Voice Banking and Voice Reconstruction

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

Cambridge, 28 May 2015

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

THE EUAN MACDONALD CENTRE

FOR MOTOR NEURONE DISEASE RESEARCH

Outline

- **Personalised VOCAs**
- **Clinical trial**: the voice banking project
- Overview of different approaches for voice reconstruction
- Speaker clustering to create age and accent specific average voice models
- **Voice reconstruction (Model interpolation)**
- Voice reconstruction (Multiple AVMs interpolation)
- Subjective experiments and results
- Perspectives

Building personalised VOCAs

Degenerative diseases (MND, Parkinson, MS)

- MND may steal the voice very rapidly (within a few months)
- Some patients may already have speech disorders at the time of diagnosis

Personalisation of VOCAs

- facilitate social interaction
- greater dignity and improved self-identity for the individual and their family

Voice banking

- Capturing the voice before it starts to degrade

HMM based speech synthesis for voice building

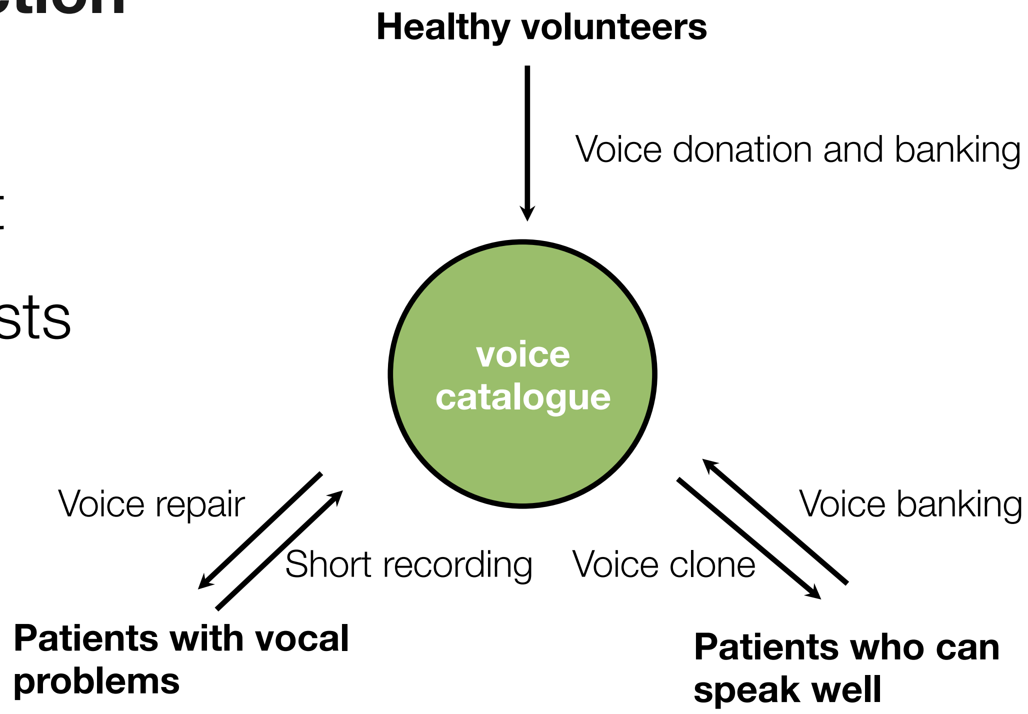
- helps to reduce complexity and to increase the flexibility of the voice building process (*adaptation of pre-trained AVMs, voice reconstruction*)

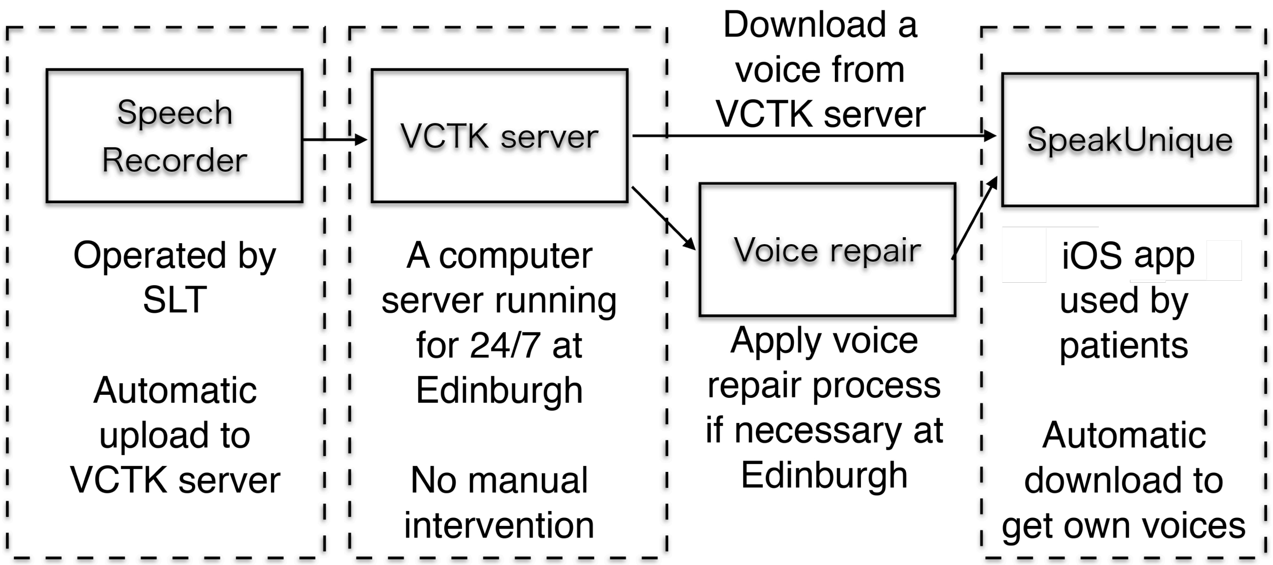
Voice banking project

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

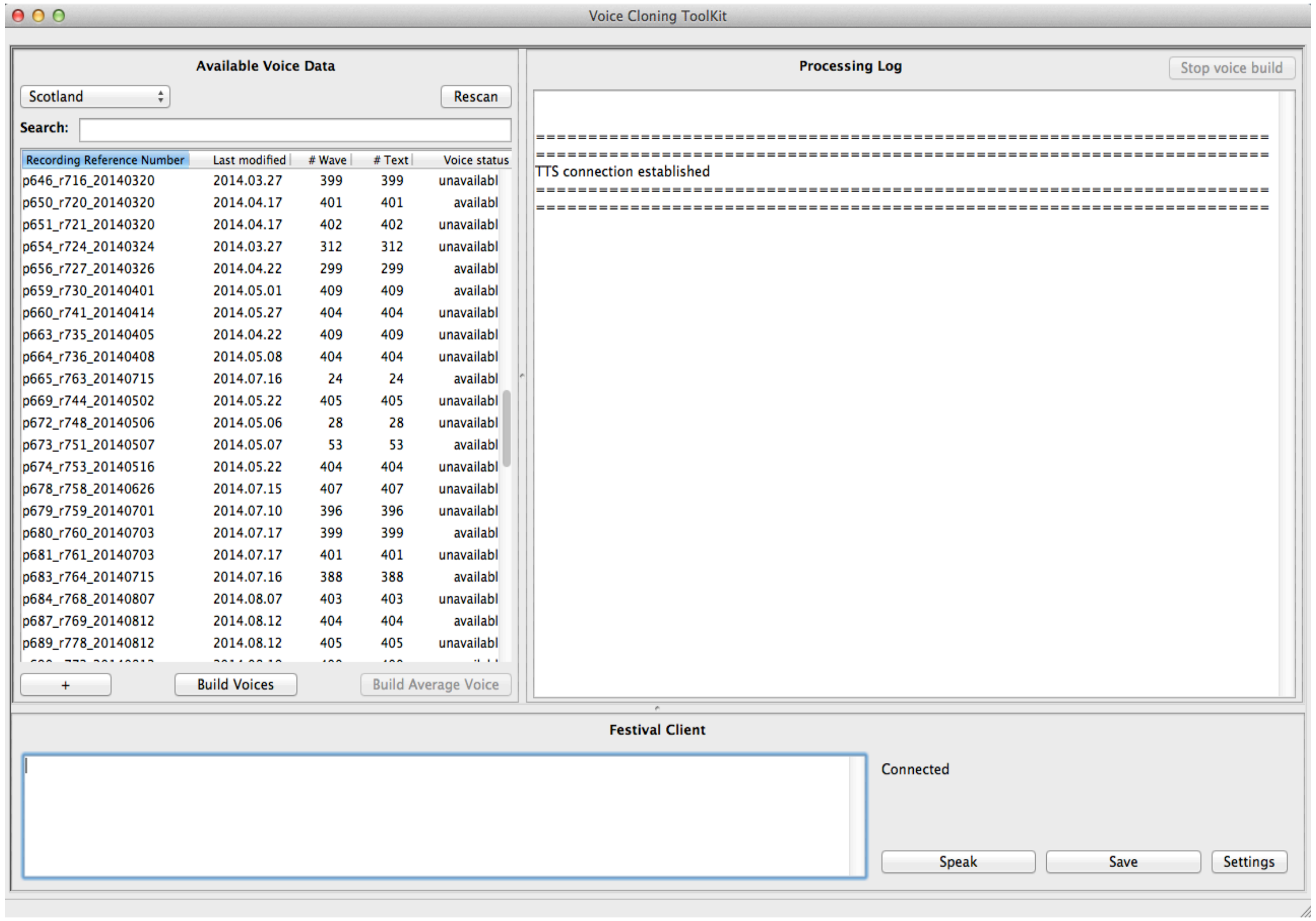
Clinical trial for voice banking and voice reconstruction

