Automatic Speech Recognition: where AI meets Human Intelligence

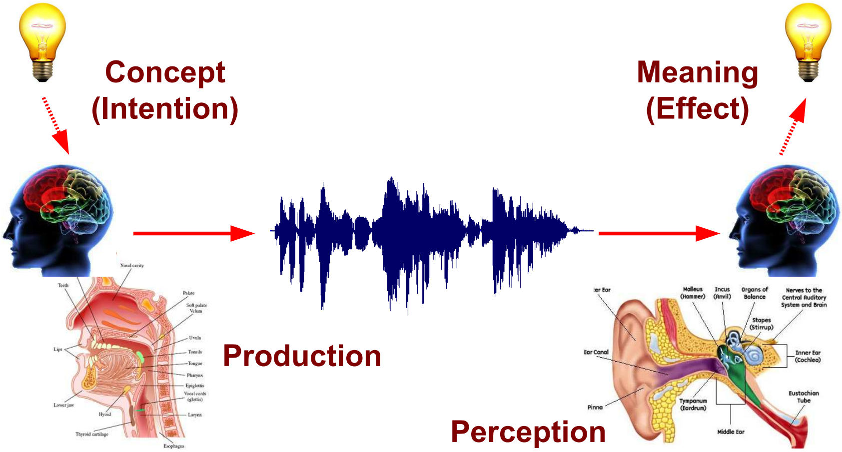
Phil Weber – Aston University

Forensic Speech Science Laboratory
Forensic Data Science Laboratory

p.weber1@aston.ac.uk | https://weberph.bitbucket.io

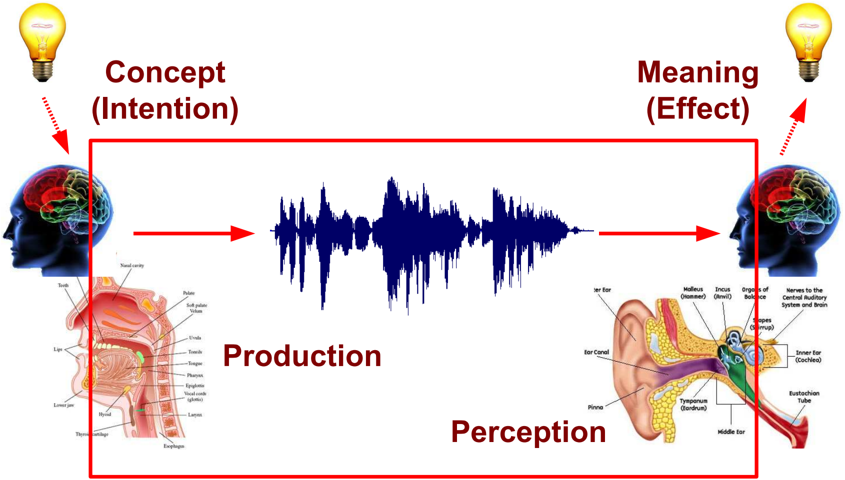
11 February 2020 – BrumAI

Telepathy?

