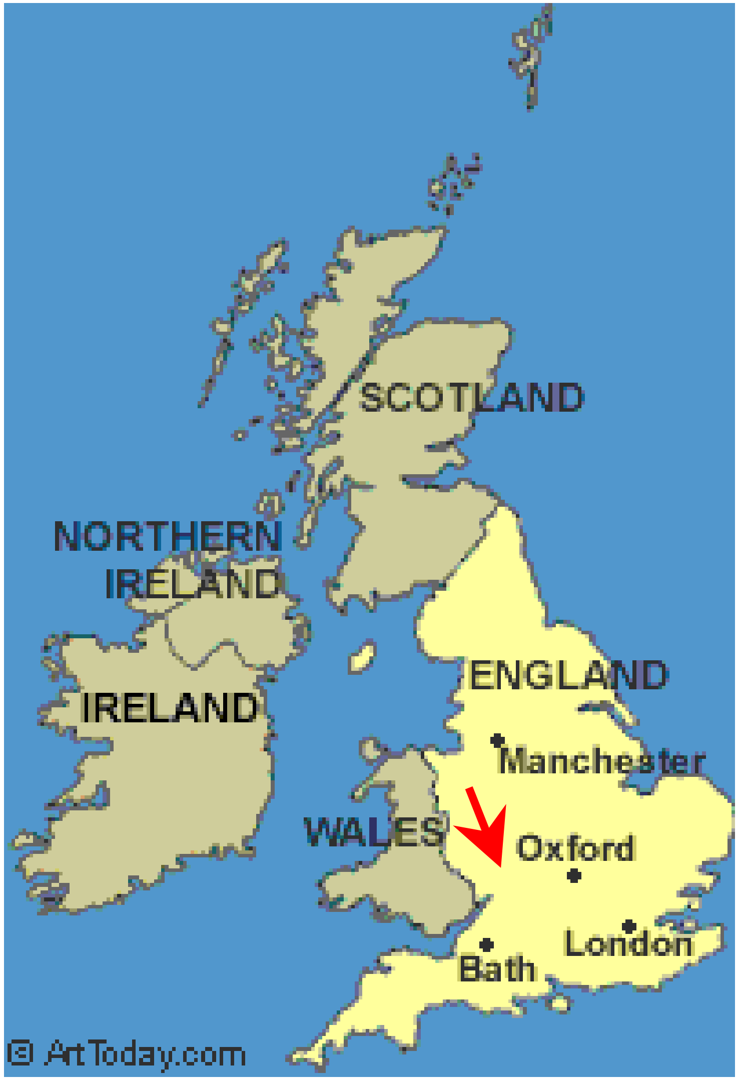
www.novauris.com

Appendix:
Brief Background on Novauris UK

www.novauris.com

+44 1242 678581

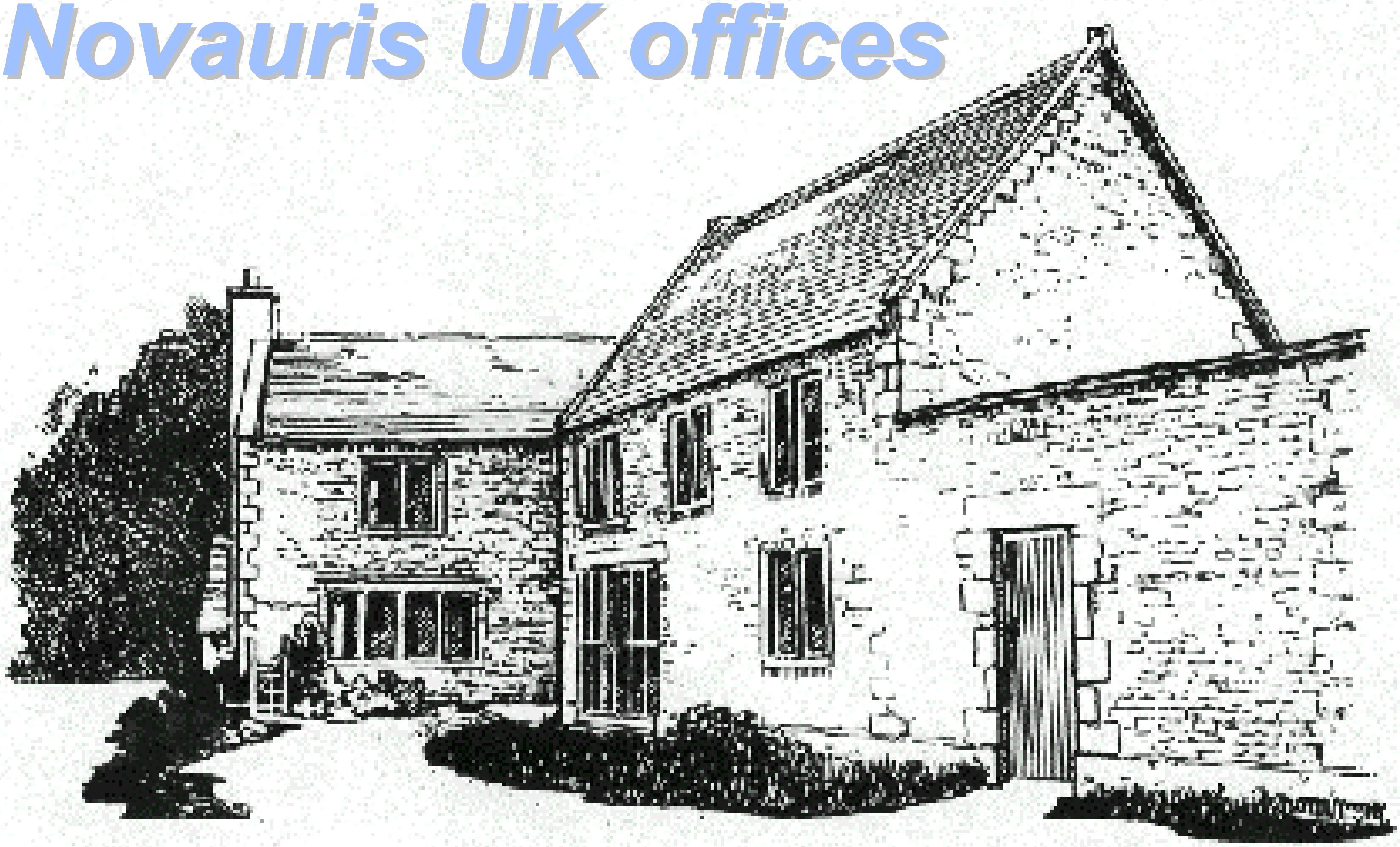
Where Novauris is located



www.novauris.com

Background

Novauris UK offices



How Novauris Began

- Novauris (= “new ear”) founded in March 2002
  - to create a new generation of ASR capabilities and applications
  - initially specialised on spoken access to large databases
  - first public demonstration, April 2003
- Initial funding provided by Jim Baker, who:
  - With Janet Baker founded *Dragon Systems, which:*
    - Pioneered & led the market in general-purpose dictation products
    - Grew to ~380 people, revenue ~$70M
    - Profitable throughout Jim’s 15-yr reign
      - 2 years later, brought down by L&H crash
- Novauris is currently an independent, privately held UK company

The UK Team

- Small, cohesive team of experienced speech technologists
  + 1 administrator/book-keeper
  - Headed by John Bridle & Melvyn Hunt
  - Most have PhDs
- Largely comprise the former Dragon Systems UK R&D team

Dragon Systems UK R&D

- Worked on speech recognition over the telephone and in cars
- Independently profitable unit
  - Headed by John Bridle & Melvyn Hunt
- Developed *C-REC* speech recognizer
  - Suitable for noise-robust embedded and multi-channel telephone applications
  - Applications in: US & UK English, German, French, Japanese…