- More than 900 healthy donors voices
- 68 patients with various degrees of speech impairment
- Development of tools for speech and language therapists as well as VOCA app for patients




Speech Recorder
Operated by SLT
Automatic upload to VCTK server
VCTK server
A computer server running for 24/7 at Edinburgh
No manual intervention
Download a voice from VCTK server
Voice repair
Apply voice repair process if necessary at Edinburgh
SpeakUnique
iOS app used by patients
Automatic download to get own voices

Voice Cloning ToolKit (VCTK)



Recording Reference Number	Last modified	# Wave	# Text	Voice status
p646_r716_20140320	2014.03.27	399	399	unavailabl
p650_r720_20140320	2014.04.17	401	401	availabl
p651_r721_20140320	2014.04.17	402	402	unavailabl
p654_r724_20140324	2014.03.27	312	312	unavailabl
p656_r727_20140326	2014.04.22	299	299	availabl
p659_r730_20140401	2014.05.01	409	409	availabl
p660_r741_20140414	2014.05.27	404	404	unavailabl
p663_r735_20140405	2014.04.22	409	409	unavailabl
p664_r736_20140408	2014.05.08	404	404	unavailabl
p665_r763_20140715	2014.07.16	24	24	availabl
p669_r744_20140502	2014.05.22	405	405	unavailabl
p672_r748_20140506	2014.05.06	28	28	unavailabl
p673_r751_20140507	2014.05.07	53	53	availabl
p674_r753_20140516	2014.05.22	404	404	unavailabl
p678_r758_20140626	2014.07.15	407	407	unavailabl
p679_r759_20140701	2014.07.10	396	396	unavailabl
p680_r760_20140703	2014.07.17	399	399	availabl
p681_r761_20140703	2014.07.17	401	401	unavailabl
p683_r764_20140715	2014.07.16	388	388	availabl
p684_r768_20140807	2014.08.07	403	403	unavailabl
p687_r769_20140812	2014.08.12	404	404	availabl
p689_r778_20140812	2014.08.12	405	405	unavailabl

```
================================================================================
================================================================================
TTS connection established
================================================================================
================================================================================
```

Connected

- Software designed to be used by clinicians
- Automatises 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

Delivering Voices: Speak Unique

Speak Unique with The Voicebank Project

Voice catalogue

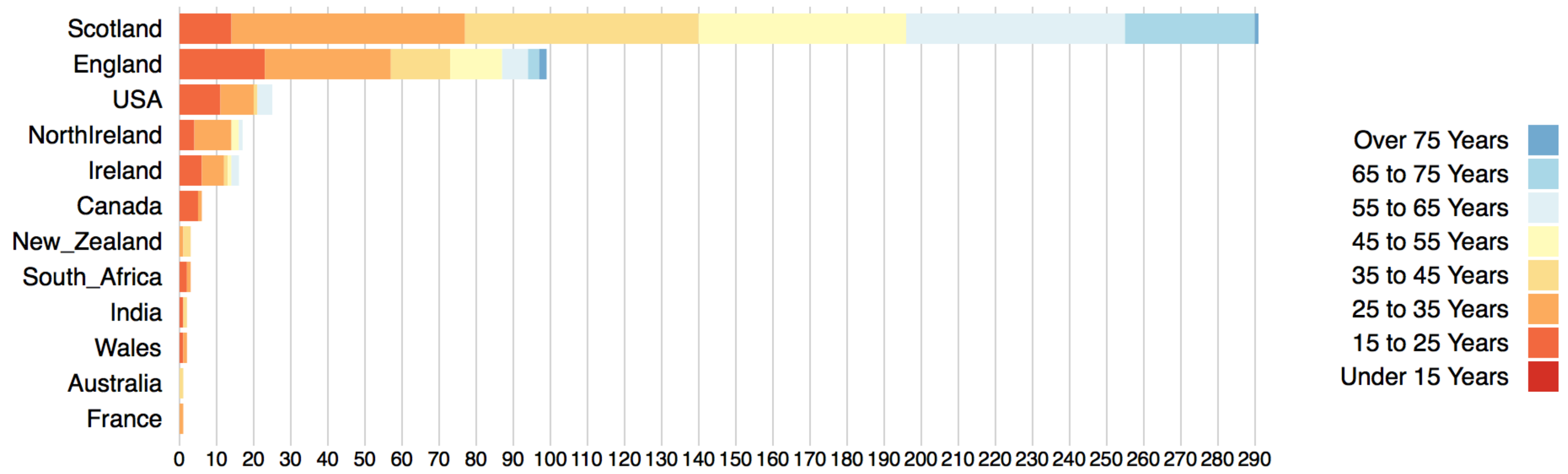
Natural
Speech
Technology

Edinburgh – Cambridge – Sheffield

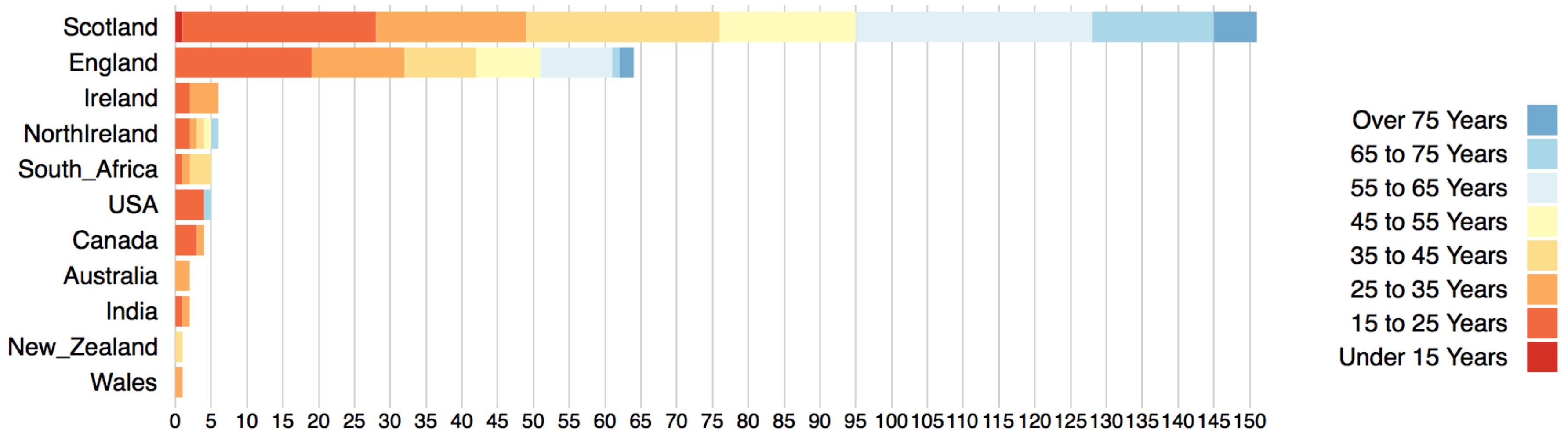
Largest speech database of British English