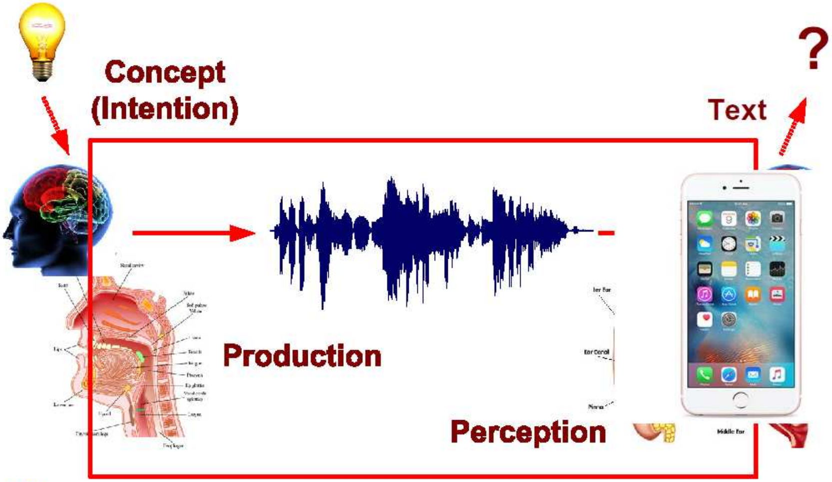


Concept
(Intention)
Meaning
(Effect)
Production
Perception
Nasal cavity
Teeth
Palate
Soft palate
Velum
Uvula
Lips
Tonsils
Tongue
Pharynx
Epiglottis
Vocal cords
(glottis)
Lower jaw
Hyoid
Larynx
Thyroid cartilage
Esophagus
ter Ear
Malleus
(Hammer)
Incus
(Anvil)
Organs of
Balance
Nerves to the
Central Auditory
System and Brain
Stapes
(Stirrup)
Ear Canal
Inner Ear
(Cochlea)
Eustachian
Tube
Tympanum
(Eardrum)
Pinna
Middle Ear

Human Speech Recognition



Automatic Speech Recognition

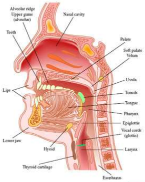


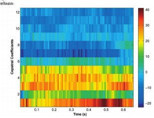

STRAW POLL

Automatic Speech Recognition

How to Make an Automatic Speech Recogniser

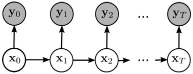
1. Feature Extraction



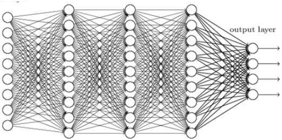


MFCCs
LPCs
...

2. Modelling



Hidden Markov Model (HMM)



Deep Neural Network (DNN)

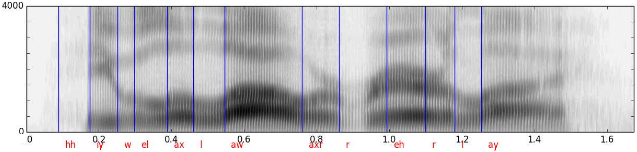
...

3. Decoding

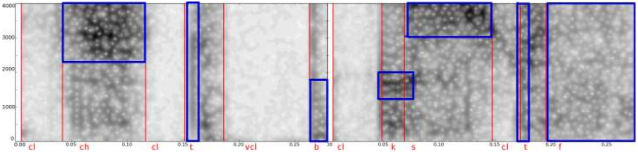
Feature Extraction – Human Speech Audio

Waveform – amplitude over time, sum of many frequencies.

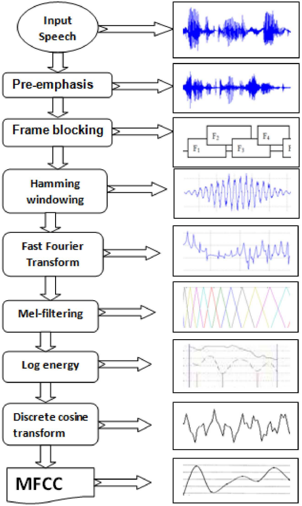
Spectrogram – frequency representation – **Vowels**.

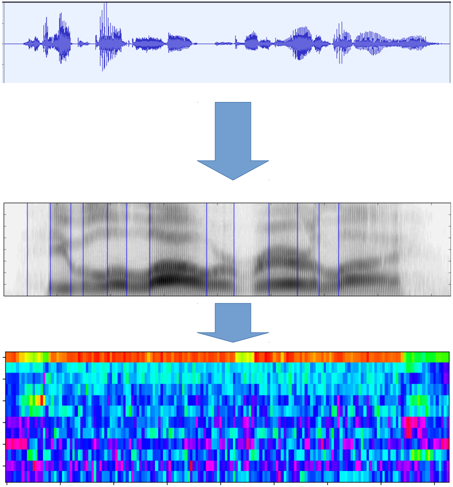


Spectrogram – frequency representation – **Consonants**.



Feature Extraction – Standardised Process (MFCCs)

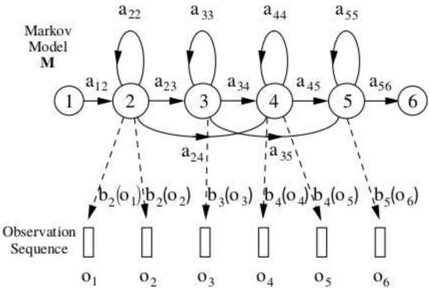




Performance analysis of isolated Bangla speech recognition system using Hidden Markov Model, Abdullah-al-MAMUN, Firoz Mahmud, IJSER 6(1):540-545, 2015.

