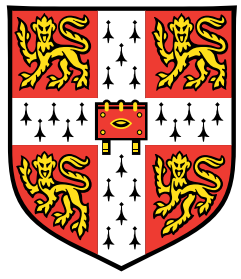
Multiple-Average-Voice-based Speech Synthesis

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

P.Lanchantin, M.Gales, S.King,
H.Lu, C.Veaux, J.Yamagishi



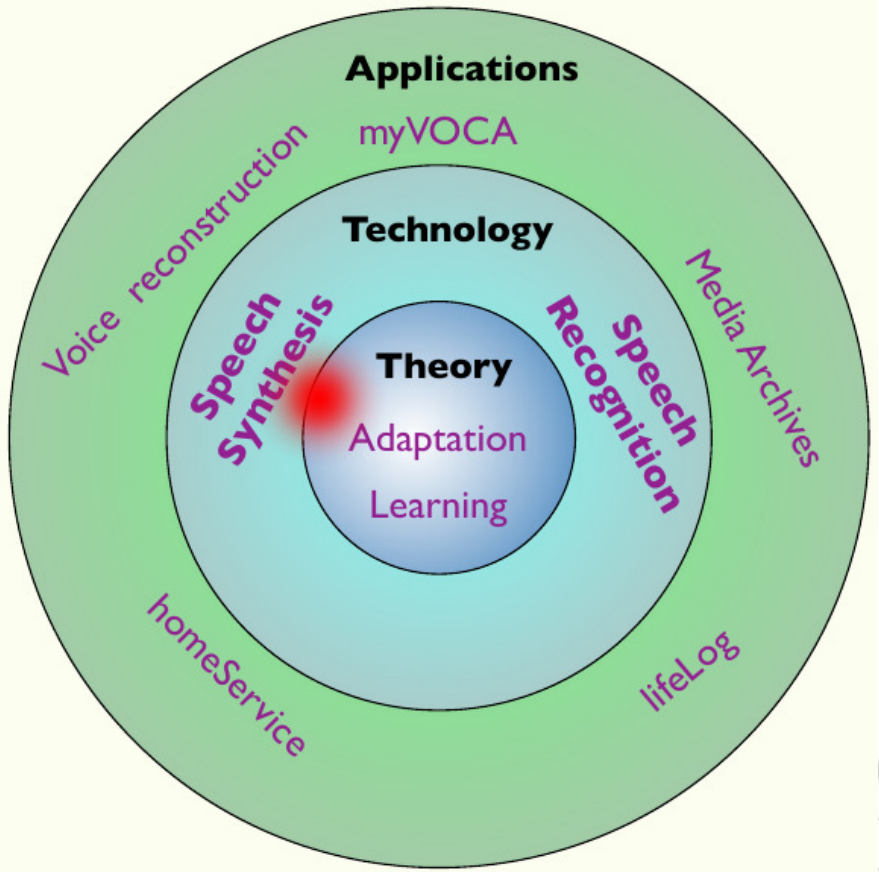
23 May 2013

Positioning in the NST project

- **Sub-project: Learning and Adaptation**
  - propose factorised canonical models, adaptation and adaptive training approaches for ASR and TTS

- **Factorised approaches**
  - separation of source of variations in speech
  - allow separate adaptation, reduce amount of training data, rapid adaptability

- **Focus of this work**
  - improve **quality** by closely matching the target speaker characteristics
  - **control** of the characteristics of the synthesized voice over the range of factors



Average-voice based Speech Synthesis

Natural Speech Technology

- **Average-Voice Model** (AVM)
  - canonical model of voices
  - seed for adaptation
- **Characteristics of the adapted voice**
  - depend directly on the AVM (topology, training methods/data)
  - quality/naturalness depend on the AVM-target voice distance[1]
  - better to use a smaller number of carefully chosen speakers than a large number of arbitrary speakers[2]
- **Way for improvement**
  - use several AVMs instead of a broad AVM
  - use factor-dependent AVMs

[1] J. Yamagishi et al. "Thousands of Voices for HMM-Based Speech Synthesis-Analysis and Application of TTS System Built on Various ASR Corpora". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.5 (2010).

[2] R. Dall et al. "Analysis of Speaker Clustering Strategies for HMM-Based Speech Synthesis". In: *Proc. Interspeech*. 2012.

Cluster Adaptive Training

Natural Speech Technology

- **Principle**
  - *estimation*: clusters of speaker dependent component parameters which define an eigenspace
  - *adaptation*: adapted voice is given by a vector of interpolation weights [1, 2]
- **Pros**
  - fast adaptation
  - 1 decision tree per cluster
  - no fragmentation of data during estimation
- **Cons**
  - computationally expensive when the amount of training data/number of clusters gets large

[1] M.J.F. Gales. "Cluster Adaptive Training of Hidden Markov Models". In: *IEEE Transactions on Audio, Speech, and Language Processing* 8 (2000), pp. 417–428.