- 1 hour recording for healthy donors (read speech)
- 20 minutes to 1 hour recording for patients

Number of female speakers per accent and age band (total: 466)



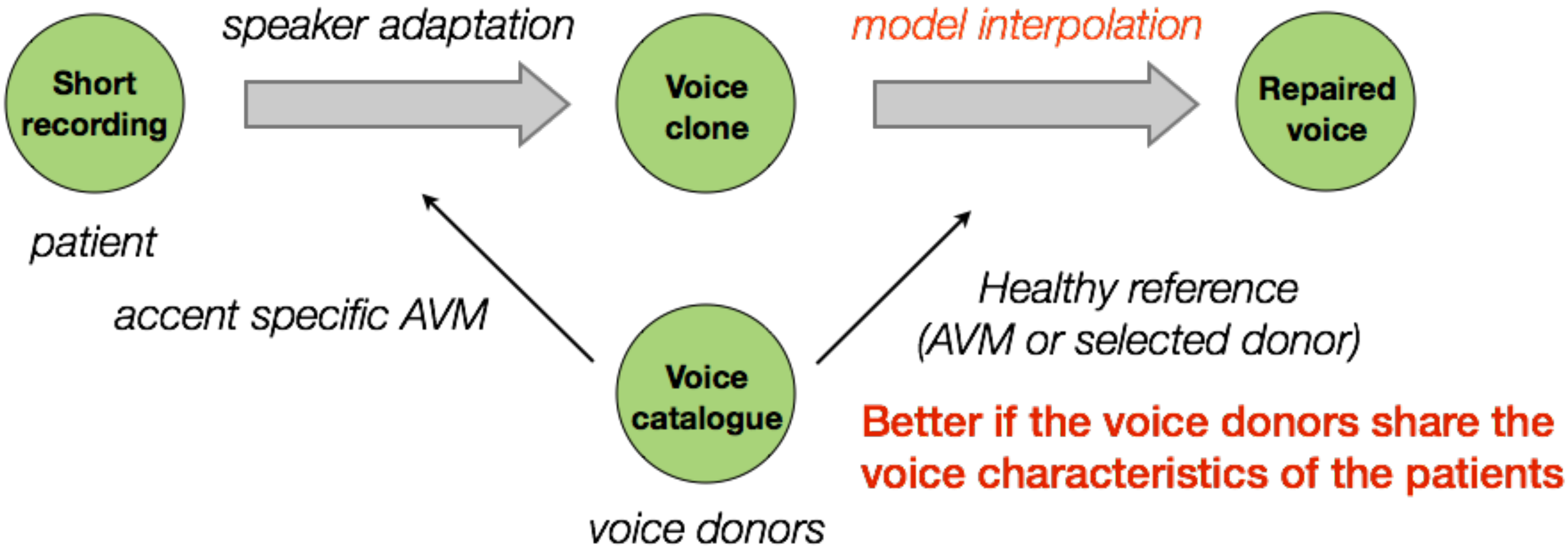
Number of male speakers per accent and age band (total: 247)



Different approaches of Voice Reconstruction

First approach: model interpolation

- Principle: fixing statistical models of the patient’s voice clone so that they can generate natural sounding speech while keeping speaker identity
- Two methods:
  - Manual tailoring of the interpolation coefficients by SLT
  - Automatic interpolation using KLD-based confidence measure



Short recording

patient

speaker adaptation

Voice clone

model interpolation

Repaired voice

accent specific AVM

Voice catalogue

voice donors

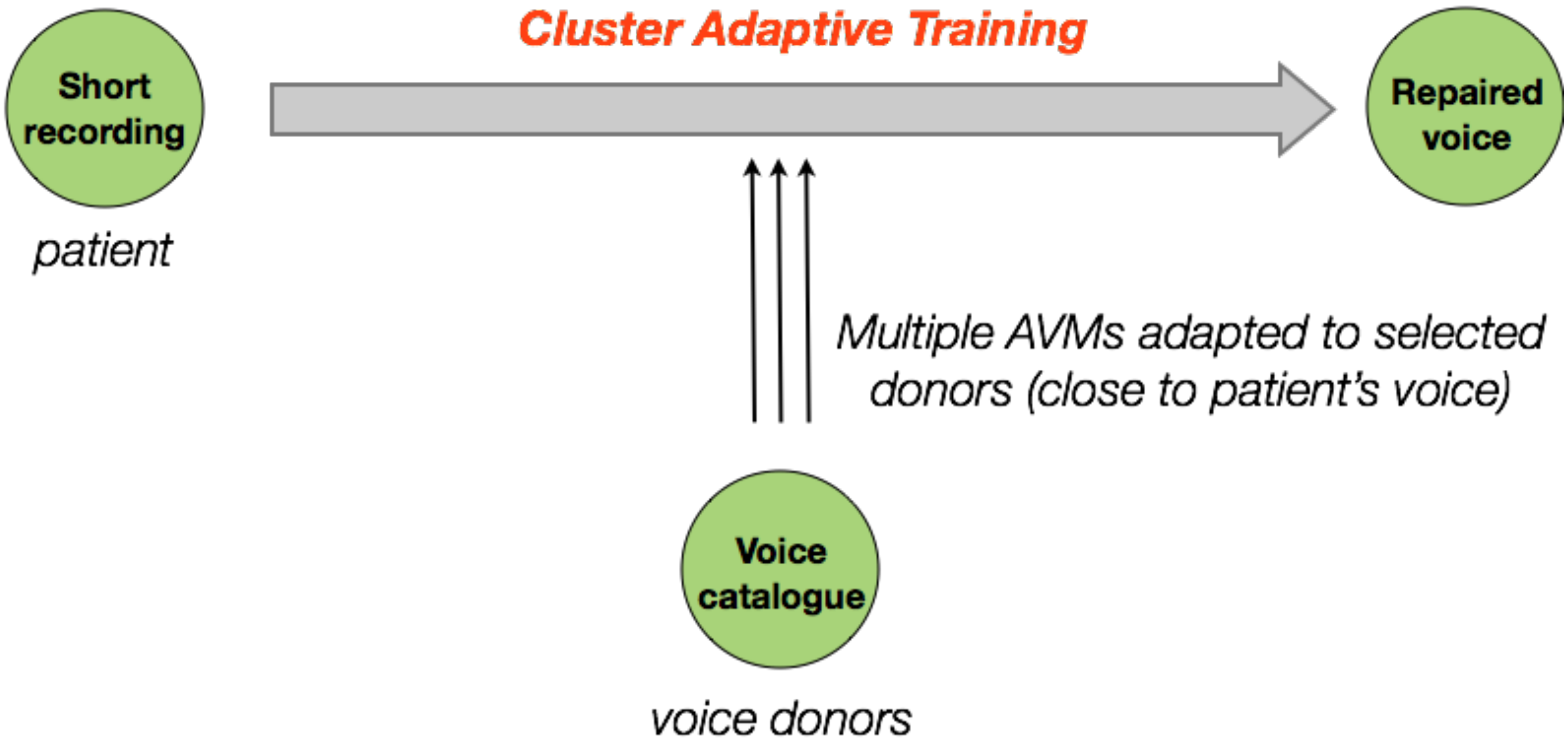
Healthy reference (AVM or selected donor)