Modelling – Until (Relatively) Recently

1. Template matching, dynamic time warping: to around 1980s.
   — limited scope, e.g. digit recognition.
2. Probabilistic modelling:
   — Hidden Markov Models dominate to 2000s.

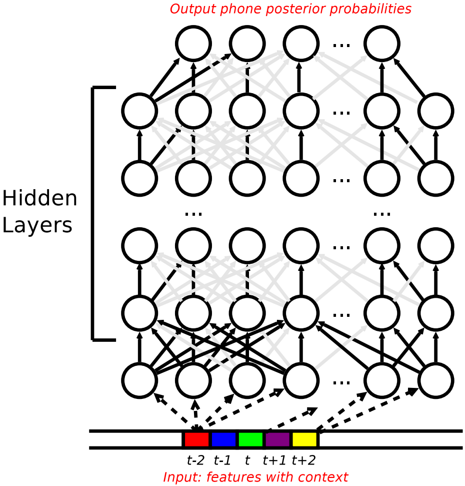


Source: The HTK Book, 2002, Fig 1.3.

3. Deep Neural Networks dominate from late 2000s.

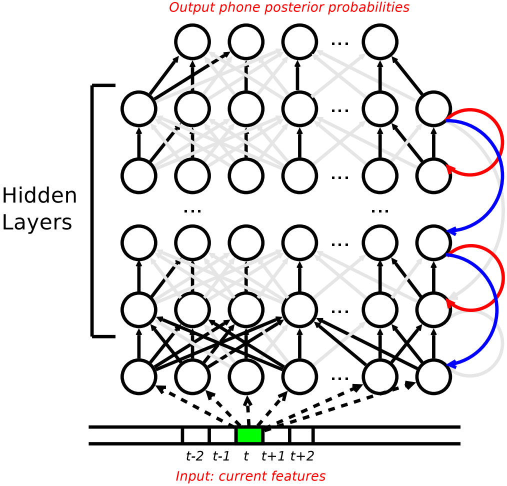
Modelling – Present

Deep Neural Networks effectively model complex acoustic distributions.



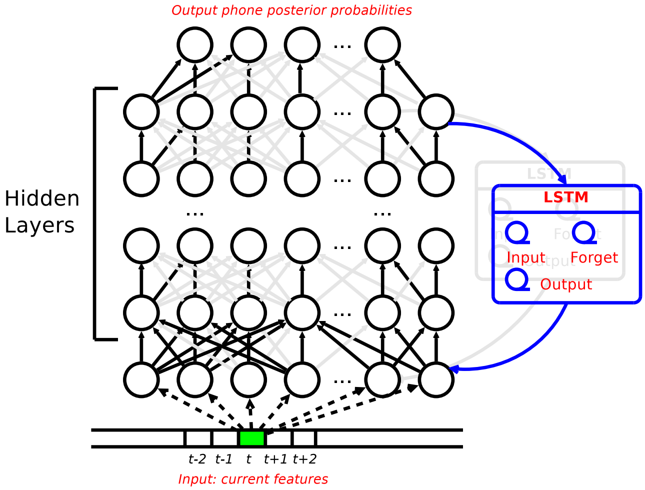
Modelling – Present

Recurrent Neural Networks additionally model (some) temporal aspects.

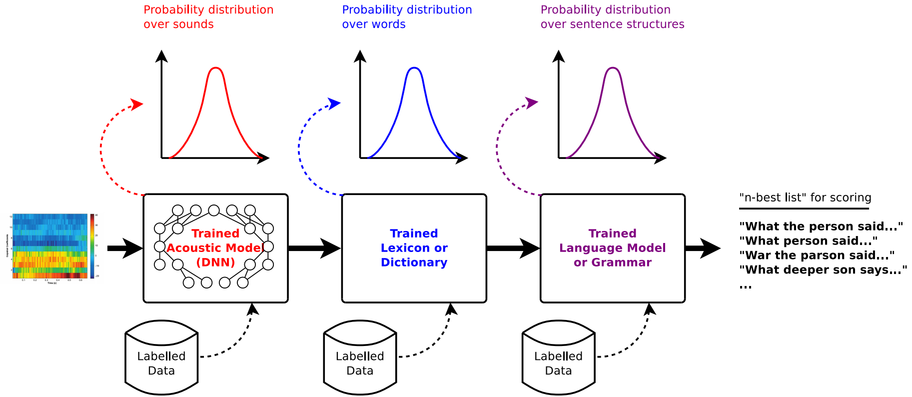


Modelling – Present

Long Short-Term Memory adds finer control of temporal modelling.



