# Reconstructing Voices within the Multiple-AVM framework

P. Lanchantin[†], C. Veaux[*] , M.J.F. Gales[†], S. King[*], J. Yamagishi[*]

[†]Cambridge University Engineering Department ,Cambridge, UK

[*]Centre for Speech Technology Research, Edinburgh, UK

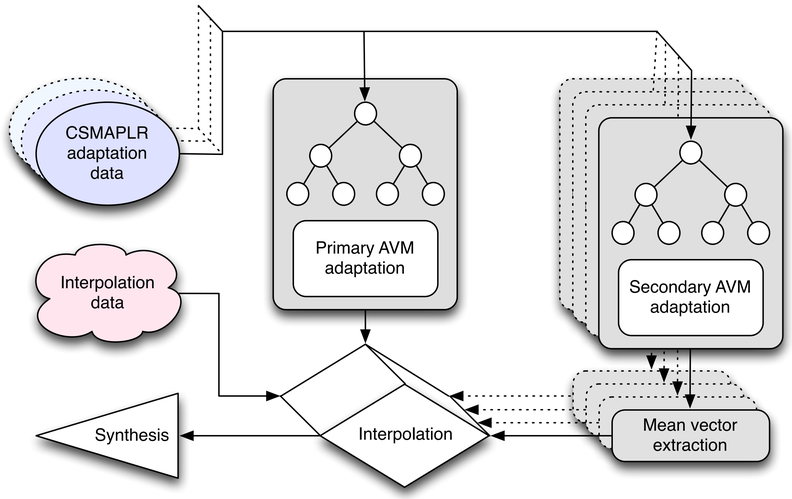
Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- **Personalisation of VOCAs**
  - facilitate social interaction
  - provision of personalised voice is associated with greater dignity and improved self-identity for the individual and their family
- **Voice Banking**
  - process of capturing the voice before it starts to degrade
  - requires large amount of recorded intelligible speech
  - problematic for patients who's voice has already started to deteriorate
  - strong motivation to reduce complexity and to increase the flexibility of the voice building process

## The Multiple-AVM framework

- hybrid approach between the AVM and the CAT framework
- **Cluster Adaptive Training (CAT [Gales00])**
  - the adapted mean vector of a component is interpolated in an eigenspace spanned by the cluster mean vectors
  - vs MAVM: clusters are AVMs which can be tuned towards the target before interpolation
- **Average Voice-based speech synthesis([Yamagishi12])**
  - AVM can be trained independently
  - vs MAVM: decision tree intersection allows a wide variety of possible contexts to be produced




- **Principle of Multiple-AVM**
  - each AVM is trained separately on data selected according to a specific factor (age, gender, regional accent, ...)
  - during adaptation **each AVM is adapted towards a speaker close to the target (or to the target itself) before interpolation** → design of the space in which the interpolation takes place depending on the application.

## Voice reconstruction in the MAVM framework

- Advantages of MAVM framework regarding the Voice Reconstruction task
- **flexibility**
  - 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
  - interpolation weights can be fine-tuned by a practician
- **complexity**
  - only a small amount of data is required to estimate the weights interpolation vector

## Experiments

- Reconstruction of a patient voice with mild dysarthia (Female, with Glasgow accent)
- 2 British accent AVMs: English (106 speakers), Scottish (181 speakers)
- Selection of 4 closest 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 (p378, p573, p044, p185) according to likelihood given the patient data
- the 2 AVMs were adapted to each of the 4 selected speakers leading to 8 adapted AVMs

**Listening Test** (38 listeners)

- comparison of 4 reconstructions of the patient's voice in terms of similarity, intelligibility and naturalness
  - ***closest*** voice: Scottish AVM adapted towards the closest p378
  - ***interp*** voice: the proposed approach
  - ***interp_sub*** voice: substitution of f0, dlf0, ddlf0, dur by closest ones
  - ***tailored*** voice: manually reconstructed by speech therapist

