Voice Recognition FTW!

Dr. Maria Aretoulaki
DialogCONNECTION Ltd, UK
maria@dialogconnection.com
@dialogconnectio @ar3toul4ki

TEDx Manchester
x = independently organized TED event

Monday 13th February 2012, Cornerhouse

Voice Recognition Apps

- Digital Dictation
  - Medical / Legal transcription
  - Speech-to-SMS, Speech-to-Tweet
- Call Centre automation
  - call routing to the right Dept.
  - Voice Self-Service (telephone banking)
- Speech-activated device control
  - Smartphones (Vlingo, SIRI)
  - Game consoles (MS XBox 360 Kinect)

Voice Recognition App Types

- ***Speaker-dependent*** can only recognise a single person / speaker
  - dictation systems
    - PC-based, hand-held
      - e.g. Medical dictation, legal dictation
    - Mobile phone control (oldskool / built-in)
      - Voice dialling, voice search through your mp3s
  - You can talk FREELY to some extent!

Voice Recognition App Types

- ***Speaker-dependent***: can be trained in as little as 5 minutes (max. 20mins)
  - speaking longer / shorter phrases
- work quite well with your voice
  - even if you've got a cold or if you speak softer than usual
- Don't work with your mates though!
  - have to be retrained in order to work with a different (single) speaker

Voice Recognition App Types

- ***Speaker-independent*** apps can recognise anyone speaking the language
  - Automated helplines
  - Voice self-service
    - Cloud-based speech recognition
    - Works from any phone and for any speaker
  - **BUT Restricted domain**
    - You can only talk about specific things (flight booking, benefit claims, online shopping)

Voice Recognition App Types

- ***Enter smartphone apps!***
  - ***Google, SIRI***
- **Speaker-independent** voice-to-text
  - Cloud-based speech recognition
  - Works from any phone and for any speaker
- Unlimited domain, general vocabulary
  - You can talk about anything as long as it's in the general dictionary for that language
    - “SIRI, will you marry me?”
    - “SIRI, where can I get a good Chinese near here?”

Voice Recognition Apps

- Speech-to-text / voice-to-text services, esp. in the context of mobile apps
  - Google voice search on smartphones
  - voice dialling your contacts
  - voicemail to sms / email
  - voice-to-sms, voice-to-email
  - Speech your Tweet / Facebook status

Voice Recognition App Types

- ***Speaker-independent***: work for everyone off the shelf (already trained on billions of data over several years)
- No need for (re)training!
- Sensitive if you've got a strong regional or foreign accent, if you've got a cold or speak softer than usual (Scots in a voice-activated Lift: ELEVEN !)

What is Voice Recognition?

- Speech-to-text conversion
- Spoken words → written words
- continuous wave signal → text string of separate words

What is Voice Recognition?

Speech signal for ".. ***and sadly crime experts predict that one day even a friendly conversation between mother and daughter will be conducted at gunpoint***" (from the Channel 4 comedy series "Brass Eye" - Season 1)

What is Voice Recognition?

- Stream of sound (spoken words)
  - */iwanttospeaktosomeoneataccounts/*
- → written representation of those words
  - *I want to speak to someone at Accounts*
- → user intention
  - Forward call to Accounts Dept.
- → next app action or prompt
  - connects you to the right person

The Voice Recognition Process

- Voice Recognition is NOT an exact science
- Even among humans, voice recognition is fraught with misunderstandings or incomplete understanding

The Voice Recognition Process

- How many times have you had to repeat your name to someone?
- How many times have you had someone cracking up with laughter, because they thought you said something different to what you actually said?

- no wonder that computers do it even less well
- It's all guesswork really!