[2] H. Zen et al. "Statistical Parametric Speech Synthesis Based on Speaker and Language Factorization". In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.6 (2012).

Multiple-Average-voice based Speech Synthesis

Natural Speech Technology

- **Principle**
  - *estimation*: CAT Clusters are AVM trained on different portions of training data
  - *adaptation*: each AVM are first adapted to the target voice before interpolation
- **Pros**
  - estimation is computationally less expensive than CAT
  - more convenient when adding a new AVM (update of the system)
- **Cons**
  - fragmentation of the training data

UNIVERSITY OF CAMBRIDGE

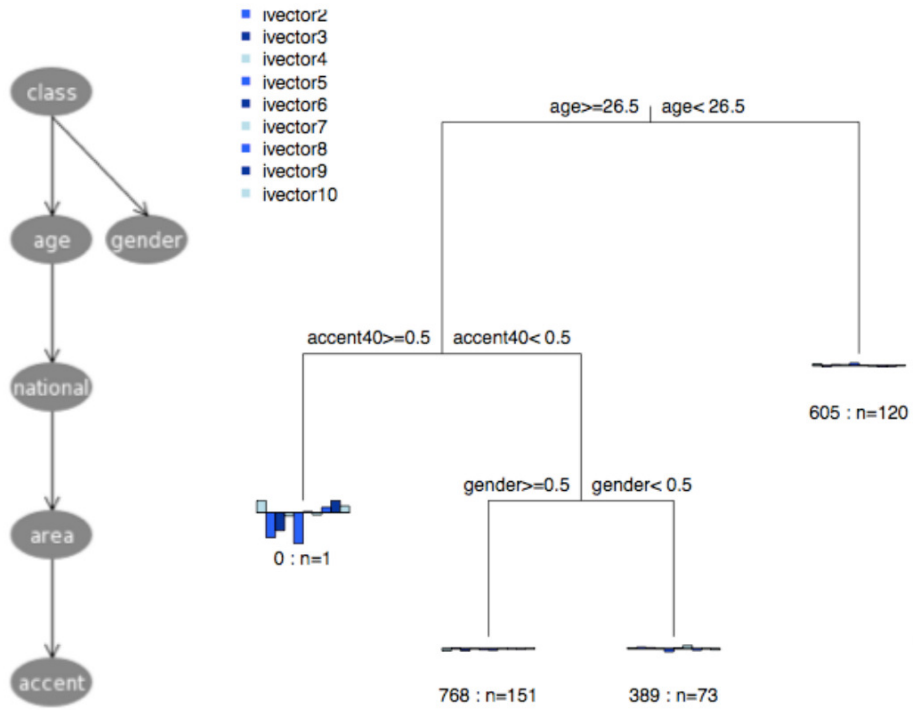
Adaptation scheme

- **Selection of the closest AVM**
  - will define the component covariances used during the interpolation
- **Adaptation of the AVMs**
  - Adaptation of each AVM separately (CMLLR+MLLR)
- **Interpolation**
  - Maximum-Likelihood based estimation
  - For each component m, only means are interpolated, covariance is the one of the closest AVM
    - $\boldsymbol{\mu}^{(m)} = \mathbf{M}^{(m)} \boldsymbol{\lambda}^{(r_m)}$
    - $\mathbf{M}^{(m)}$: the matrix of P AVM mean vectors
    - $\boldsymbol{\lambda}^{(r_m)}$: AVM weight vector for AVM weight class r_m

Factor Selection

- **Which Factors should we consider?** (will define the eigenspace)
  - most distinguishing factors
- use dynamic Bayesian network (dBN) and CART to structure diverse speakers

- VoiceBank including metadata (age, gender, accent, ...)
- use i-vector representation
- **age and gender** were find to be the most important factors
- should normalize age and gender first in order to find other important factors



