Voice in situ: creating global speech data

Bradley Music

March 4, 2014

Appen background

- Two decades of experience in Language Tech
  - Speech recognition
  - Text-to-Speech
  - Machine Translation
  - Semantic analysis
  - Search relevance
  - Document classification

    … and more
- Worldwide ‘Managed Crowd’ of over 100,000 language specialists.

Language coverage = 150+

Amharic
Arabic (Egyptian, Gulf, Iraqi, Levantine, MSA, Syrian, Maghrebi – Algerian, Libyan, Moroccan, Tunisian)
Bahasa Indonesia
Bahasa Malaysia
Basque
Bengali
Bulgarian
Cantonese (China PRC, Hong Kong)
Catalan
Croatian
Czech
Danish
Dari
Dutch
English (Australian, Canadian, Gulf, Indian, Irish, New Zealand, Singapore, South African, UK, US)
Estonian
Farsi
Finnish
French (Belgian, Canadian, French)
Gallego (Galician)
German (Austrian, German, Swiss)
Greek
Gujarati
Haitian Creole
Hausa
Hebrew
Hindi
Hungarian
Icelandic
Italian
Japanese
Kannada
Kazakh
Kermanji (Iran)
Korean
Kurdish Sorani
Laki (Iran)
Latvian
Lithuanian
Luri (Iran)
Malagasy
Malayalam
Mandarin (China, Taiwan)
Maori
Marathi
Mazanderani (Iran)
Oriya
Norwegian (Bokmål, Nynorsk)
Pashto
Portuguese (Brazilian, European)
Romanian
Russian
Serbian
Slovak
Slovenian
Somali
Spanish (Columbia, Costa Rican, European, Mexican, Peruvian, US, Venezuelan)
Swedish
Sylheti
Tagalog
Tamil
Telugu
Thai
Turkish
Ukrainian
Urdu
Uzbek
Vietnamese
Wu
Xiang

A look upstream

Speech tech's dramatically increased consumer appeal → greater demands on **Data Creation** (DC)

(i.e. collection and transcription of speech data)

Agenda:

- Factors in Data Creation (w/ examples)

Goal:

- Interesting and useful for anyone dependent on speech data

(Consumer) speech tech today

- More scenarios of increasing breadth and depth, particularly for:
  - Home
  - Auto
  - Mobile

- Manufacturers race to enable these scenarios.

The need for speed(y data)

Manufacturers enable language tech integration through licensing, building, acquiring, partnering – but …

- ➔ **Problem**: If you're talking to a *game controller* vs. a *TV* vs. a *personal assistant*, words, phrases, intonations, demographics, and environments are different and **OOB tech may not have satisfactory accuracy**.

- ➔ **Solution**: Large volumes of transcribed data for training & tuning SR for each scenario.

- ➔ **Updated problem**: Manufacturers, electronics companies serving them, and even technology providers themselves struggle to keep up with the data demands.

Common client challenges in DC

- Data needed via **specialized hardware/software**
- **Specialized audio data** is difficult to come by, e.g. gunshots, aggressive speech, baby crying
- Insufficient data **quantity/quality**
- **Specialized transcription conventions**, e.g. semantic markup, overlapping speech, what's meant vs. what's said
- **TTS talent** hard to find
- Tight **timeframes**
- Poor vendor **communication**
- Data needed in **many languages**, including **difficult, low-density languages**

Data Creation – key factors

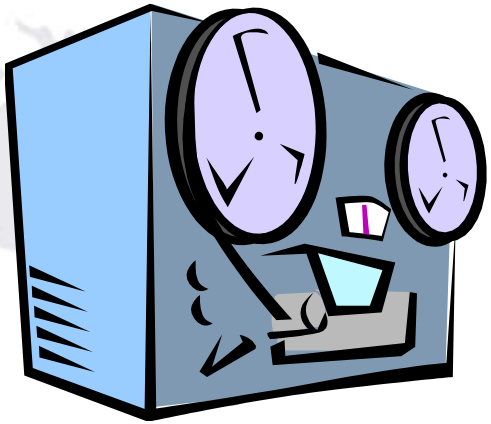
What will they say?

Where will they say it?

What did they say?

Who will say it?

What equipment will they use?



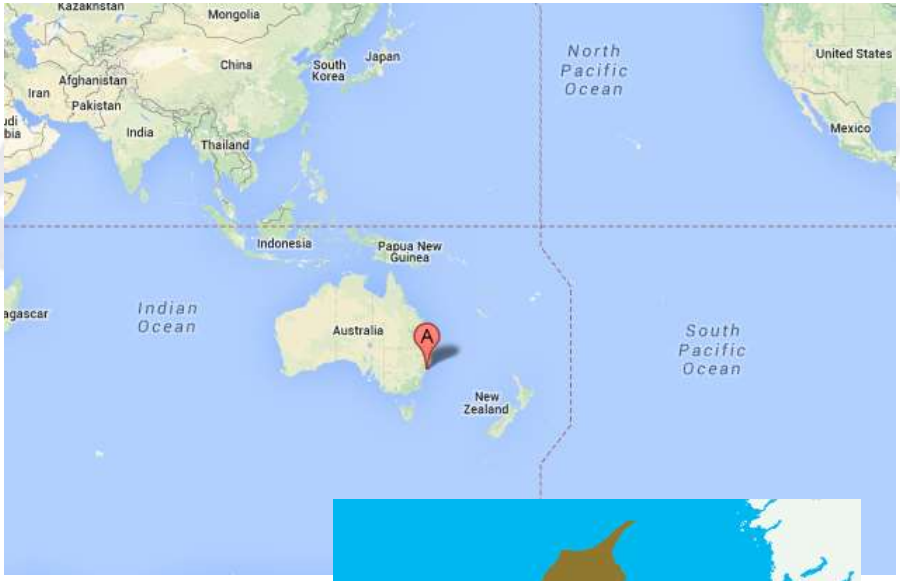
What will they say?