## How does the Multiple-AVM approach perform in the voice reconstruction task ?

### Estimated interpolation weights

- the range of weights assigned to duration and $f_0$ streams reveals the atypical characteristics of these patient's voice components;
- characteristics have been reproduced during the interpolation despite having only a small degree of freedom

Table 1: Estimated interpolation weights for each model stream.

| AVM.tgt | mcep | $lf_0$ | $dlf_0$ | $ddlf_0$ | bap | d1 | d2 | d3 | d4 | d5 |
|---|---|---|---|---|---|---|---|---|---|---|
| Sco.378 | 1.39e-1 | 2.68e+4 | 1.83e+5 | -7.94e+4 | 4.57e-1 | 1.26e+5 | -2.06e+5 | -4.24e+4 | -7.53e+4 | -3.54e+4 |
| Eng.378 | 1.42e-1 | 4.84e+2 | -2.10e+2 | -1.31e+4 | 1.15e-1 | -4.10e+3 | 1.07e+5 | 5.14e+4 | 7.33e+3 | 3.47e+4 |
| Sco.573 | 5.91e-1 | -2.32e+4 | -1.55e+5 | -9.11e+4 | 3.22e-1 | -6.59e+4 | -1.47e+5 | -1.20e+4 | 7.80e+4 | 3.95e+4 |
| Eng.573 | -5.54e-2 | 4.47e+2 | -2.54e+4 | -3.69e+3 | 1.14e-1 | -4.98e+2 | -1.74e+5 | -1.62e+5 | -2.43e+5 | -1.29e+4 |
| Sco.044 | 8.97e-2 | -1.73e+4 | -2.07e+5 | 3.99e+4 | -5.71e-2 | 4.62e+4 | -7.35e+4 | 9.30e+4 | 1.31e+4 | 3.55e+2 |
| Eng.044 | -2.31e-3 | 4.34e+3 | -7.77e+4 | -1.77e+5 | 3.41e-2 | 4.10e+4 | 2.13e+5 | 1.66e+5 | 2.46e+4 | -3.32e+4 |
| Sco.185 | 4.76e-2 | 2.13e+4 | 2.56e+5 | 1.65e+5 | 2.03e-1 | -1.01e+5 | 4.24e+5 | -1.84e+4 | 2.52e+4 | -7.37e+3 |
| Eng.185 | -1.94e-2 | -8.35e+4 | 1.14e+5 | 1.07e+6 | -1.41e-1 | -4.39e+4 | -1.17e+5 | -8.84e+4 | 1.51e+5 | 2.93e+3 |

### Listening tests

**Similarity**

- healthy version of the voice was unavailable
- reference: AVM directly adapted towards the patient voice
- 10-points scale MOS, 30 randomly chosen pairs

**Intelligibility**

- transcription test
- 20 utterances played just once

**Naturalness**

- AB comparison test
- asked to judge which sample sounds more natural

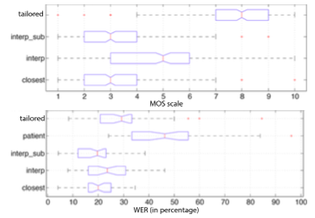



Figure 2: Results of the similarity test (top) and of the intelligibility test (bottom).

Table 2: Naturalness evaluation, (95% error margin=5.73).

| A | pref A | pref B | B |
|---|---|---|---|
| tailored | 39.38 | **60.62** | interp_sub |
| tailored | **83.56** | 16.44 | interp |
| tailored | 35.62 | **64.38** | closest |
| interp_sub | **94.86** | 5.14 | interp |
| interp | 14.04 | **85.96** | closest |
| interp_sub | 53.42 | **46.58** | closest |

- **Conclusion**
  - the Multiple-AVM framework is well-suited to the reconstruction task
    - requires small amount of patient's data
    - can be fine tuned by a speech practician
    - interpolation in clean voice eigenspace
  - evaluations show improvement in naturalness and intelligibility compared to a voice reconstructed by a practician but further evaluation is required for similarity.

This Research was supported by ESPRC Programme Grant, grant no. EP/I031022/1 (Natural Speech Technology)