Voice Banking and Voice Reconstruction

Christophe Veaux, Pierre Lanchantin, Gergely Bakos, Junichi Yamagishi, Simon King

Edinburgh, 28 June 2016

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

THE EUAN MACDONALD CENTRE
FOR MOTOR NEURONE DISEASE RESEARCH

Building personalised VOCAs

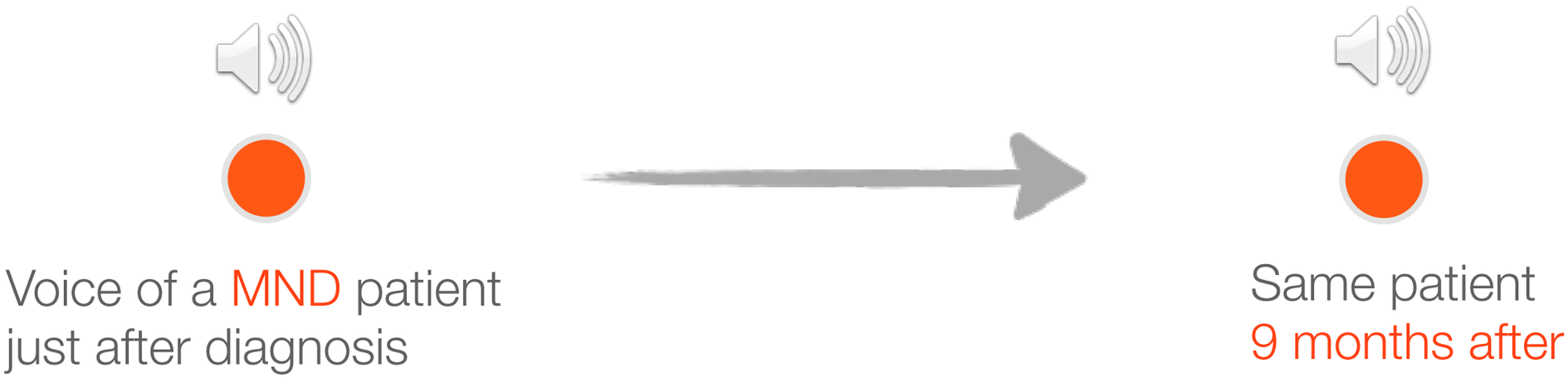
Natural Speech Technology
Edinburgh – Cambridge – Sheffield

Degenerative diseases leading to dysarthria (MND, Parkinson, MS)

- Dysarthria influences all levels of speech production: phonation, respiration, articulation, resonance and prosody

- The deteriorations proceed individually and the variation in the quality of speech problems is large between adults [Yorkston et al., 1993]

- Word intelligibility can decrease from 95 % to 88 % during a 6 months period [Watts and Vanryckeghem, 2001)]



Building personalised VOCAs

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Voice Output Communication Aid (VOCA)

- Current VOCAs only provide a small range of voice
- not a good match of age, accent, speaking style

Personalisation of VOCAs

- facilitate social interaction
- Speech is not just a mean of communication but also a display of personal and group identity
- greater dignity and improved self-identity for the individual and their family

Personalised voices is a long standing request from VOCA users

Building personalised VOCAs

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Voice banking

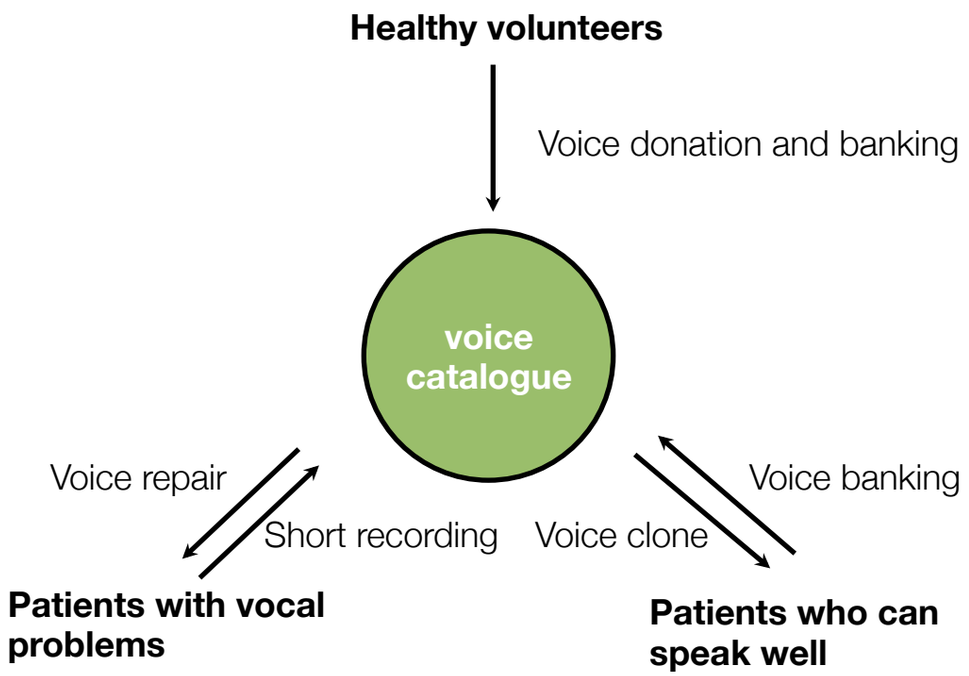
- Capturing the voice before it starts to degrade

Voice donation

- In order to build average voice models for speaker adaptation

Voice reconstruction

- For patients who already have speech disorders at the time of recording



Pre-NST

Natural Speech Technology

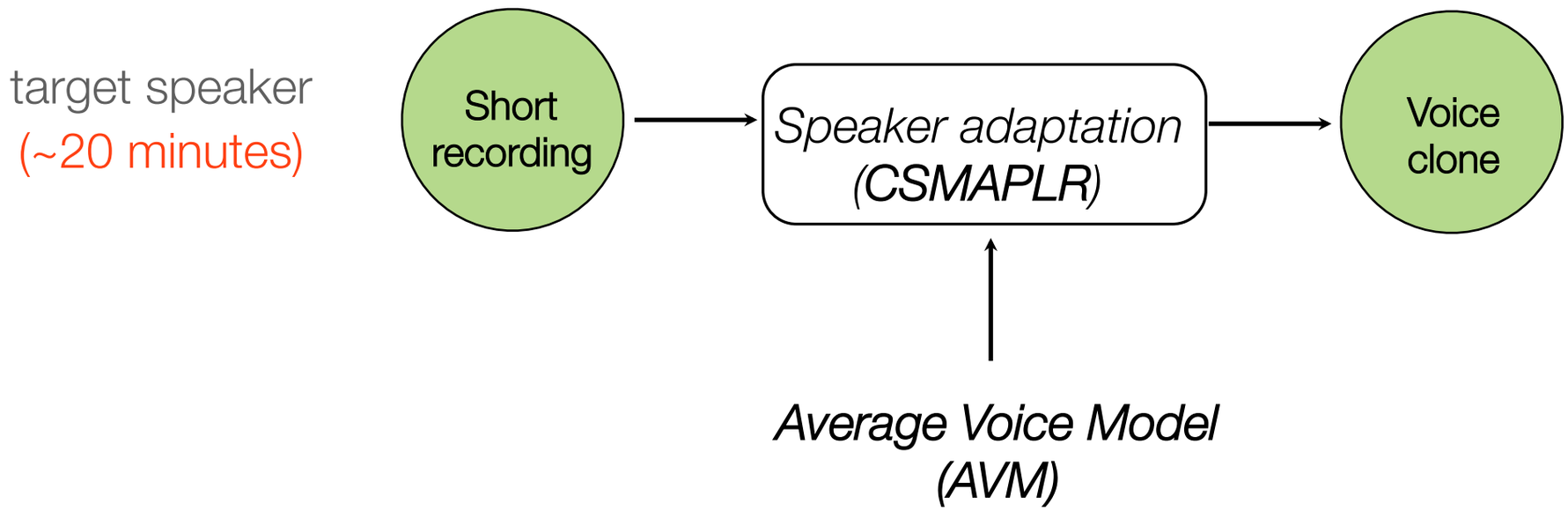
Edinburgh – Cambridge – Sheffield

Pilot study on Voice banking / Voice Reconstruction

- Recording of 100 donors and 7 patients

HMM based speech synthesis for voice building

- helps to reduce complexity and to increase the flexibility of the voice building process (*adaptation of pre-trained AVMs, voice reconstruction*)



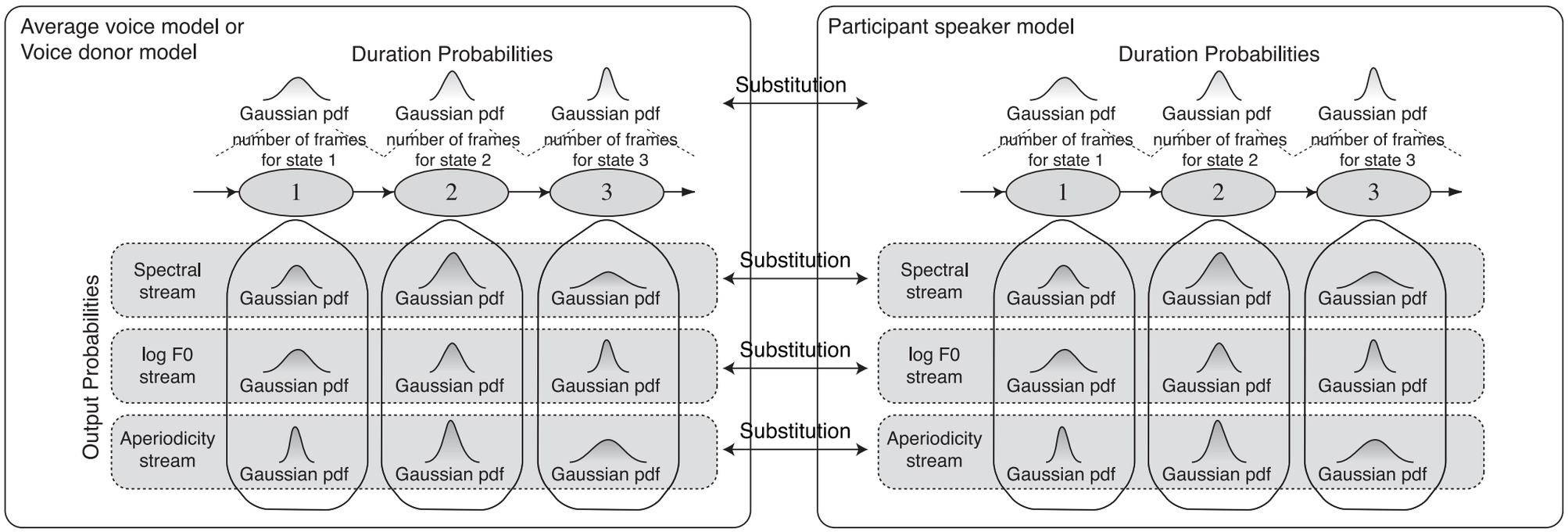
Pre-NST

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

“Manual” voice reconstruction

- fixing statistical models of the patient’s voice clone so that they can generate natural sounding speech while keeping speaker identity