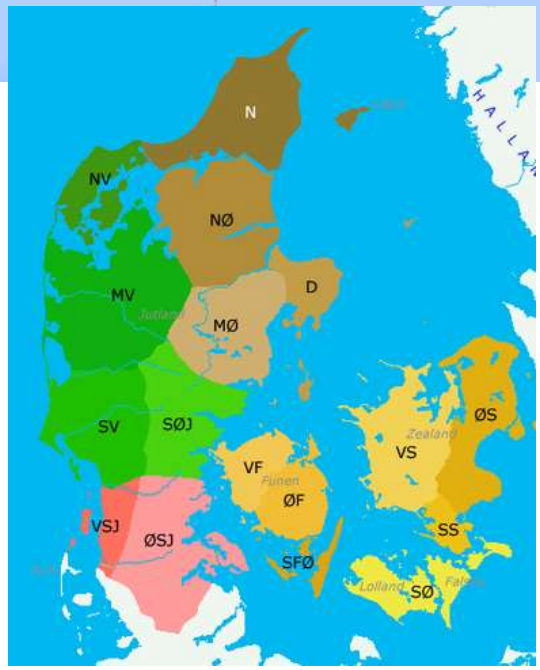
- Scripting for non-existent interfaces
- Balancing response frequency and variety
- Non-leading instructions
- Primary vs. secondary interface
- (Demographic considerations)

Who will say it?

Demographics

- Number of speakers
- Age
- Gender
- Dialect representation

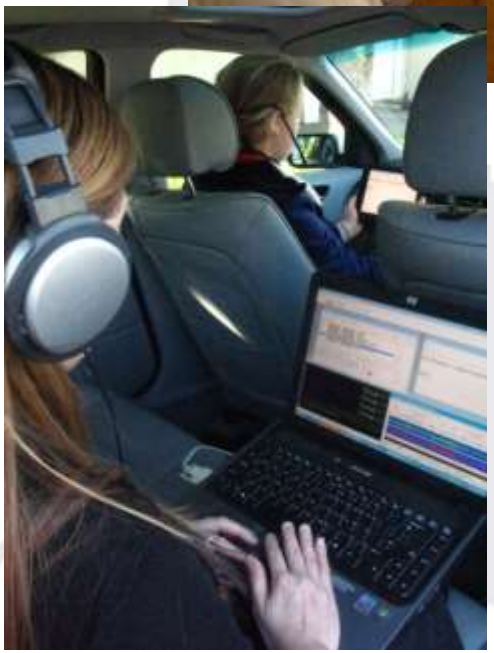




Where will they say it?

Environments (w/ options)

- Home
- Mobile
- Car
  - Models
  - ‘Noise options’
  - Convolution problem
  - The non-existent interface

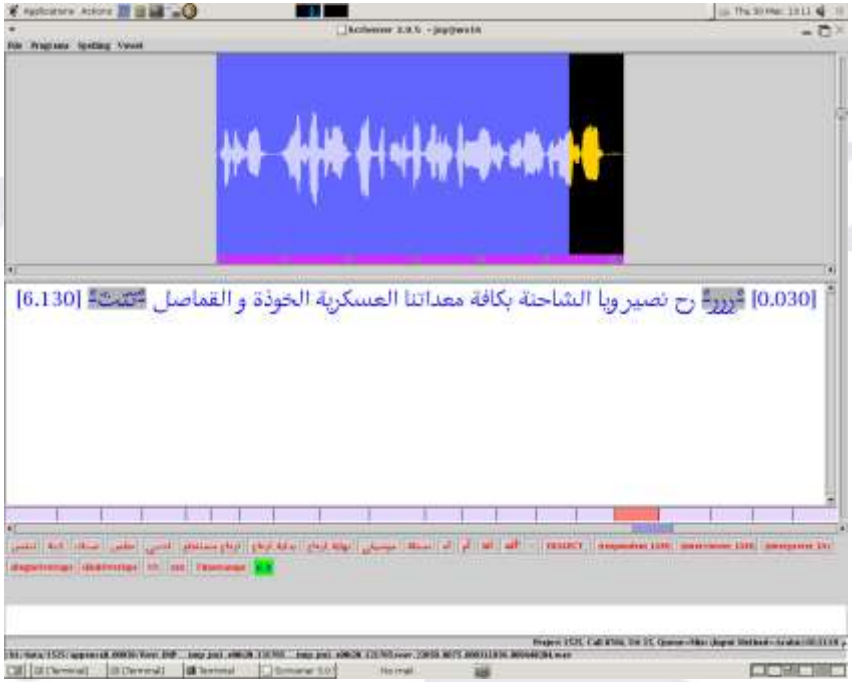


What equipment will they use?

Equipment

- Microphones
- Telephones
- Smartphones
- Computers
- Specialized equipment

What did they say?



Transcription

- In country of origin?
- Conventions
- Quality Assurance
- Automation: QA Segmentation, input filtering

Summary

Lots of factors to consider when creating data!