Better if the voice donors share the voice characteristics of the patients

Different approaches of Voice Reconstruction

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

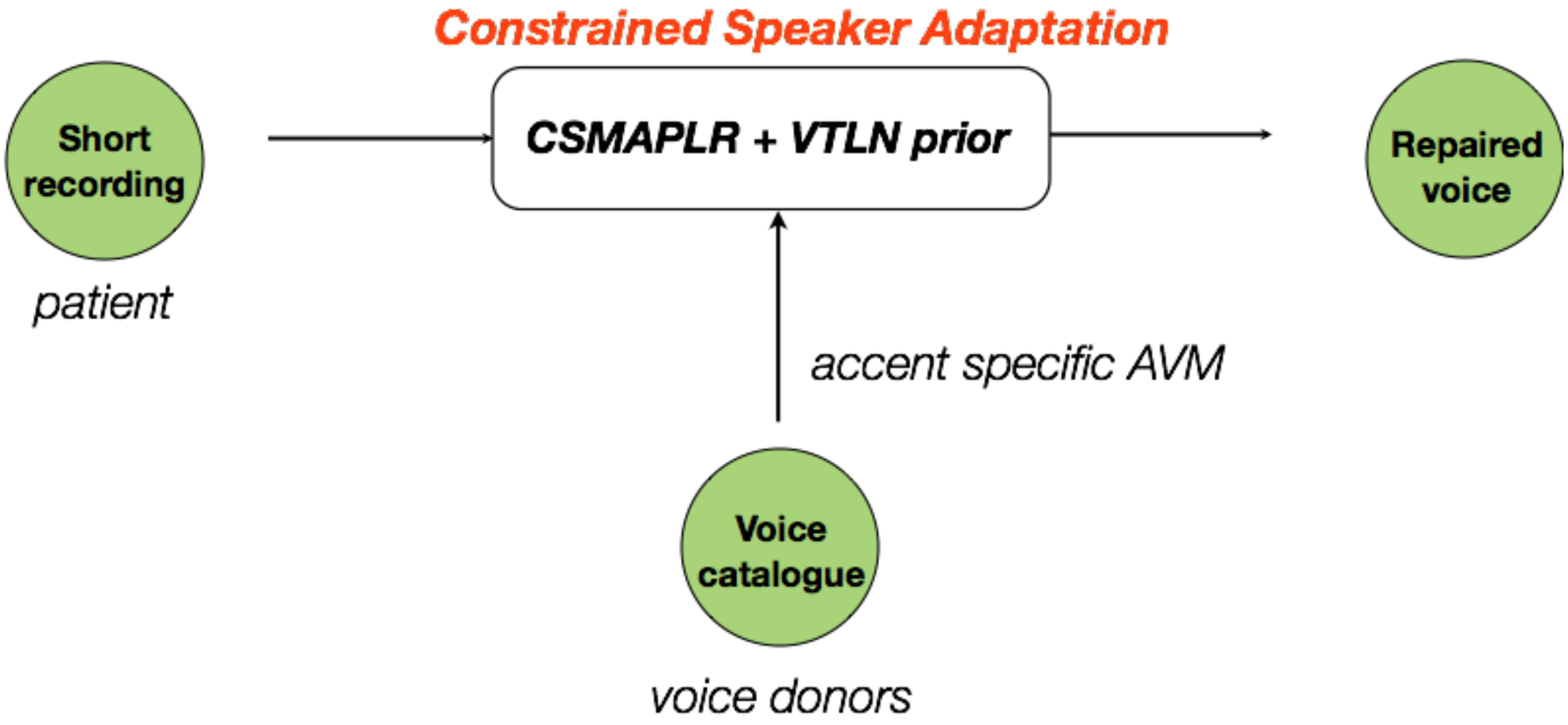
- the adapted mean vector of a component is interpolated in an eigenspace spanned by the cluster mean vectors
- but clusters are AVMs which can be tuned towards the target before interpolation


Cluster Adaptive Training
Short recording
patient
Repaired voice
Multiple AVMs adapted to selected donors (close to patient's voice)
Voice catalogue
voice donors

Different approaches of Voice Reconstruction

Third approach: constrained adaptation (*on-going*)

- estimation of the VTLN parameters (global transform) on the most reliable data (e.g. vowels)
- the VTLN transform is used as a prior to constraint the speaker adaptation
- KLD-based confidence measure can be used to adjust the weight of the VTLN prior



Constrained Speaker Adaptation

Short recording

patient

CSMAPLR + VTLN prior

Repaired voice

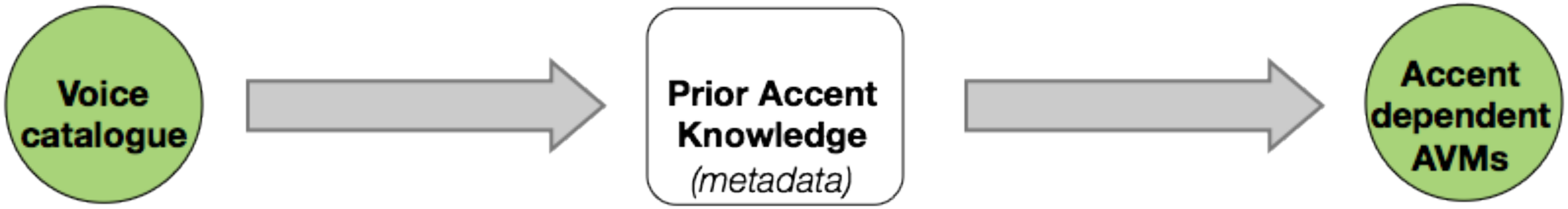
accent specific AVM

Voice catalogue

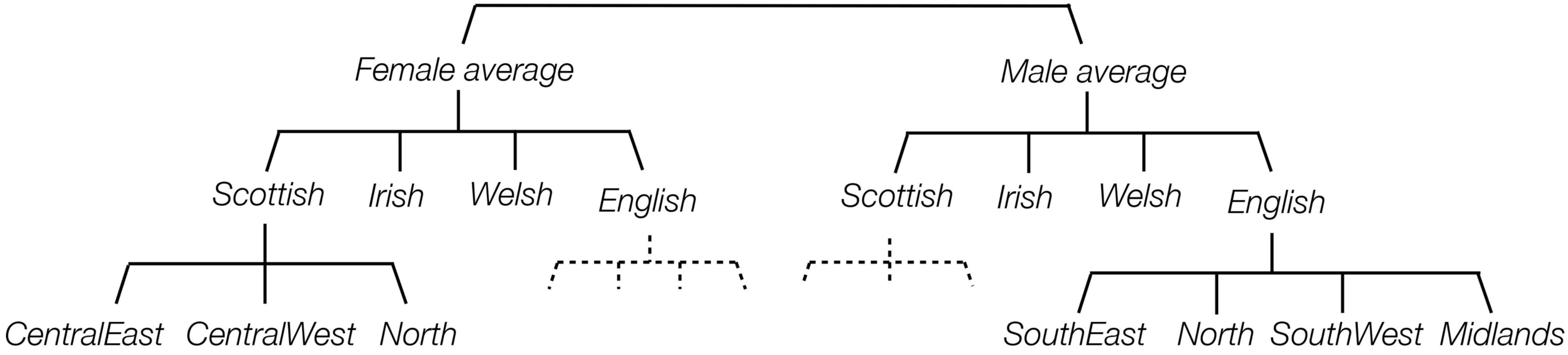
voice donors

Speaker clustering

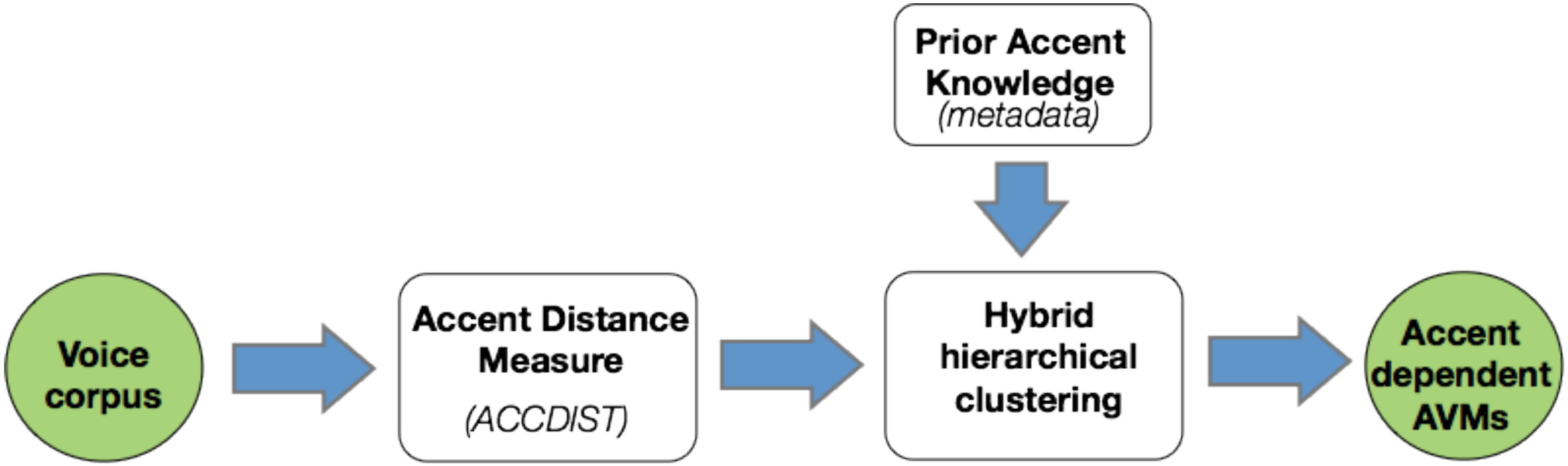
- Voice donors are pooled into clusters to create average voice models (AVM) with specific accent / gender
- Approximately 10 speakers (4000 sentences) required to build an average voice
- First approach is based on meta-data:



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



Speaker clustering



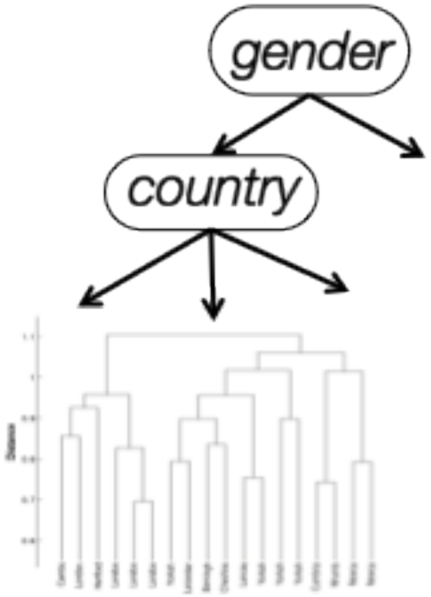
Voice corpus

Accent Distance Measure
(ACCDIST)

Prior Accent Knowledge
(metadata)

Hybrid hierarchical clustering

Accent dependent AVMs



Hierarchical clustering based on metadata

Hierarchical clustering based on acoustic distance

Speaker clustering

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

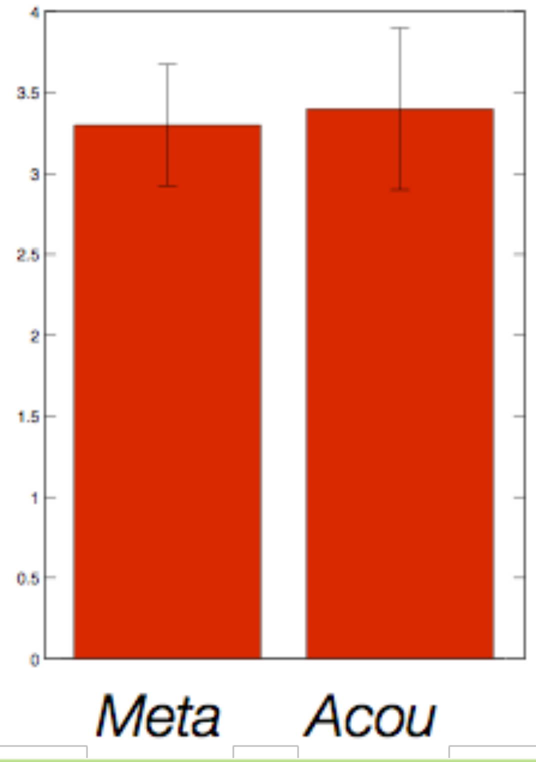
Experiment

- 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

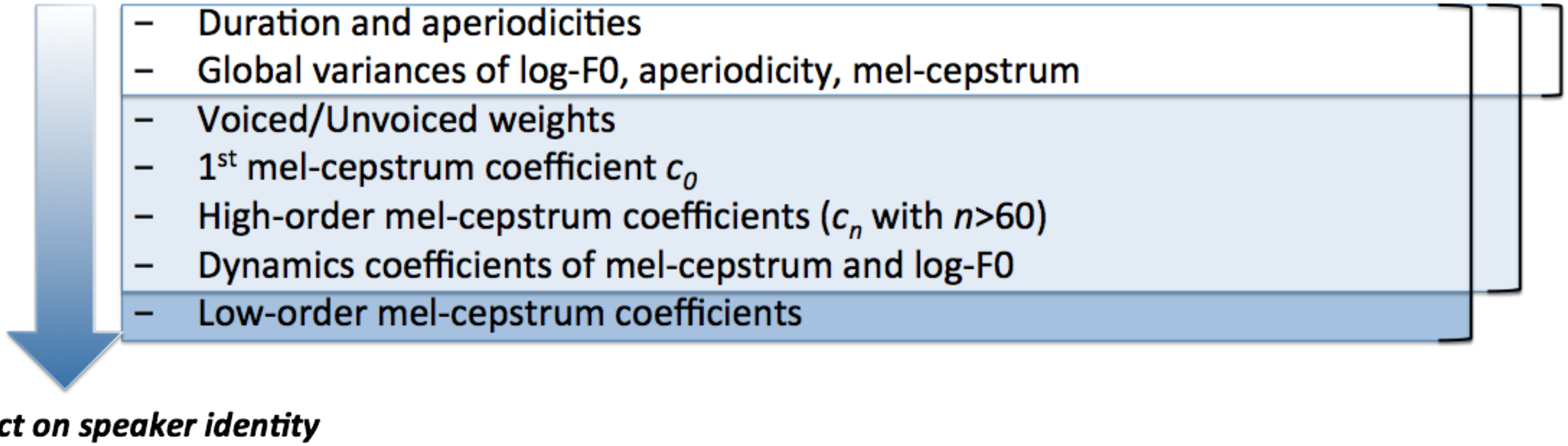
- Comparison of speaker adapted voices using the best AVM derived from meta-data versus the best AVM derived from acoustic data (hierarchical clusters)
- Reference is the target speaker voice

Similarity (MOS)



Model interpolation

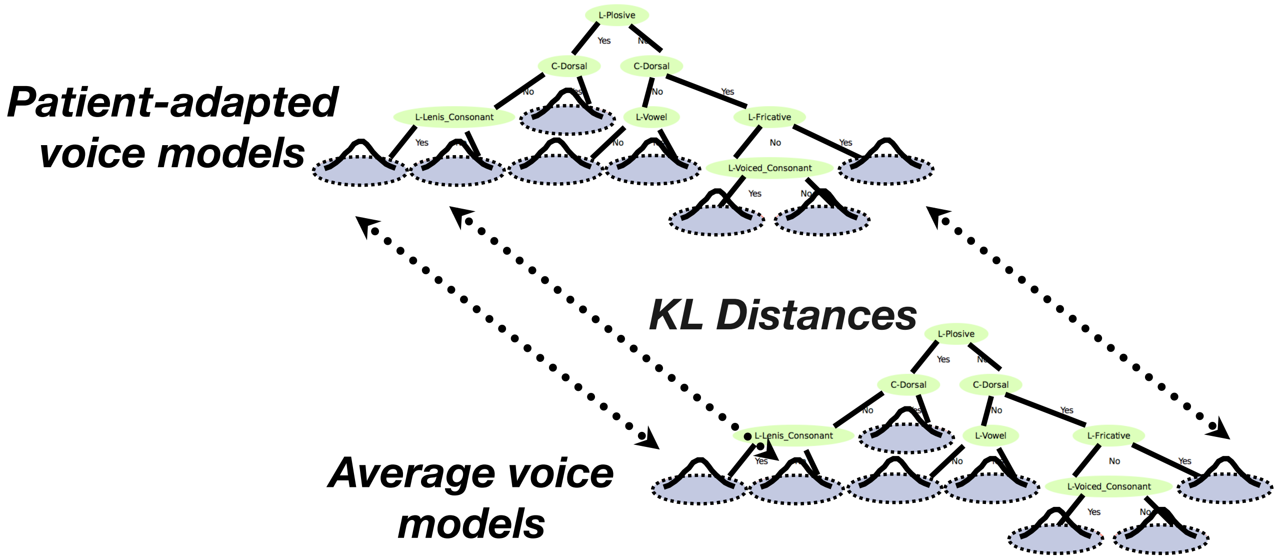
Manual: Interpolation weights are set manually by SLT



- Duration and aperiodicities
- Global variances of log-F0, aperiodicity, mel-cepstrum
- Voiced/Unvoiced weights
- 1st mel-cepstrum coefficient c_0
- High-order mel-cepstrum coefficients (c_n with n>60)
- Dynamics coefficients of mel-cepstrum and log-F0
- Low-order mel-cepstrum coefficients