Voice banking project (NST)

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Objectives

- Automates voice reconstruction
- Better voice similarity through better coverage of accents
- Development of tools for speech and language therapists as well as VOCA app for patients

Large scale clinical trial

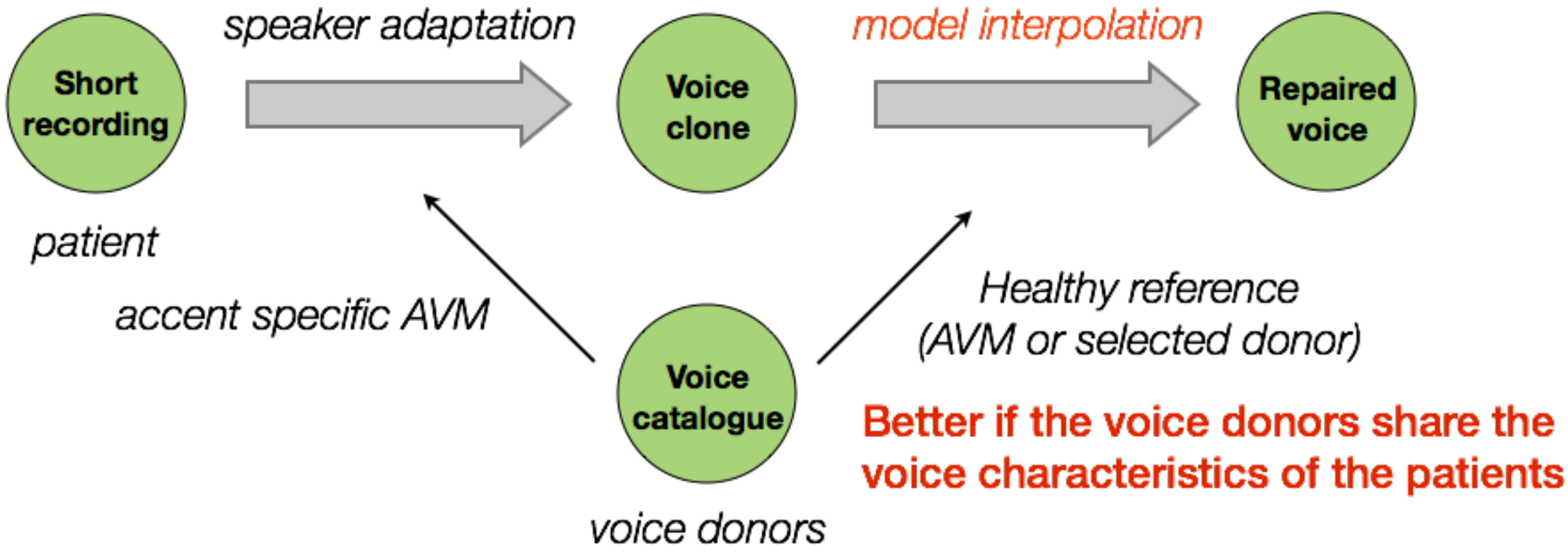
- More than 900 healthy donors voices
- More than 100 patients
- Feedback from all patients and their families on the personalised voice and its impact on their quality of life

Voice Reconstruction

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

First approach: model interpolation

- Post-process after speaker adaptation
- Two methods:
  - Manual tailoring of the interpolation coefficients (Pre-NST)
  - Automatic interpolation using KLD-based confidence measure

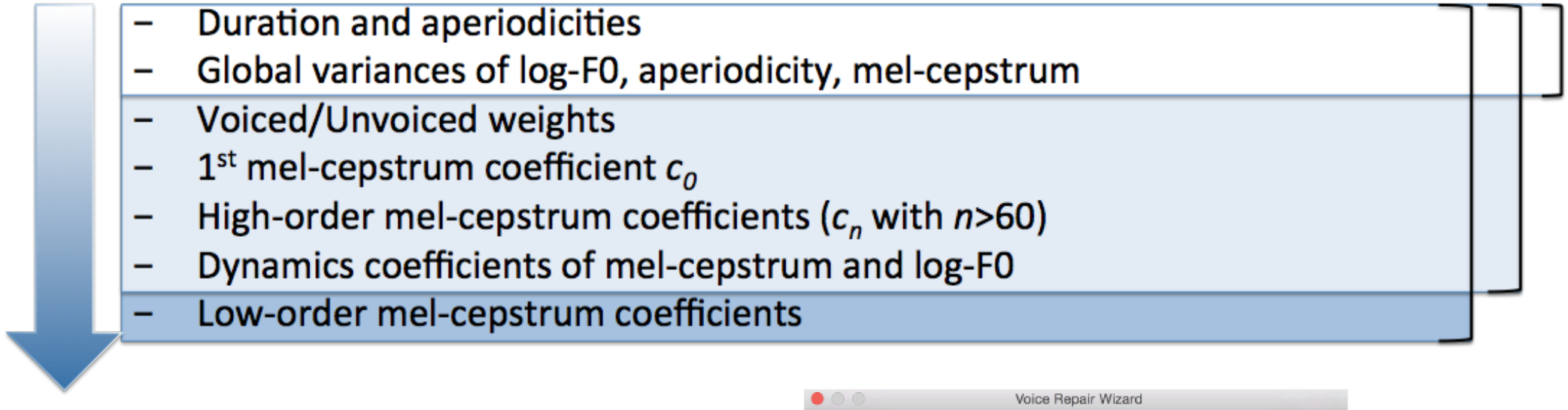


Voice Reconstruction

Natural Speech Technology

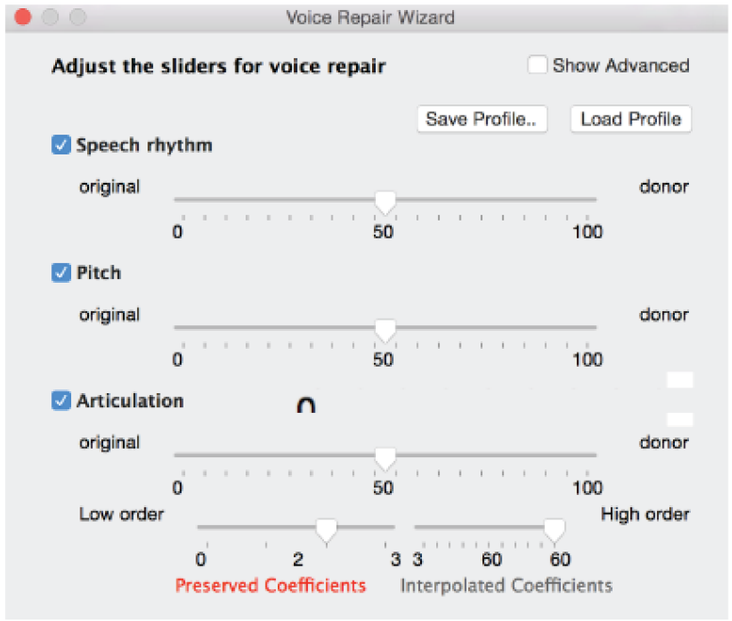
Edinburgh – Cambridge – Sheffield

- Original voice (MND patient)
- Interpolation of the *less* speaker-dependant model components