Decoding



Evaluate probability of 'all possible' sequences of
sounds into words, into sentences, into ... meaning?

State of the Art

State of the Art (examples):

- **Google** 2.6% Word Error Rate (WER) on 'LibriSpeech960h' dataset:
  - SpecAugment + Listen Attend Spell:
  - 6.8% on **conversational** speech ('Switchboard').
- **Facebook** 3.5% WER:
  - 4-layer CD-HMM-LSTM, 800 'memory cells' per layer, 6,133 outputs
  - Language Model: 80,000 words, 200 million n-grams.

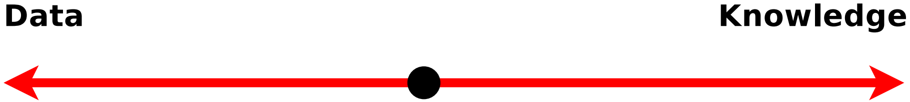
Toolkits: KALDI

Issues? — What about the real world?

1. *"PwC: Lack of trust in AI assistants like Alexa could hinder adoption".*
2. 100,000s to millions of parameters ... Why?
3. What happened to our knowledge of human speech?

Park et al., SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition, Interspeech, 2019.
Serdyu et al., Towards end-to-end spoken language understanding, Facebook AI Research, c2018.

Where is the balance?



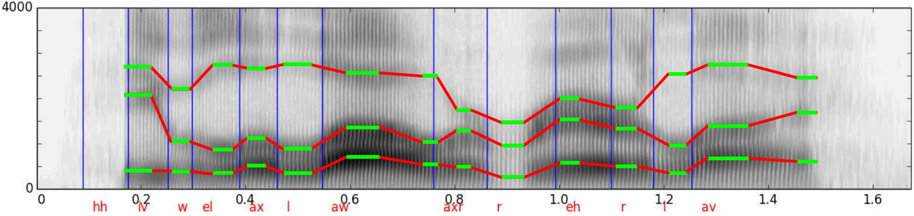
Models inspired by human speech perception and production

(Work with colleagues at the University of Birmingham)

Models Inspired by Human Speech

Continuous-State Hidden Markov Model (CSHMM).

Vowels: 'Dwell-Transition' model tracking formants, or ...

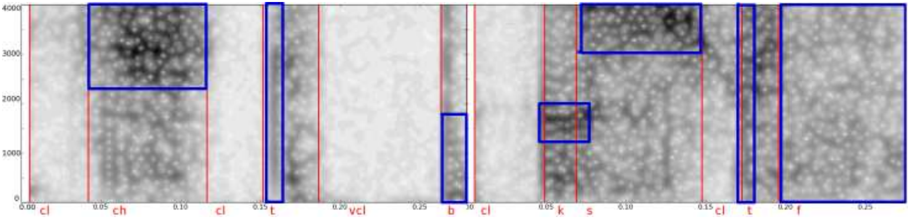


Human voice stationary in target vowel, smooth transition to next.

Models Inspired by Human Speech

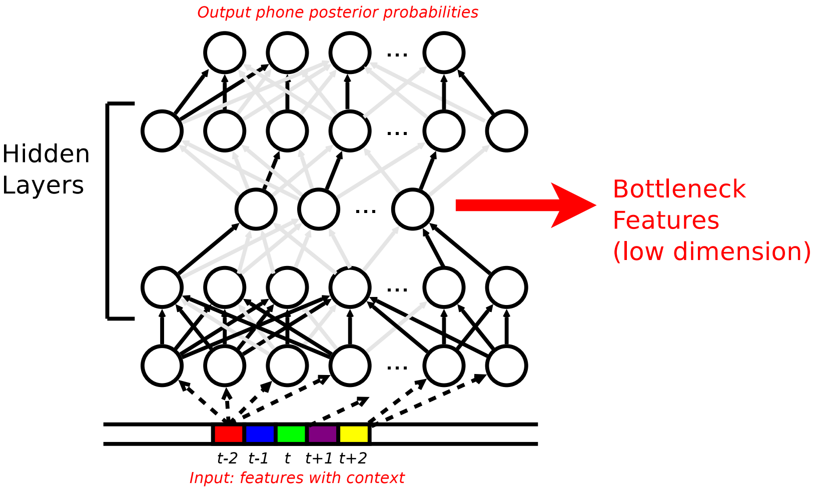
Continuous-State Hidden Markov Model (CSHMM).