Impact on speaker identity

Automatic: Interpolation weights are derived from *KLD-based confidence measure*



Model interpolation

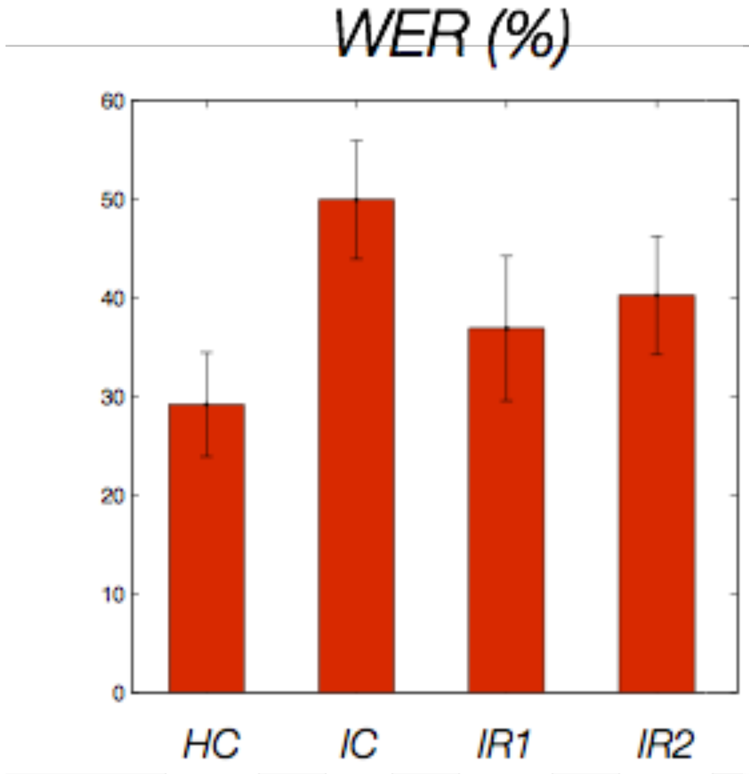
Listening tests (40 listeners)

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

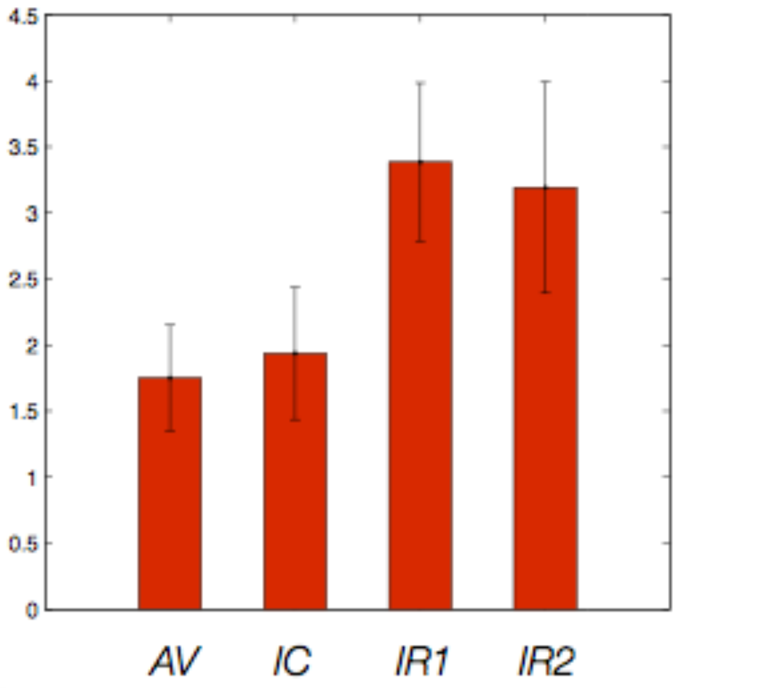
Compared synthetic voices:

- 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

WER (%)



Similarity to reference voice HC (MOS)



Model interpolation

Feedback from 15 patients and their families (manual method)

- Comments: too quick, voice slightly robotic, not able to reproduce “strong” accent, missing naturalness of spontaneous speech