Impact on speaker identity

Interpolation weights can be adjusted manually by a SLT



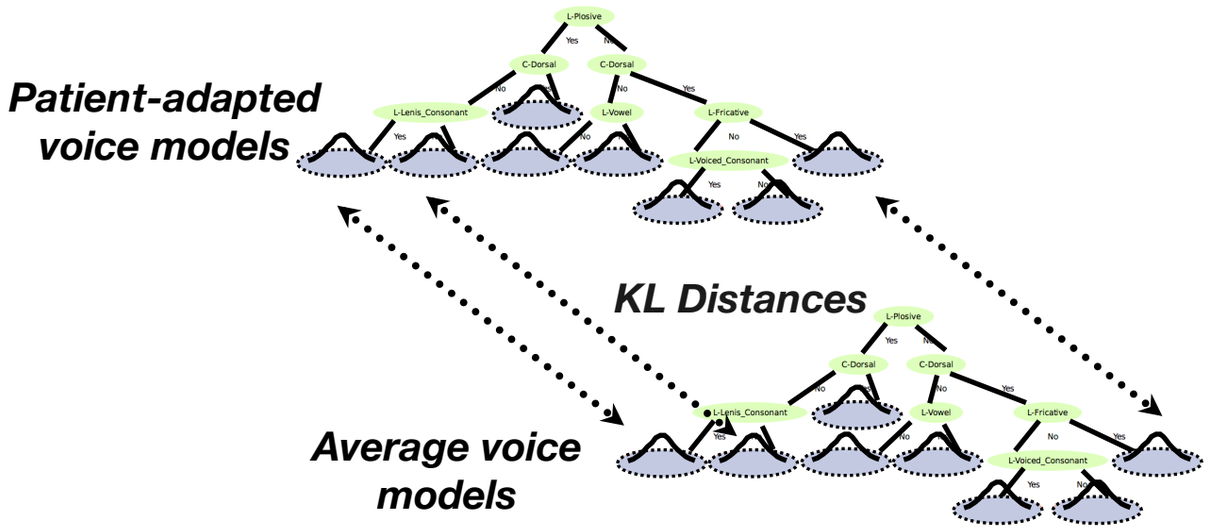
Voice Reconstruction

Natural Speech Technology

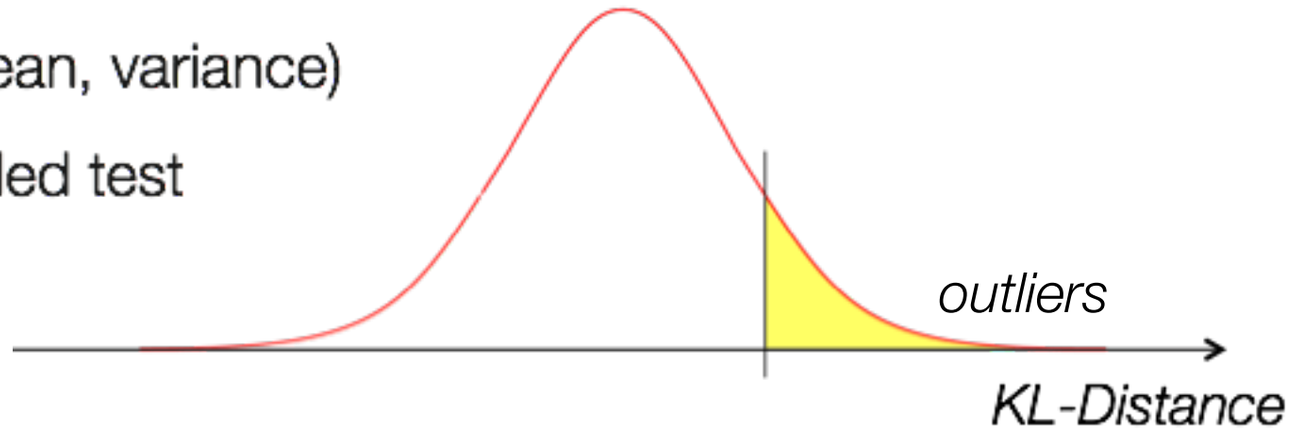
Edinburgh – Cambridge – Sheffield

KLD-based confidence measures

- KL distances between the context-dependent models of the patient voice model and the AVM



- Statistics of KLD between models (mean, variance)
- Confidence measure based in one-tailed test
- Interpolation weights for each model



Voice Reconstruction

Natural Speech Technology

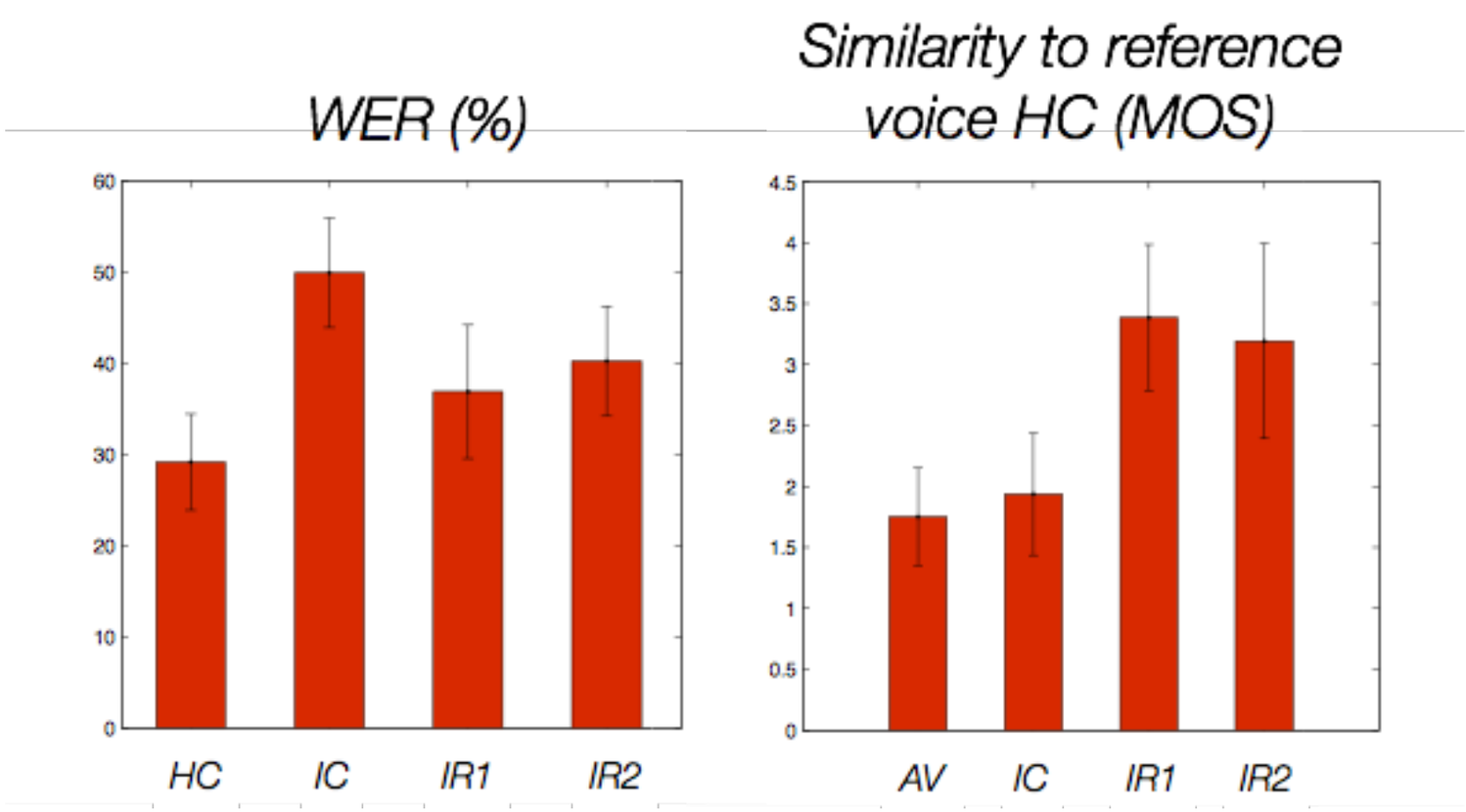
Edinburgh – Cambridge – Sheffield

Listening tests (40 listeners)

- Two recordings of a same MND patient
- one “healthy voice” recording (just after diagnosis)
- one “disordered voice” recording (10 months later)

Compared synthetic voices:

- HC: Voice clone of “healthy speech”
- IC: Voice clone of “impaired speech”
- IR1: **Manual** (tailored) model interpolation
- IR2: **Automatic** model interpolation
- AV: Average voice model



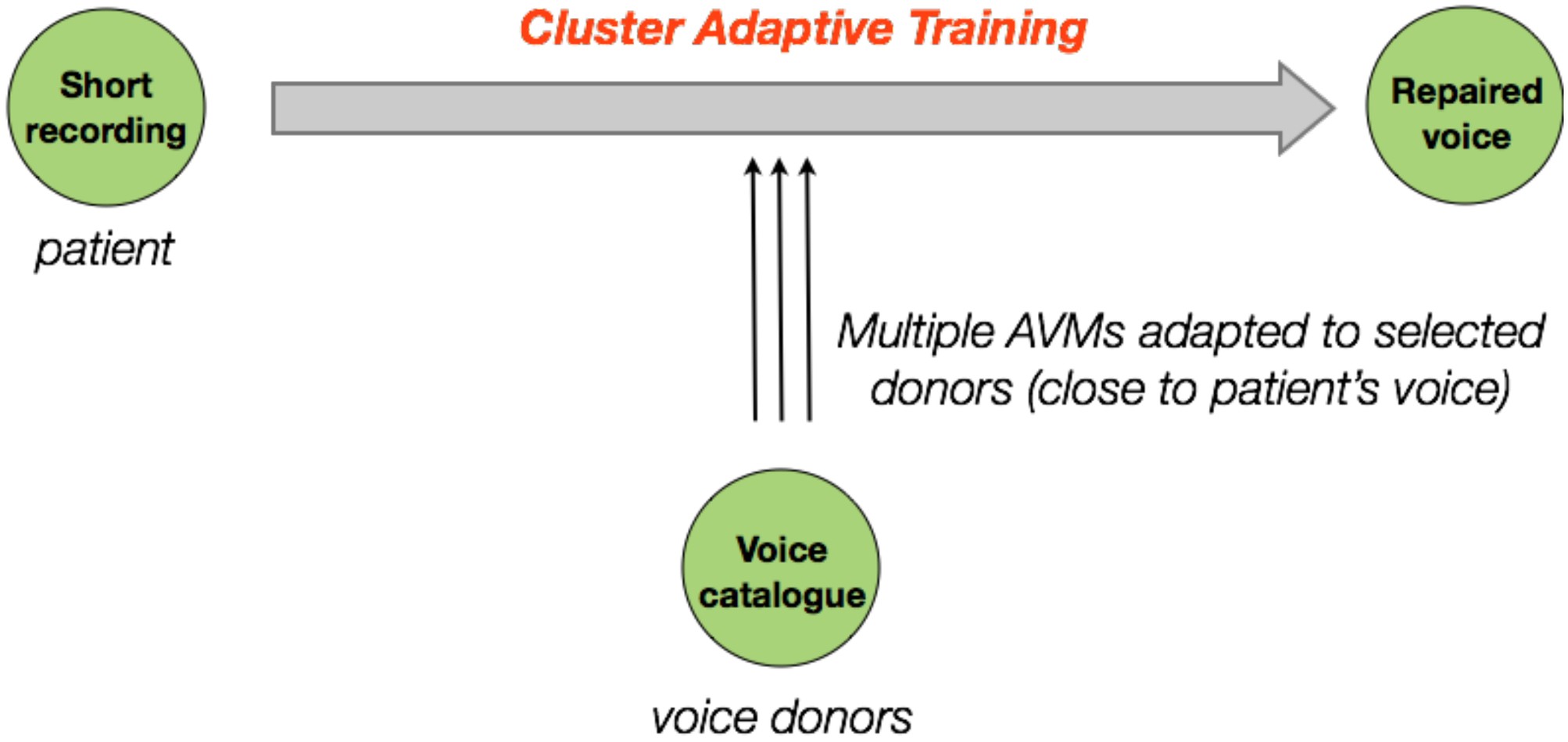
Voice Reconstruction

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Second approach: multiple AVMs interpolation (hybrid between AVM and CAT)

- the adapted mean vector of a component is interpolated in an eigenspace spanned by the cluster mean vectors
- but clusters are AVMs which can be tuned towards the target before interpolation