Consonants: 'Dwell-only' model tracking high-energy bands



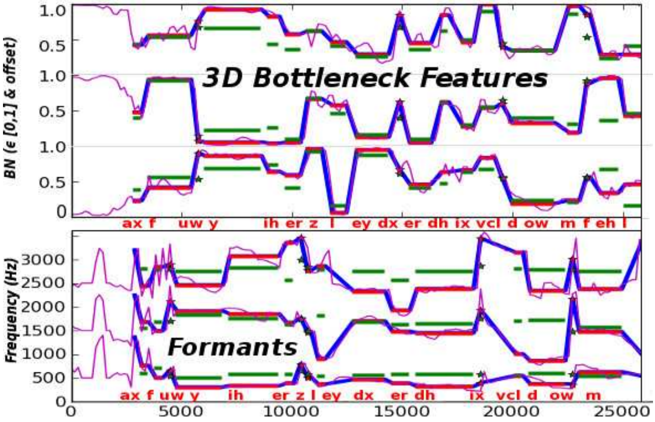
Turbulence in specific frequency bands, abrupt transitions.

Bottleneck Network Feature Extractor



Bottleneck Features fit the CSHMM well

Top: Bottleneck features (pink) fit the model much better (blue).



Bottom: Formants (pink) extremely variable, especially for non-vowels. Initial phone estimates (green) not discriminating.

Phone Recognition Results

Phone error rates (% sounds classified correctly):

Model	Features	Dimension	%Err	# Parameters
DNN state-of-art	MFCC	39	**18.0**	??? $\times 10^6$
HMM traditional	MFCC	39	29.1	1.4×10^7
HMM traditional	BNF	9	29.4	2.3×10^5
CSHMM faithful	BNF	9	36.5	**535**

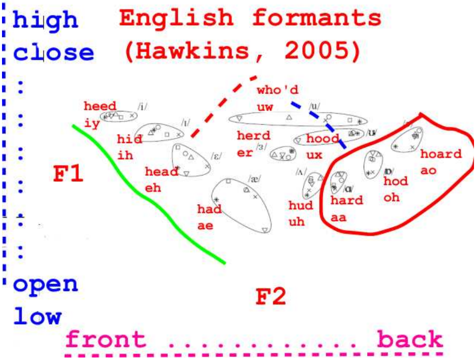
- Bottleneck features perform equally well with lower dimension.
- 'Faithful' CSHMM model not (yet) competitive with state of the art.
- — but using very few parameters.

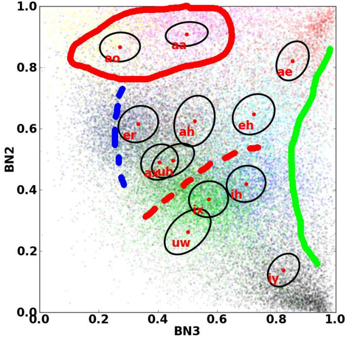
Relating human speech science and automatic speech models

(Work with colleagues at the University of Birmingham)

Automatic : Human Speech recognition – Vowels

Human vowel space diagram (left) *vs* two of bottleneck feature (2 of 3D):



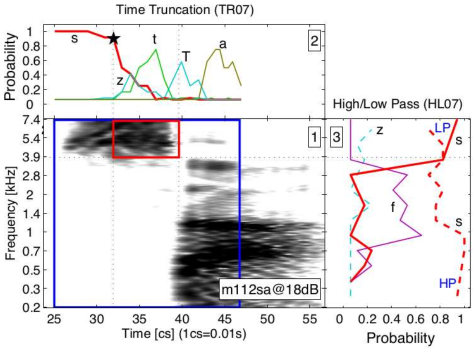


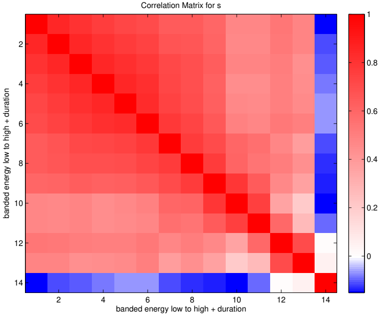
[Hawkins, 2005]

Neural network (right) learns something very similar.