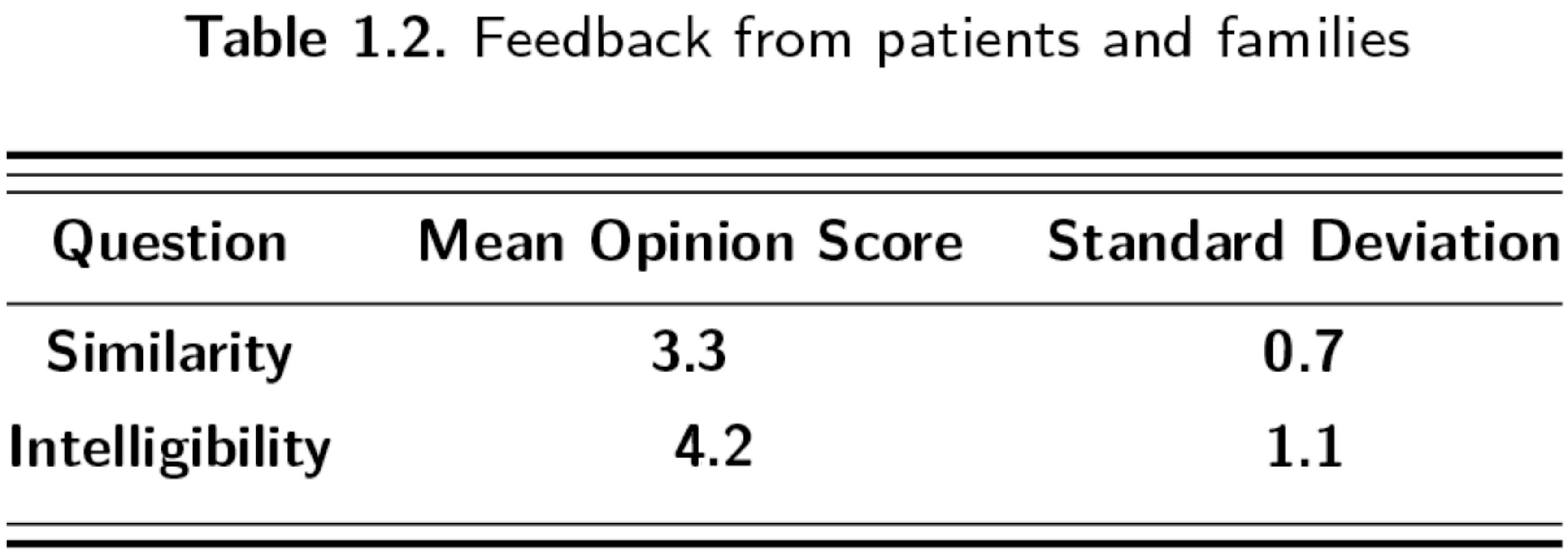


Table 1.2. Feedback from patients and families

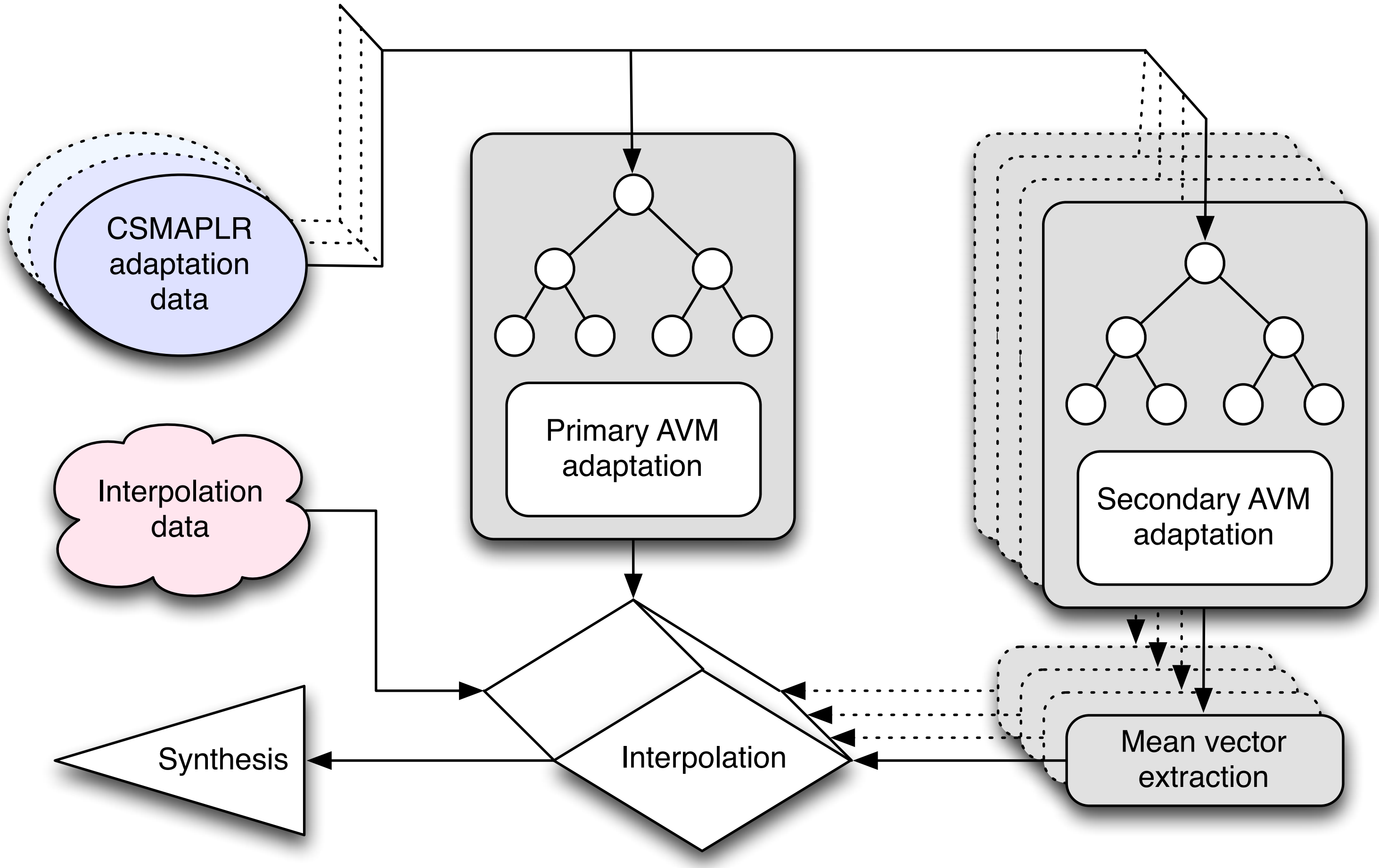
Question	Mean Opinion Score	Standard Deviation
Similarity	3.3	0.7
Intelligibility	4.2	1.1

(Naturalness Average Score of 3.1 out of 5)

➡ **On-going perceptual evaluation with 60 patients, comparing manual and automatic methods**

Multiple AVM interpolation

- 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

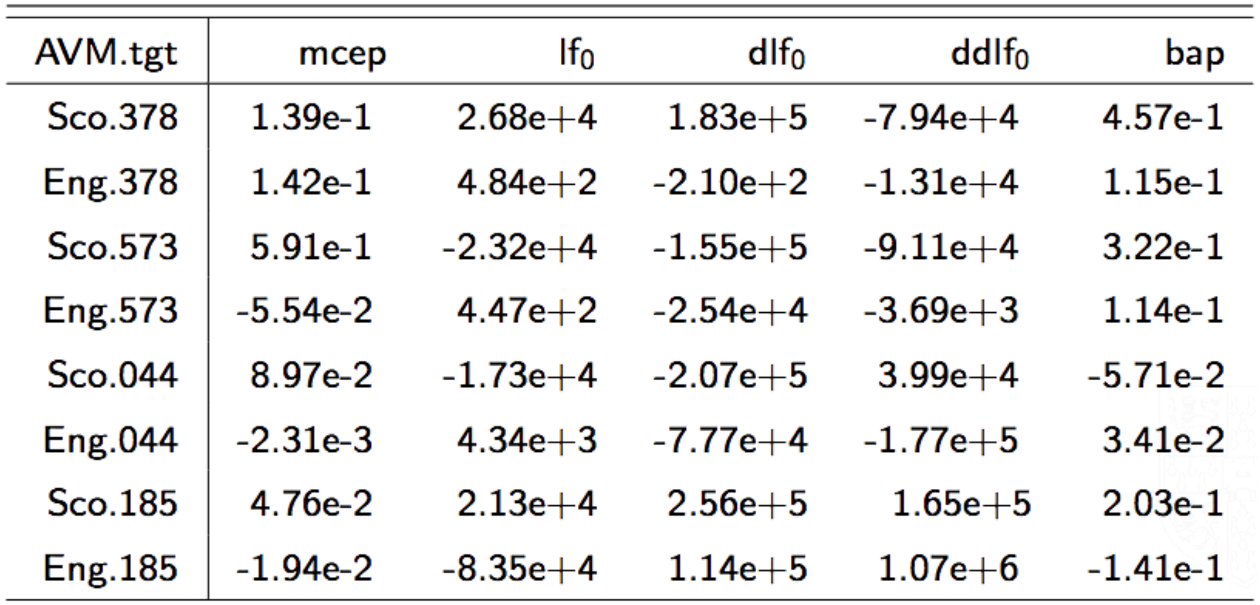
Experiment:

- Reconstruction of a patient voice with mild dysarthia: Female, Scottish (Glasgow)
- 2 British accent AVMs: English (106 speakers), Scottish (181 speakers)
- Pre-selection of 21 female voices with glasgow accent aged 23 to 68 years
- Adaptation of the scottish AVM towards each of these 21 voices
- Selection of the 4 closest voice donors according to likelihood given the patient data

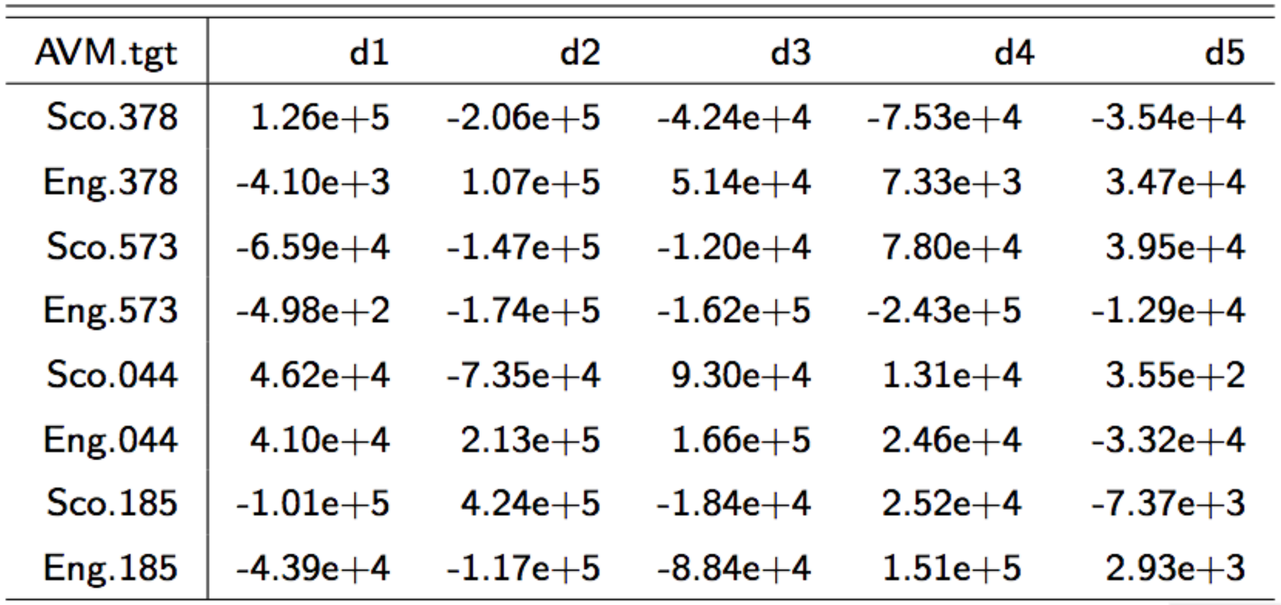
- The 2 AVMs were adapted to each of the 4 selected speakers leading to 8 adapted AVMs