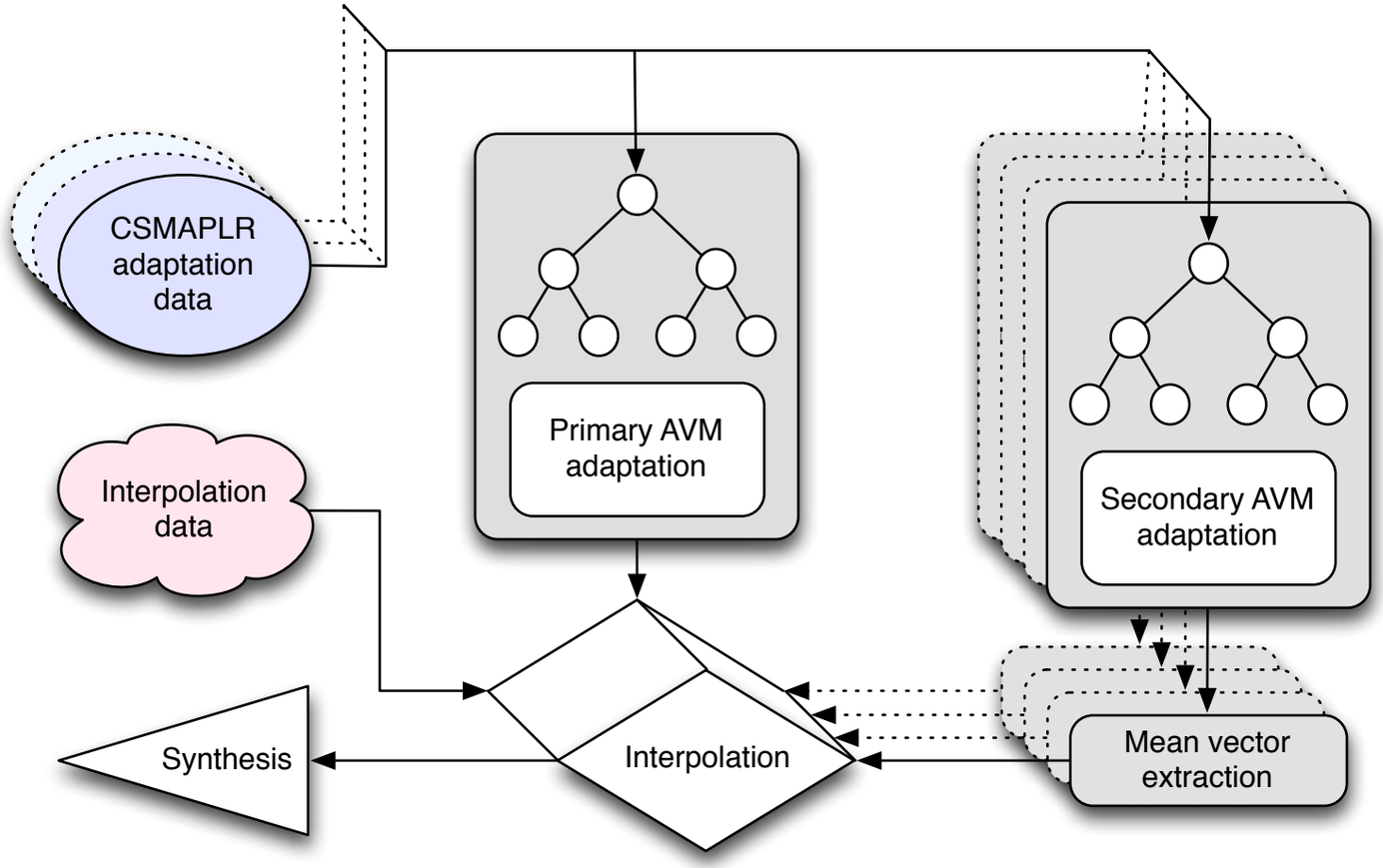


Multiple AVM interpolation

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Interpolation eigenspace can be designed using different combination of AVM/target voices
- Interpolation can be done in a clean space by selecting healthy target voices close to the disordered one
- Constrained interpolation: limited degrees of freedom helps to reduce the “noise” due to disorders in the adaptation data



Multiple AVM interpolation

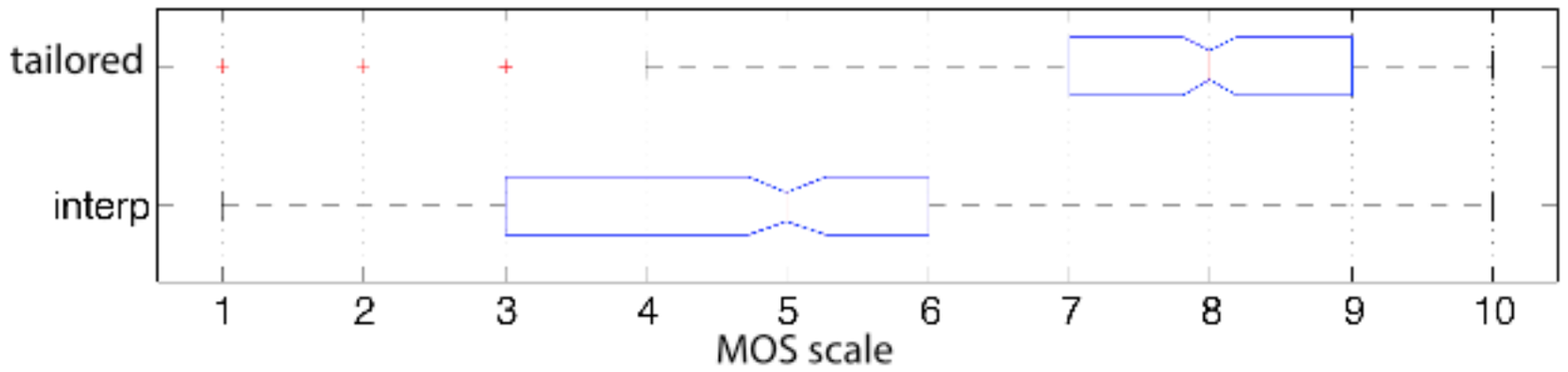
Natural Speech Technology

Edinburgh – Cambridge – Sheffield

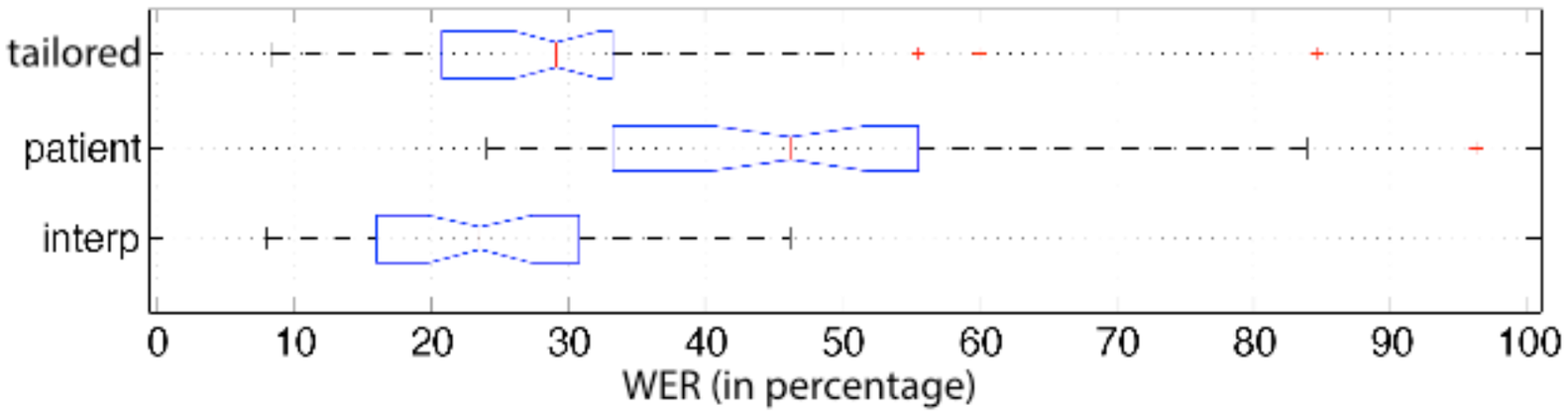
Listening tests (38 listeners):

- interp: Multiple AVM interpolation
- tailored: manually reconstructed by speech therapist

Similarity Test



Intelligibility Test

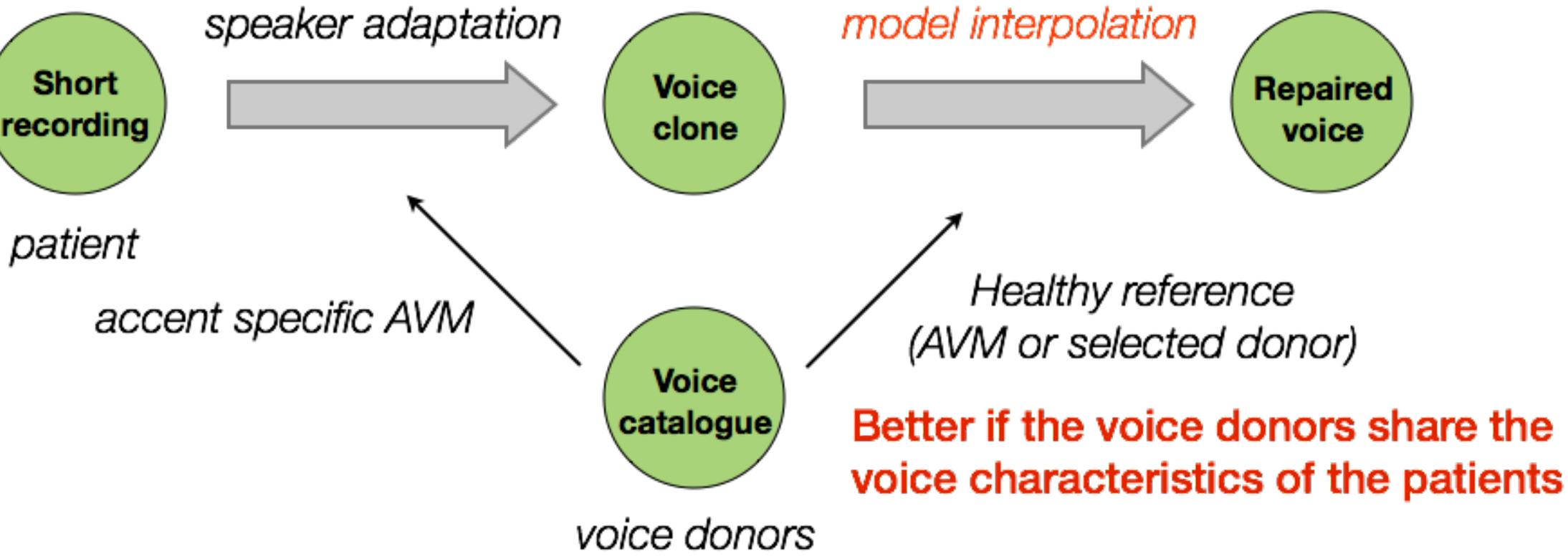


Voice Reconstruction

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Selected approach: model interpolation with KLD-based confidence measure