  - Sold to *Visteon*
    - including command & control and navigation
    - Now fitted in *Jaguar* cars — and others
  - Subsequently licensed by *SpeechWorks (now Scansoft)*

Dr James Baker

- Chairman of Novauris
- Initial investor
- Mathematician by training (Princeton)
- Introduced *HMMs* to speech recognition
- Co-Founder & Chairman of Dragon Systems
- Now lives in Florida

Dr Melvyn Hunt

- ❑ Joint MD — Functions as CEO
- ❑ Physicist by training (Oxford)
- ❑ Honorary Fellow, Dept of Phonetics and Linguistics, University College London
- ❑ Introduced LDA for acoustic representations and MLLR for speaker adaptation.
- ❑ His team in Canada developed the world's first helicopter piloted by voice
- ❑ Introduced what may be the world's first commercial telephone ASR system with barge-in (*Flightline* — 1991)
  - While Chief Scientist, Marconi Speech & Information Systems
- ❑ Served on the IEEE Speech Technical Committee

John Bridle, FIA

- Joint MD — Functions as CTO
- Pioneer in using dynamic programming time warping in the West (early 70s).
- Provided the algorithmic design for the world’s first commercial truly continuous speech recognizer – *Logos* (early 80s).
- Pioneer in using neural networks for speech recognition (mid 80s).
- Formerly head of the UK government’s Joint Speech Research Unit
- Jointly headed Dragon Systems UK
- Fellow of the Institute of Acoustics
- Served on IEEE Speech Technical committee