The Voice Recognition Process

- Voice recognition is usually based on a “lexicon”
  - A text representation of all relevant words and their phonetic “transcription” (pronunciation)
    - Apply → @plai
    - Applicant → Aplik@nt
  - All the ways that people are most likely to pronounce this specific word

The Voice Recognition Process

- Regional accents
  - The same word is pronounced completely differently depending on whether you are from London, Liverpool, Newcastle, Edinburgh, Dublin, Sydney, New York, or New Orleans
- Foreigners speaking the language
  - the very same English letter combinations and word will sound even more different when spoken by a Greek, a German or a Japanese native speaker

The Voice Recognition Process

- recognition lexica are augmented with additional “pronunciations” for each problematic word
  - 3 different versions of the same word spoken by different people are still recognised as one and the same word! (Yay!)
- only for words relevant to your specific app (and domain) and for accents representative of your end-user population

The Voice Recognition Process

- If an app is going to be used mainly in England, you're better off covering Punjabi and Chinese pronunciations of your English app words rather than Japanese or German variants
  - There will of course be Japanese and German users of your system, but they represent a much smaller percentage of your user population and we can’t have everything!!

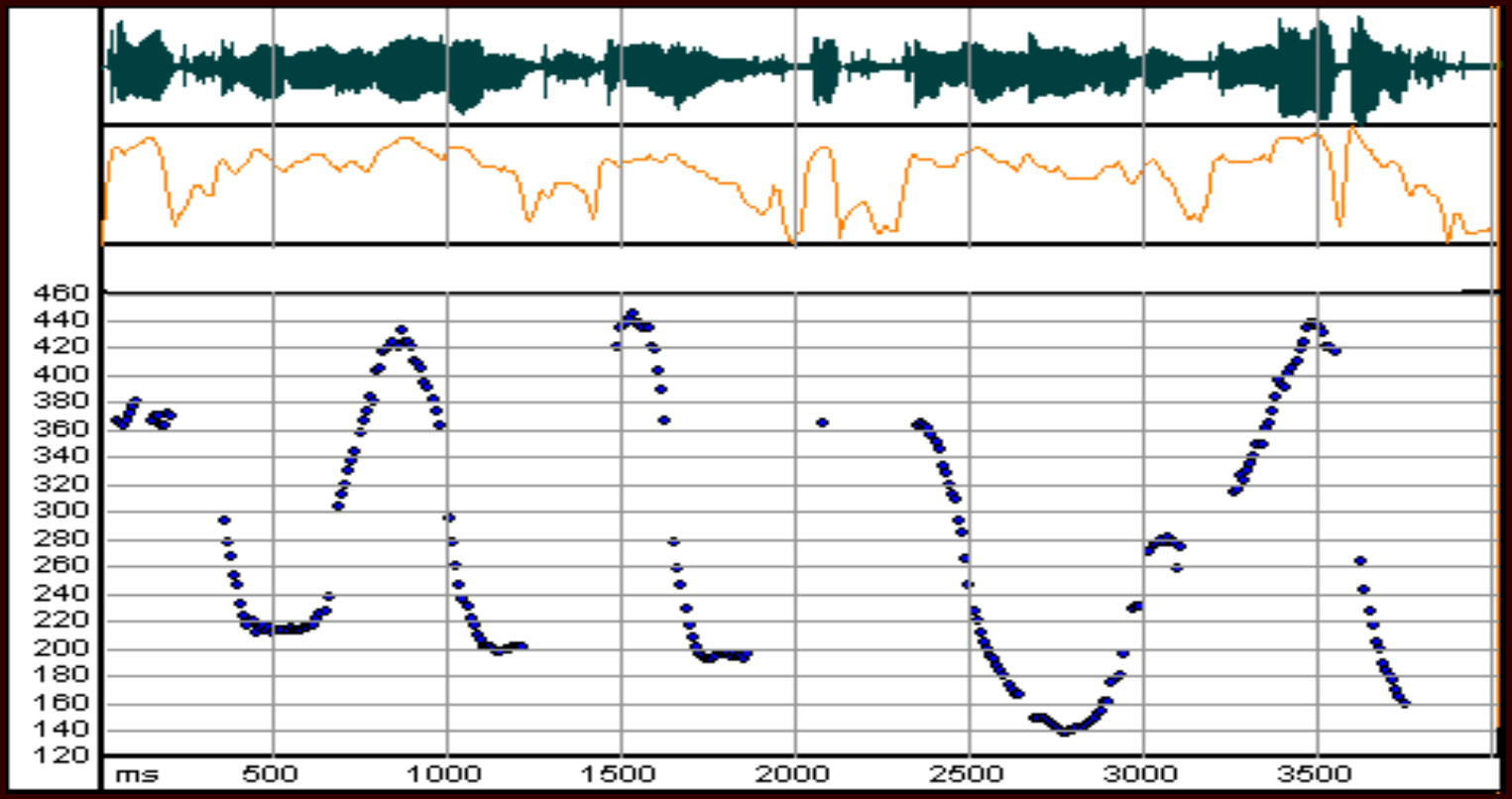
The Voice Recognition Process

- Recognition is usually based on a “lexicon”
- but the whole process is actually statistical
  - The recogniser has to figure out what you’re saying, chopping this wave signal up into parts, each representing a word that makes sense IN THE CONTEXT OF the surrounding words

The Voice Recognition Process

- Unfortunately, the same signal can potentially be chopped up in several different ways
- each representing a different string of words
- and of course a different meaning!

The Voice Rec Process



"How to recognise speech" vs "How to wreck a nice beach"!

The Voice Recognition Process

- Ambiguity of interpretation
  - “How to recognise speech”

    OR
  - “How to wreck a nice beach” !!!
- Thankfully, the app usually defines the domain (science vs holidays)
- Which in turn defines lexicon & grammar
  - “Recognise” + “speech” are more likely than “wreck” + “beach”

Manual Grammar-based Rec

- hand-crafted **lexica**: words + their pronunciations