The need of accent-specific AVMs

Natural Speech Technology

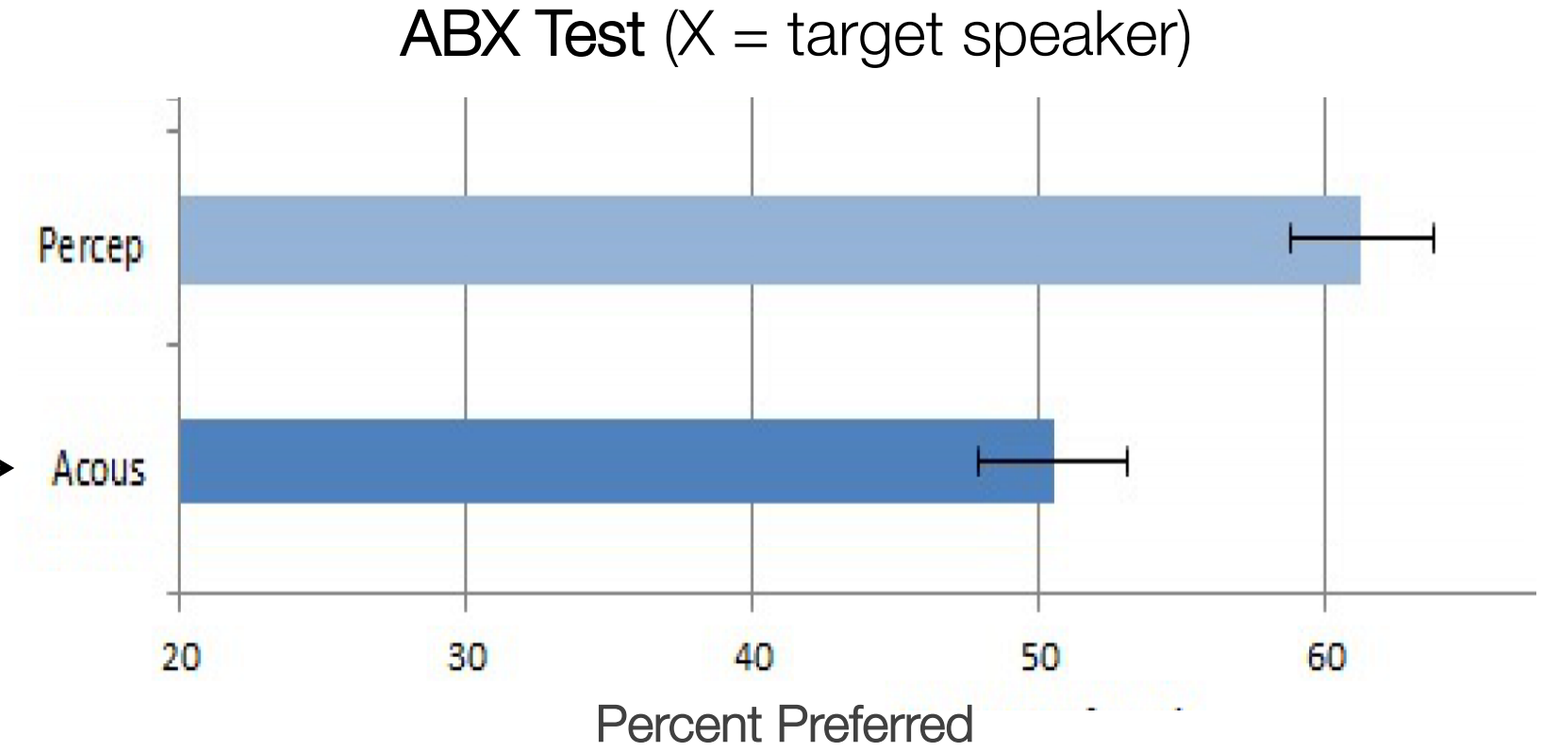
Edinburgh – Cambridge – Sheffield

Improves speaker similarity

- An average voice learned over a **small number** of speakers **perceptually close to the target** gives better results than a **large average voice model**

10 closest donors selected using **perceptual similarity scores**

Global average voice trained over **70 unselected donors**



Needed for voice reconstruction

- Model interpolation require an AVM close to the patient's voice or a set of AVMs

Large scale voice recordings

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Record a large number of speakers with different age, gender and accent

Semi-anechoic chamber of School of Informatics,

Anne Rowling Regenerative Neurology Clinic (Jan 2013): Voice banking studio

Large scale voice recordings

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Record a large number of speakers with different age, gender and accent

Corpus design

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Text materials

- 400 sentences in average for each speaker (1 hour recording session)
- Sentences taken from a corpus of newspaper articles (1300000 sentences)
- Rainbow passage (covers a wide variety of consonant clusters)
- Accent elicitation (phonetic shifts) sentences from the Speech Accent Archive

Metadata

- Age, gender, accent (from childhood location), occupation (education level)

Corpus design

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Specifically designed for the training of accent specific average voice models

- Different lexicons (change of phonetic inventory and segmental structure)

  Combilex lexicon (RPX, Scottish English, US English)

- Each speaker records a different text script
- Phonetic and prosodic coverage is optimized across several speakers

Greedy selection of the best set of sentences that

- Maximize the trigram and phone coverage (Most frequent unit first)
- Balance the distribution of number of syllables and phrases (Less frequent unit first)

Coverage optimization favors more complex sentences

➡ Readability constraints

Corpus design

Natural Speech Technology

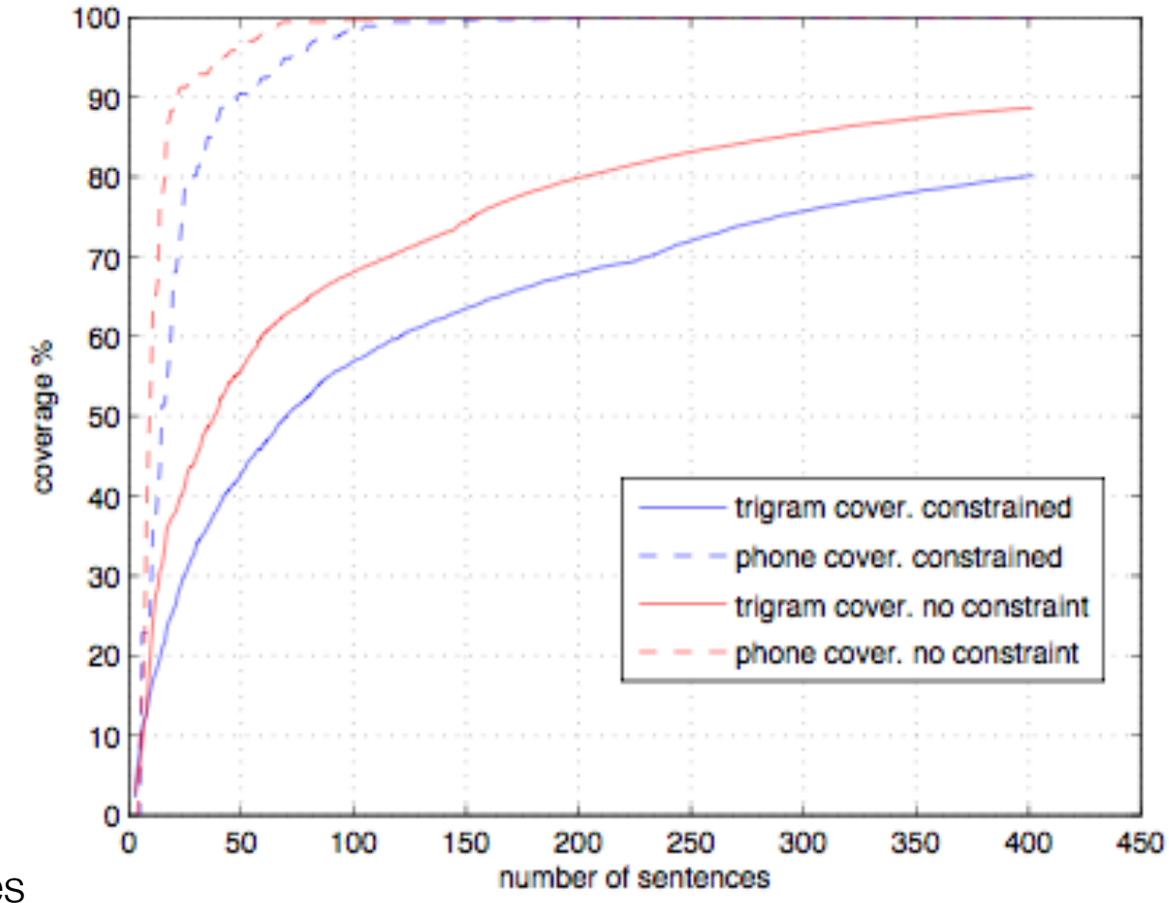
Edinburgh – Cambridge – Sheffield

Principal correlates of sentence complexity [Tanguy & Tulechki, 2009]

- Number of words per sentence
- Number of syllables per sentence
- Length of noun phrases
- *Syntactic complexity*

  (POS trigram frequencies)
- *Lexical complexity*

  (word frequencies)