Automatic : Human Speech – Consonants

Research in human perception (left) tells us what energy cues are important:



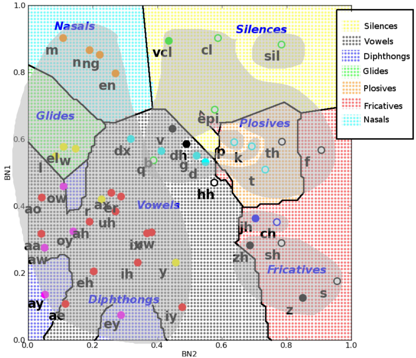


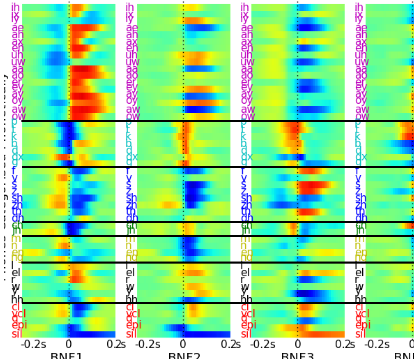
[Li & Allen, 2012]

Features based on these bands work best for CSHMM recognition.
Similar features appear in models learned from data (right).

Automatic : Human Speech – All Sounds

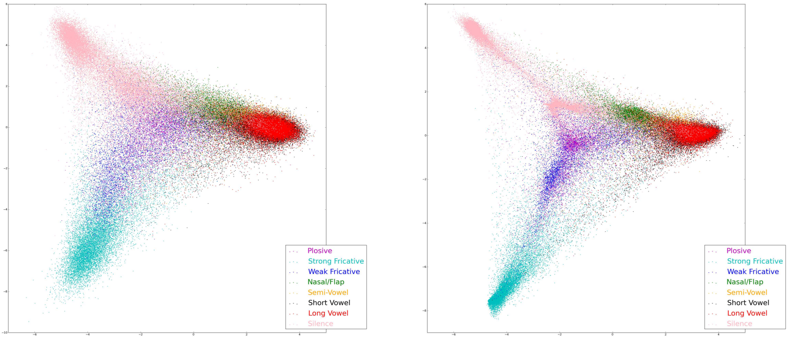
- 2 dimension bottleneck feature mapping (left) is relatable to human speech production and perception,
- 9 dimension neuron activations over time (right) recall the science of perceptual cues and suggest learned roles for different neurons.





Automatic : Human Speech – Other Layers

- LDA compressions of 512 dimension (non-bottleneck) hidden layers show interpretable mappings,
- increasing sharpness with closeness to the output layer (left to right) recalls theories relating DNNs to cognition.

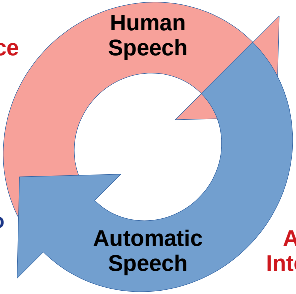


Linxue Bai, PhD Thesis, 2017, UK Speech 2019

So What?

- **Human speech science informs ASR development**
- **ASR development can inform human speech science**
- **What we learned may help us do better with ASR**

Human Intelligence

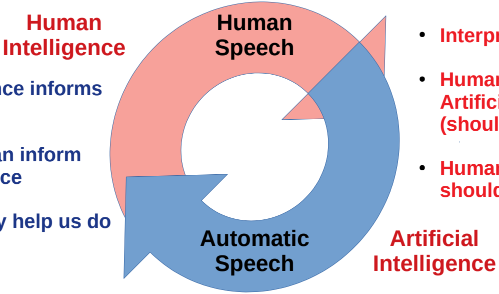


Artificial Intelligence

- **Interpretable AI is crucial**
- **Human [intelligence] and Artificial [intelligence] can (should) inform each other**
- **Humans and technologists should talk!**

So What?

- **Human speech science informs ASR development**
- **ASR development can inform human speech science**
- **What we learned may help us do better with ASR**



- **Interpretable AI is crucial**
- **Human [intelligence] and Artificial [intelligence] can (should) inform each other**
- **Humans and technologists should talk!**

Thank you!

`p.weber1@aston.ac.uk | https://weberph.bitbucket.io`