Multiple AVM interpolation

Interpolation weights for each speaker and each stream



AVM.tgt	mcep	lf$_0$	dlf$_0$	ddlf$_0$	bap
Sco.378	1.39e-1	2.68e+4	1.83e+5	-7.94e+4	4.57e-1
Eng.378	1.42e-1	4.84e+2	-2.10e+2	-1.31e+4	1.15e-1
Sco.573	5.91e-1	-2.32e+4	-1.55e+5	-9.11e+4	3.22e-1
Eng.573	-5.54e-2	4.47e+2	-2.54e+4	-3.69e+3	1.14e-1
Sco.044	8.97e-2	-1.73e+4	-2.07e+5	3.99e+4	-5.71e-2
Eng.044	-2.31e-3	4.34e+3	-7.77e+4	-1.77e+5	3.41e-2
Sco.185	4.76e-2	2.13e+4	2.56e+5	1.65e+5	2.03e-1
Eng.185	-1.94e-2	-8.35e+4	1.14e+5	1.07e+6	-1.41e-1



AVM.tgt	d1	d2	d3	d4	d5
Sco.378	1.26e+5	-2.06e+5	-4.24e+4	-7.53e+4	-3.54e+4
Eng.378	-4.10e+3	1.07e+5	5.14e+4	7.33e+3	3.47e+4
Sco.573	-6.59e+4	-1.47e+5	-1.20e+4	7.80e+4	3.95e+4
Eng.573	-4.98e+2	-1.74e+5	-1.62e+5	-2.43e+5	-1.29e+4
Sco.044	4.62e+4	-7.35e+4	9.30e+4	1.31e+4	3.55e+2
Eng.044	4.10e+4	2.13e+5	1.66e+5	2.46e+4	-3.32e+4
Sco.185	-1.01e+5	4.24e+5	-1.84e+4	2.52e+4	-7.37e+3
Eng.185	-4.39e+4	-1.17e+5	-8.84e+4	1.51e+5	2.93e+3

- the range of weights assigned to duration and f0 streams reveals the atypical characteristics of these patient’s voice components;

- some voice symptoms have been reproduced during the interpolation despite having only a small degree of freedom

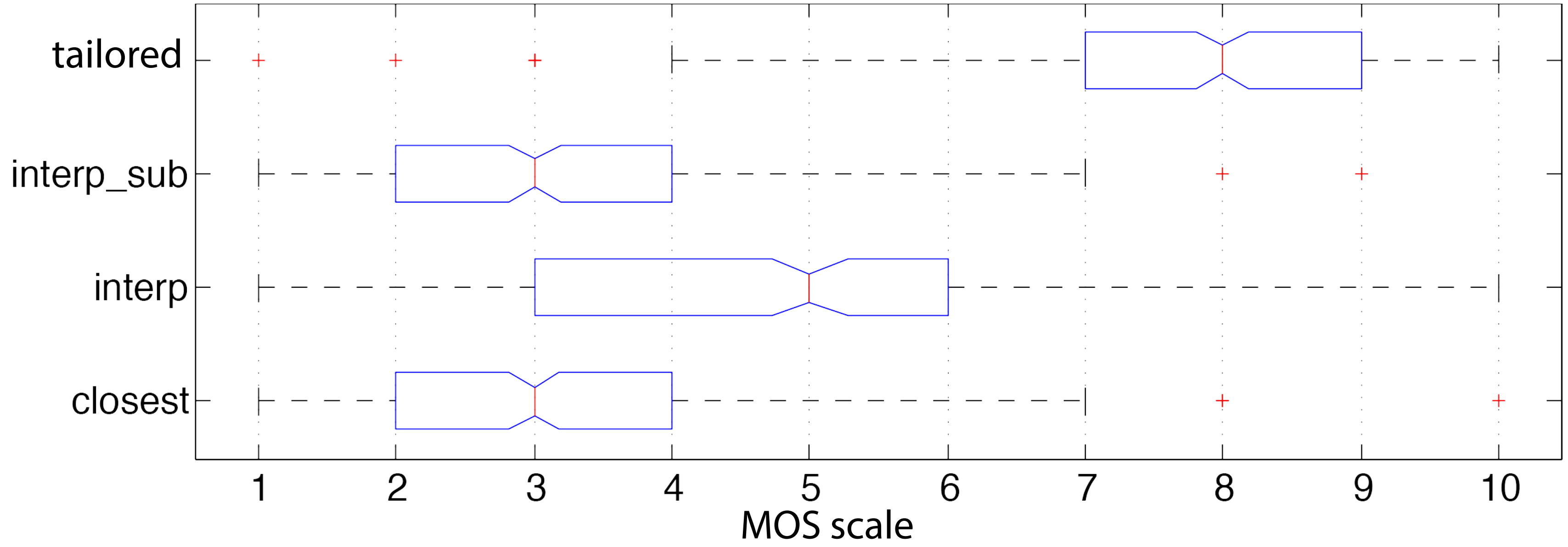
Multiple AVM interpolation

Listening tests (38 listeners):

- closest: Scottish AVM adapted towards the closest voice donor
- interp: Multiple AVM interpolation
- interp sub: interp + substitution of f0, dlf0, ddlf0, dur from closest donors
- tailored: manually reconstructed by speech therapist

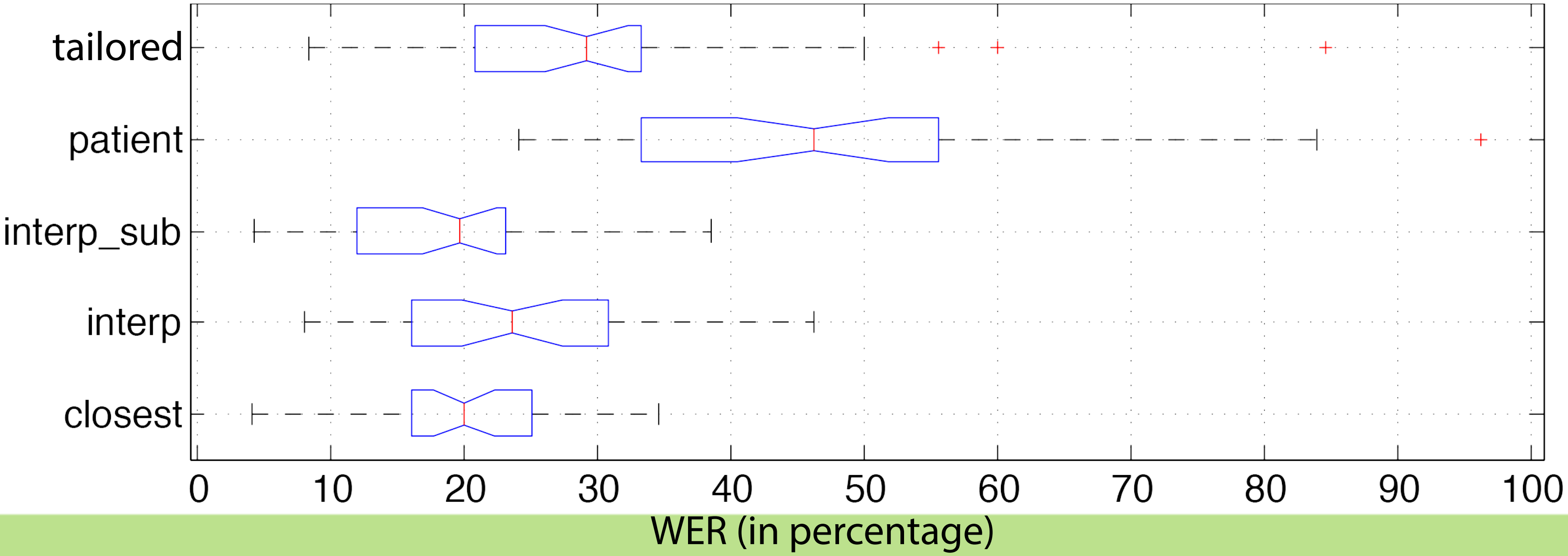
Similarity Test

reference: AVM directly adapted towards the patient voice



Intelligibility Test

semantically unpredictable sentences



Perspectives

- Proof of concept is daily running in Anne Rowling Clinic
- Repaired voices delivered to 19 patients
- Assessment of the improvement in terms of Quality of Life
- Improving the voice repair process
- Spread out of the tools to company or communities / associations