Readability filtering ratio ~ 30 %

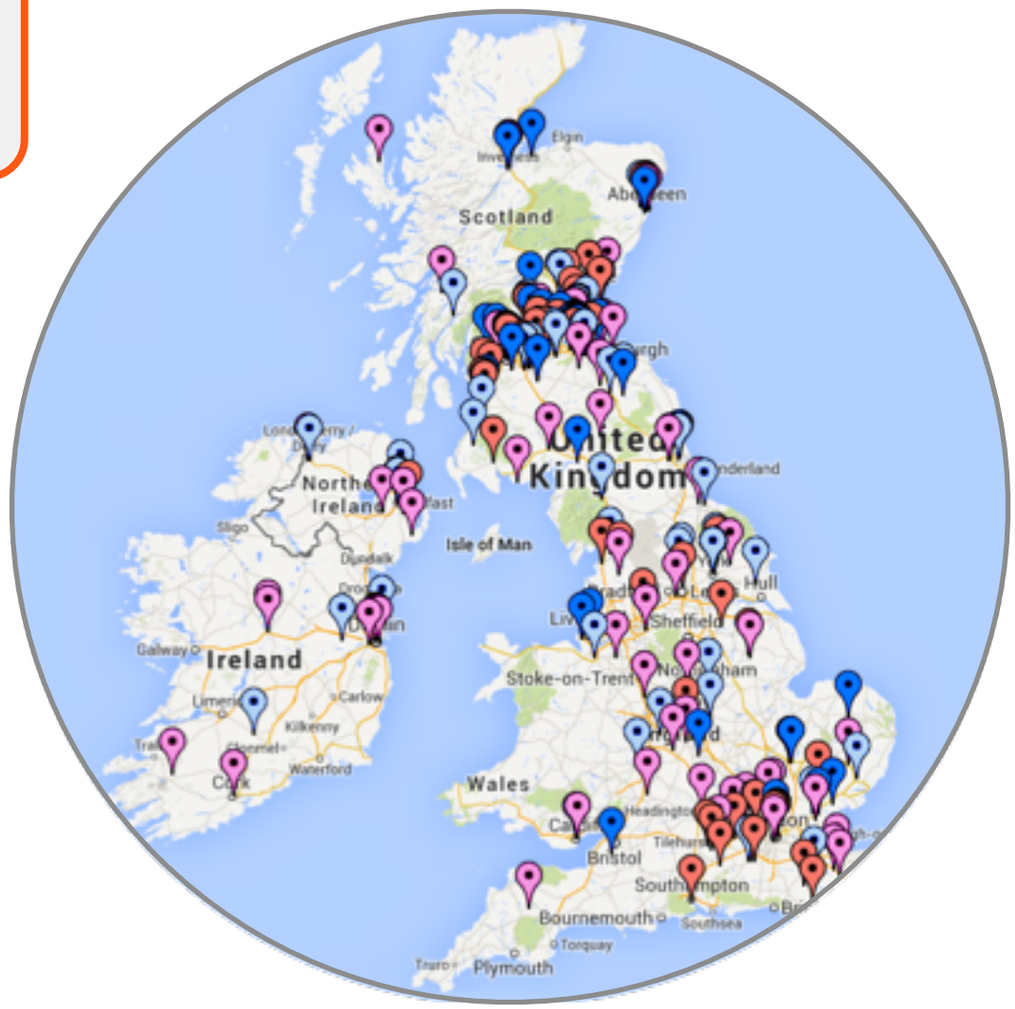
- 99% trigram coverage reached after ~3500 sentences

Voice corpus

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

More than 900 voice donors already



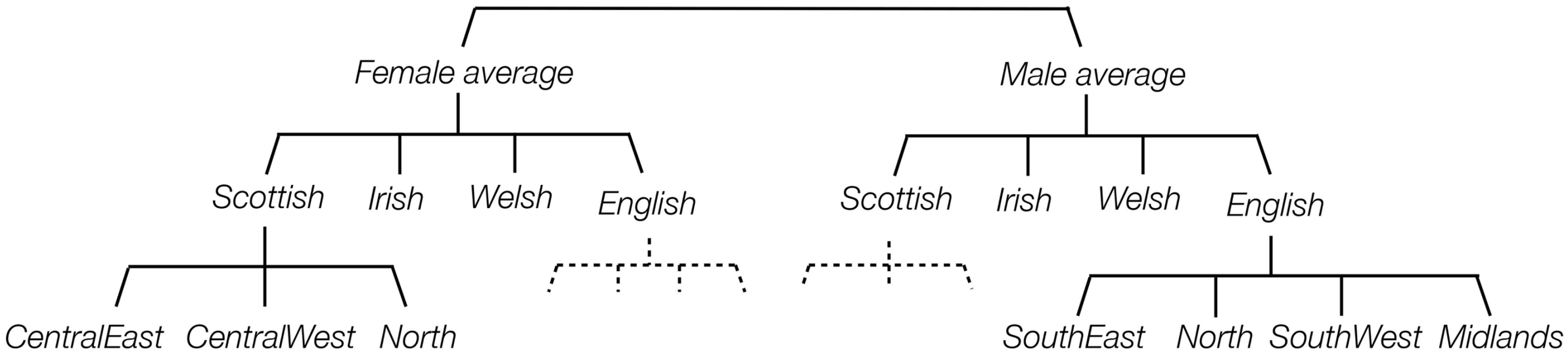
Speaker clustering

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Voice donors are pooled into clusters to create average voice models (AVM) with specific accent / gender
- Approximately 10 speakers (4000 sentences) required to build an average voice
- First approach is based on meta-data:

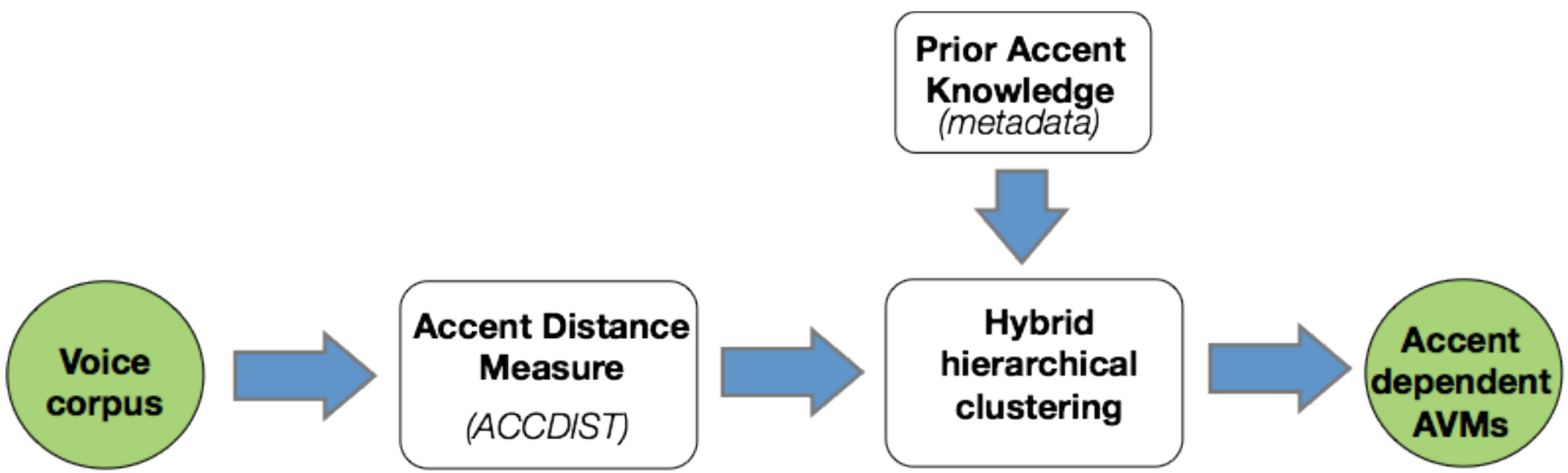
- **Hierarchy: Gender >> Country >> Broad accent >> Regional accent**



Speaker clustering

Natural Speech Technology

Edinburgh – Cambridge – Sheffield



Hierarchical clustering based on metadata

Hierarchical clustering based on acoustic distance

Speaker clustering

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

ACCDIST [Huckvale, M., 2007]

- For each speaker, acoustic distances between same vowels in different contexts

cat, father, after

➡ Vowel distance tables for each speaker

(60 mcep and dmcep coefficients at the center of the vowel)

SouthEast		
Vowel Distance	**father**	**cat**
after	2.27	3.21
father	0	3.71

- Correlation between distance tables of pairs of speakers
  - ➡ Pair-wise similarity measure of the phonological systems between speakers
  - ➡ Removes influence of speaker identity variation

Speaker clustering

Natural Speech Technology

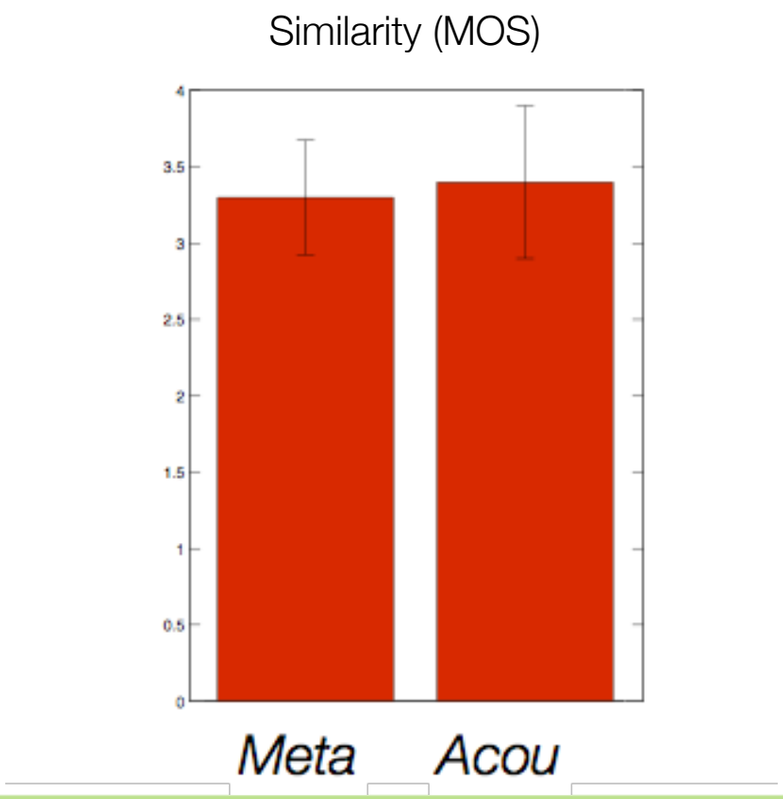
Edinburgh – Cambridge – Sheffield

Experiment

- Hierarchical clustering of Scottish female speakers based on ACCDIST
- Only clusters with more than 20 speakers are considered
- AVM are learned over each cluster of speakers ➡ 7 AVMs
- 10 target Scottish female speakers selected in different geographical regions
- For each target speaker, the best AVM is selected based on likelihood