- hand-crafted **grammars**: word combos that make up legal sentences in the language
- "manual" approach is sufficient for very **limited domains** (e.g. ordering a printer or getting your account balance)
  - Lexica and grammars describe most relevant phrases that are likely to be spoken by the user population
  - Any other phrases will be just irrelevant one-offs that can be safely(?) ignored

Statistical Voice Recognition

- For anything more complex and advanced (smartphone apps)
- Collecting **large amounts of real-world speech data** (human-human dialogues, human-machine dialogues) for training
- machine learning of the most likely and meaningful combinations of sounds for that language
- Much more robust and accurate

Statistical Voice Recognition

- much better coverage of what people actually say (descriptive)
  - Vs what the developer thinks that people should say (prescriptive)
- They can accurately **predict** sound and word combinations that could not have been pre-programmed in a hand-crafted grammar!

Statistical Recognition Grammar

the real world vs the ideal world

- **coverage**
  - synonyms vs limited keywords;
  - spontaneous speech [erm, uhm, I'd like – tell me ...] vs grammatical sentences
  - Colloqualisms vs „proper English“
- **Grammar rigidity**
  - Whole sentences vs single command words

Manual Voice Recognition Wins!

- Automated **Call Centre apps** use the **manual** approach
  - Statistical Data collection and data analysis is very time-consuming
  - Data cost and data privacy issues are often prohibitive
- **Smartphone apps** are based on the **statistical** approach
  - Only works for standard language use (feta vs fetish!)
  - Everything can be a wild guess!

The Future: Voice Assistants with an attitude!

- Context-based voice recognition FTW!
  - Robust statistical recognition + device / use context and user data
  - voice recognition → intention understanding → reading your soul? :)
  - Reaction –> initiative and self-activation!
    - Unsolicited (!) Reminders, warnings, recommendations, suggestions
    - Multimodality, multi-device, multi-location

A good Voice recognition app

- Usability
  - Simple to use, intuitive, self-explanatory, learnable
- User acceptability
  - feeling understood (without expecting too much off the system: „**It's better to be a good machine than a bad person**“) and served
  - Liking the app and wanting to come back!
  - At the very least, the user should NOT feel irritated!