Similarity test

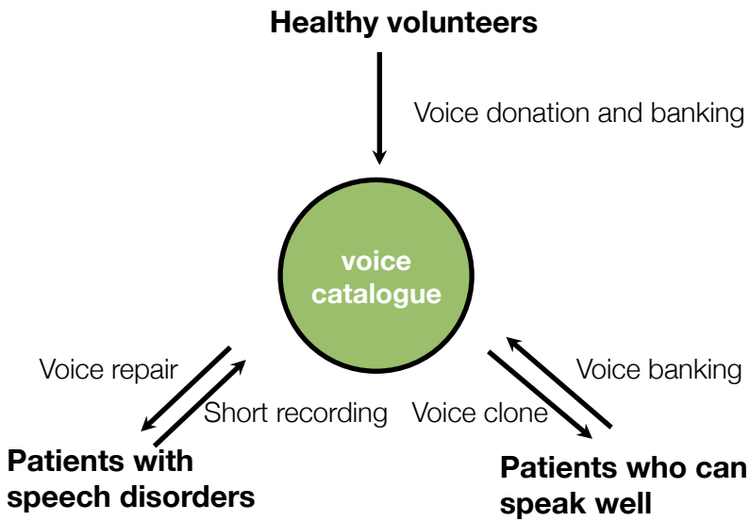
- Comparison of speaker adapted voices using the best AVM derived from meta-data versus the best AVM derived from acoustic data (hierarchical clusters)
- Reference is the target speaker voice

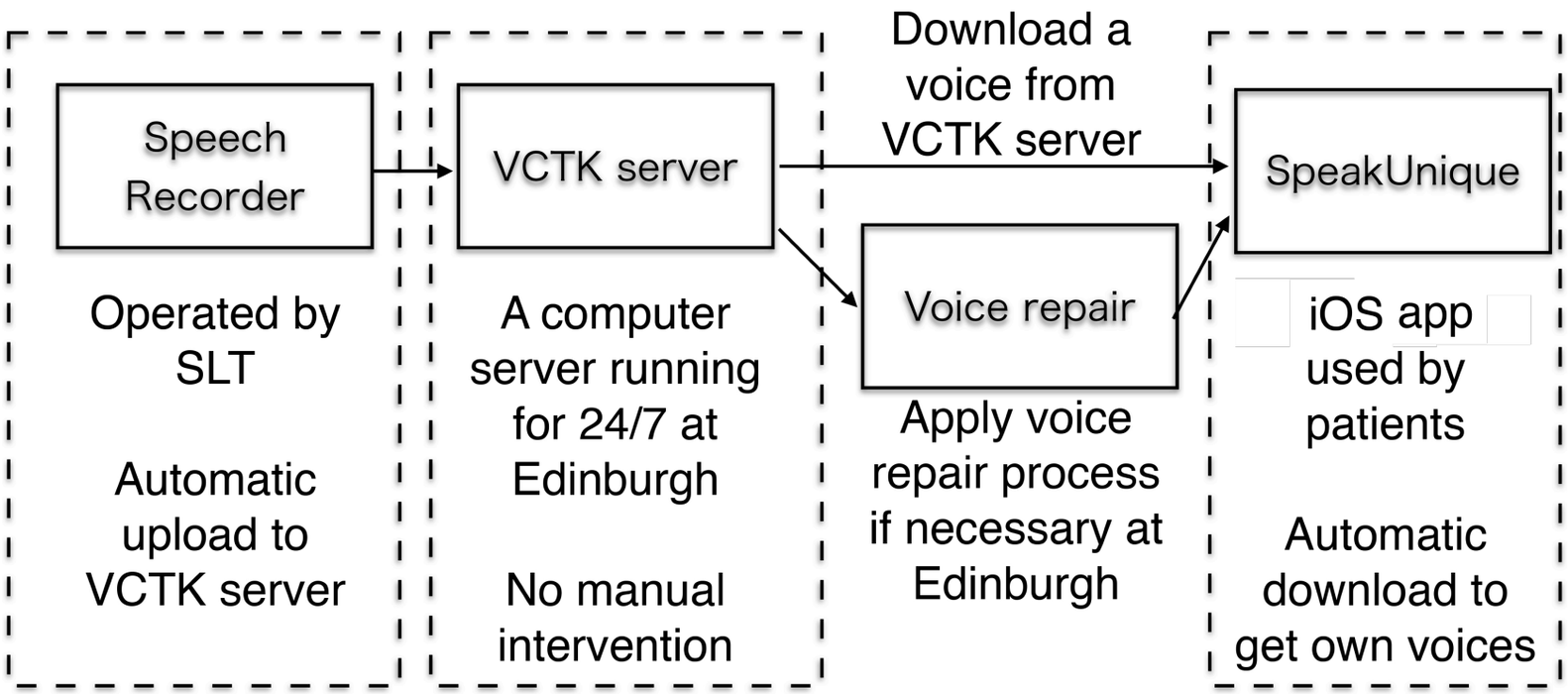


Tools for clinical trial

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

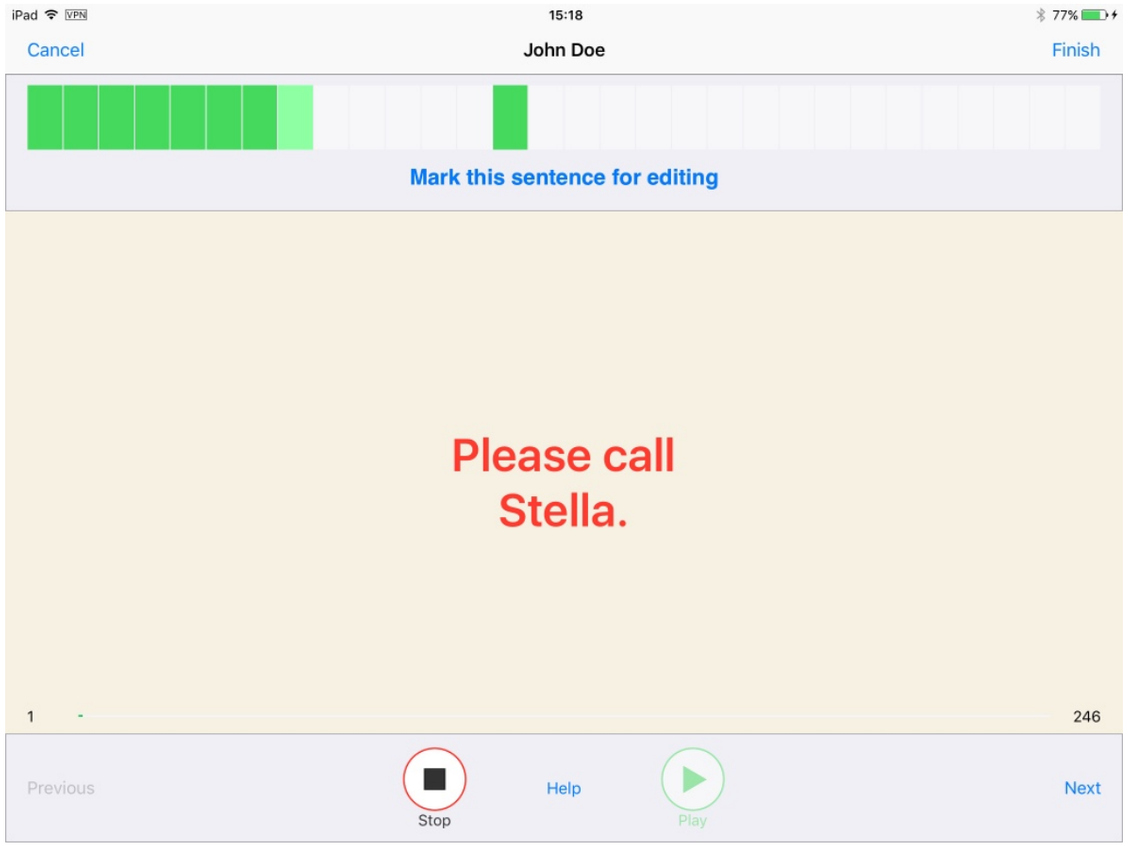




Speech Recorder

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

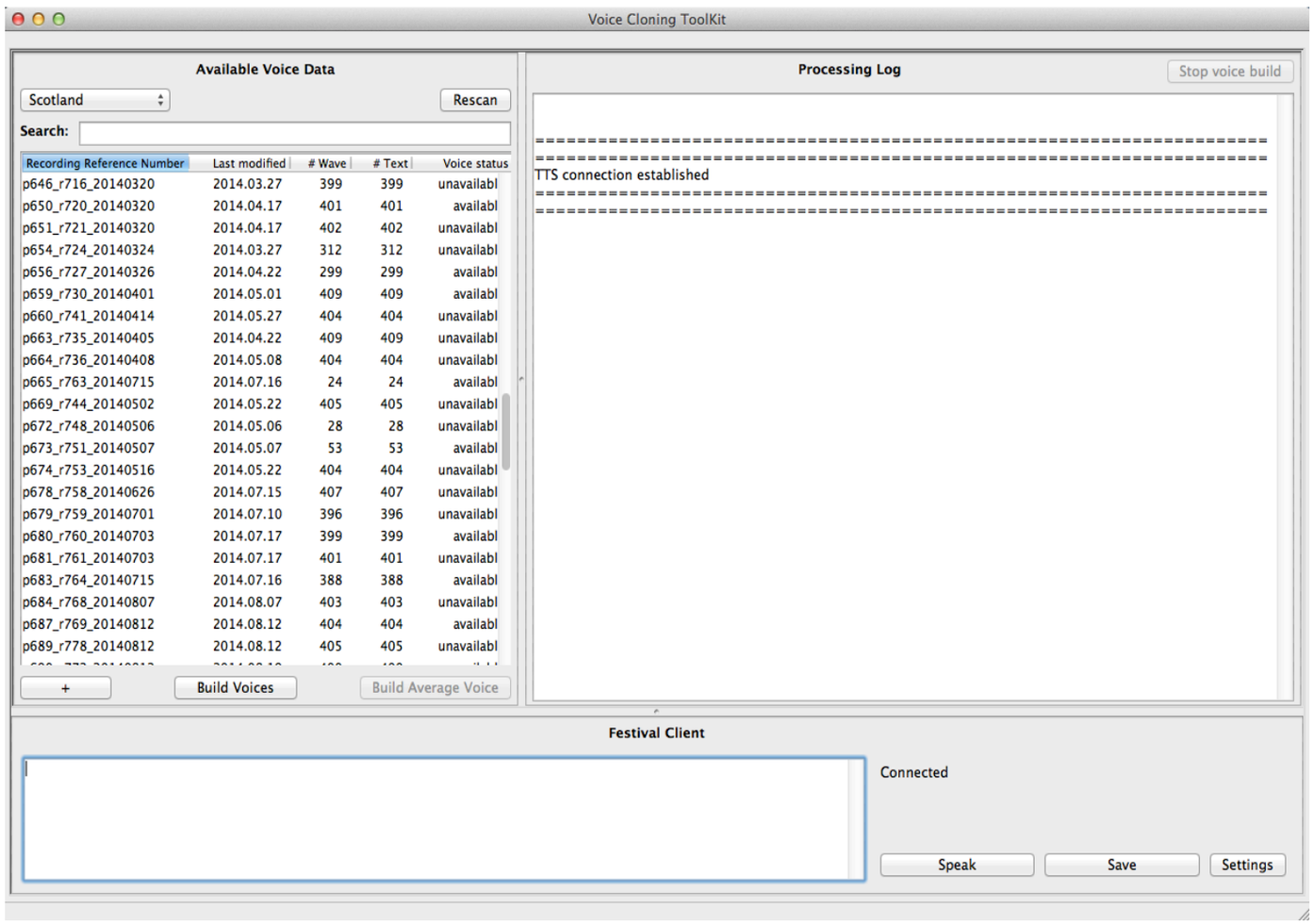


- iOs application
- Automatic monitoring of a recording
- Can be use without assistance of a SLT
- Texts optimised for triphone coverage
  but with a readability constraint (syllable bigrams and word / sentence length)
- The recordings are automatically uploaded to the server

Voice Cloning ToolKit (VCTK)

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

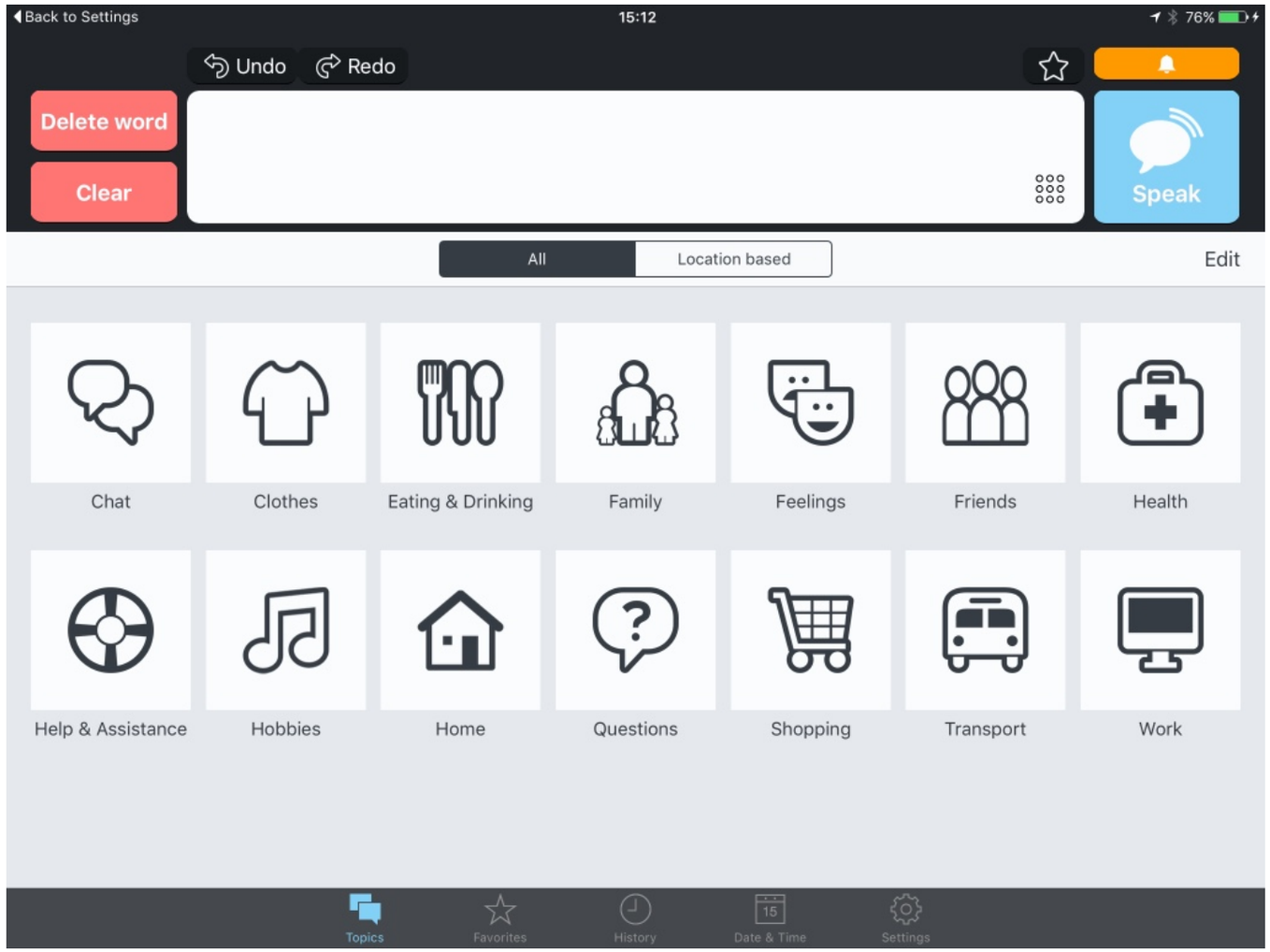


- Software designed to be used by clinicians
- automates the recording and voice building process
- Voices can built in a couple of hours
- Once built, voices can be repaired in a couple of minutes

SpeakUnique

Natural Speech Technology

Edinburgh – Cambridge – Sheffield



- iOs application
- Automatic download of the repaired voice model
- Offline synthesis
- Feedback form

Delivering Voices: Speak Unique

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Voice evaluation

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Online comparison of the personalised voice

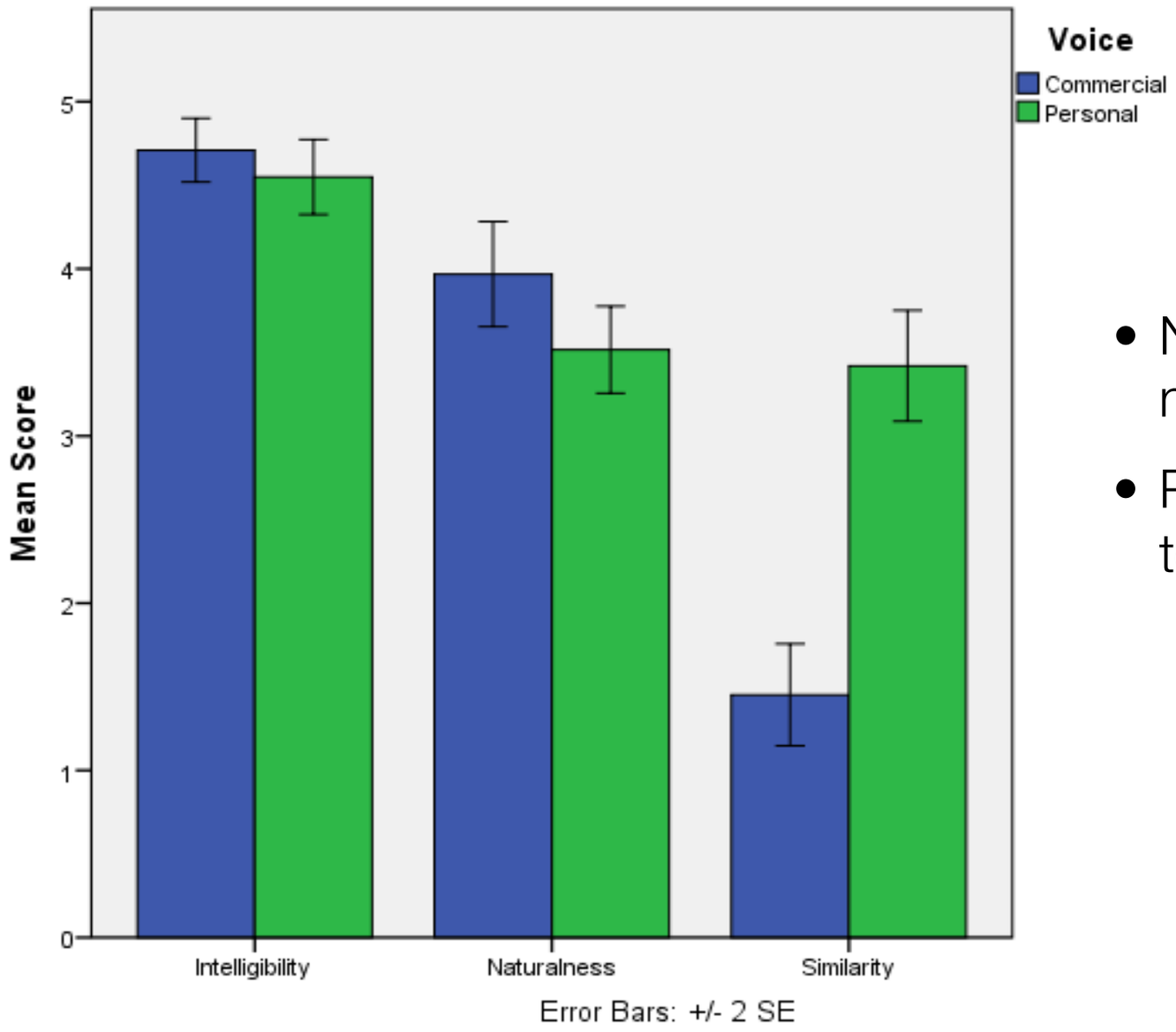
- 6 samples sentences
- Personalised voice is compared to a generic voice for VOCA (Cereproc unit selection voice built from 7h recording of voice talent)
- 40 patients completed the online evaluation
- 28 had no repair required, 12 repaired voices

- Rating of intelligibility, naturalness and similarity to own voice
- Personal overall preference

Voice evaluation

Natural Speech Technology

Edinburgh – Cambridge – Sheffield



- No significant difference in intelligibility and naturalness
- Personalised voices significantly more similar to patient's own voice

Voice evaluation

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Overall preference

- 80% of the 40 participants expressed a preference for personalised synthetic voice over the generic alternative
- However only 56% of those with ‘repaired’ voice preferred personalised
- *Comments: voice slightly robotic, not able to reproduce “strong” accent, missing naturalness of spontaneous speech*

Voice banking and delivery

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Conclusions

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Proof of concept is daily running in Anne Rowling Clinic
- Repaired voices delivered to 100 patients
- Large survey of the feedback form patients and their families
- Assessment of the improvement in terms of Quality of Life
- Spread out of the tools to company or communities / associations