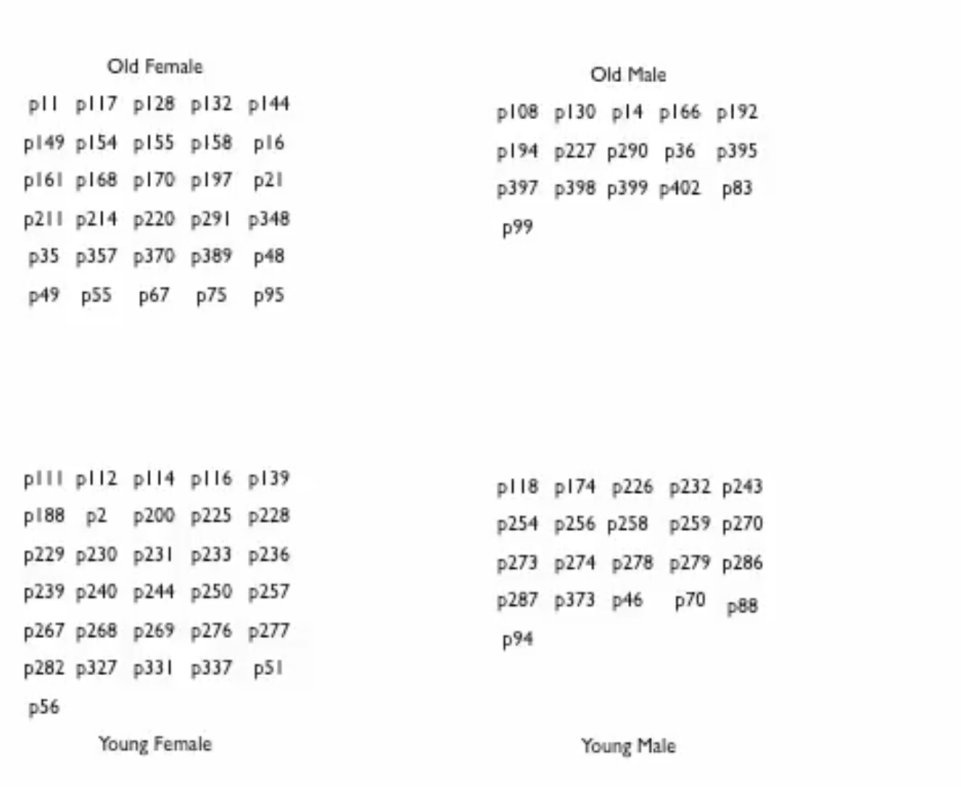
dBN (left) and Decision tree (right) trained using i-vectors

Data Selection

Natural Speech Technology

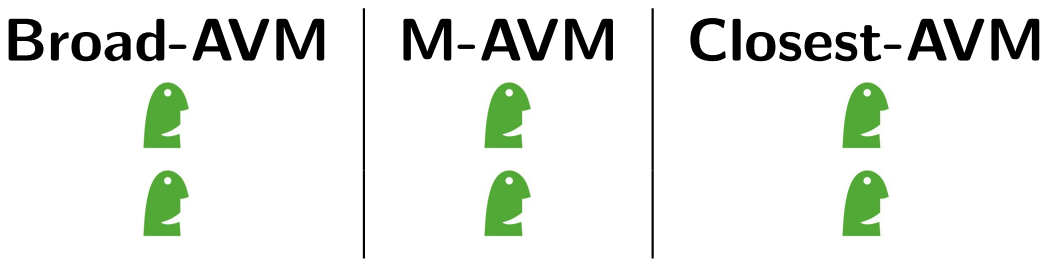
- **How to select the training data ?**
  - Metadata can be used but might not be reliable
  - $\rightarrow$ speaker reassignement

1. *Initialisation*: initial speaker clusters are built according to metadata;
2. *Multiple AVM training*: AVM are trained for each clusters using SAT;
3. *Speaker re-assignment*: each speaker is re-assigned to clusters according to the likelihood of each AVM given adaptation data of the speaker



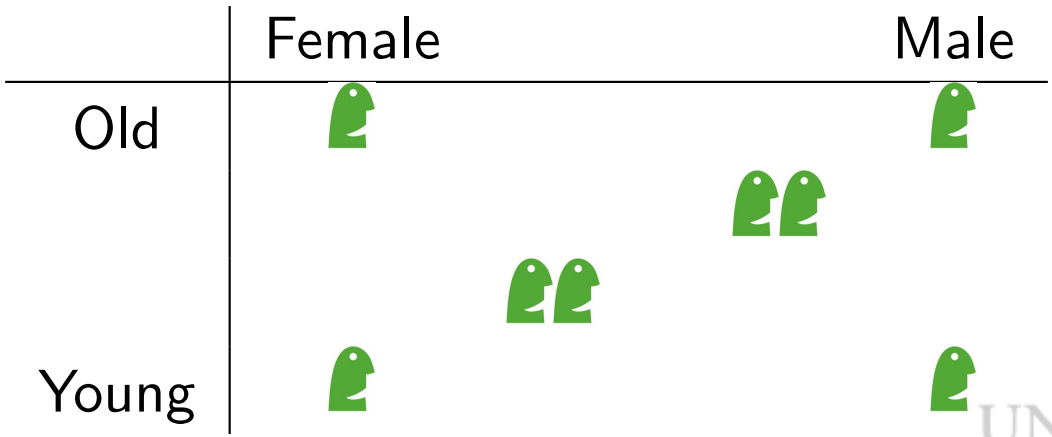
Preliminary results: quality

- Consider Gender/Age factors (Male/Female, Young/Old)
- Compare **broad AVM** with **Multiple-AVM** (M-AVM)
- Quality of the M-AVM voice is perceived better
- But currently not significatively better than the **closest AVM** voice (hard decision) $\rightarrow$ bias due to the choice of the component covariances



Preliminary results: controllability

- The proposed approach can also be used to **design voices** by simply modifying weight vector
- In this experiment:
  - AVM **are not adapted** to the target speaker in order to keep a wide range of variability
  - 4 AVM are used (before re-assignement): YoungFemale, YoungMale, OldFemale, Oldmale



Conclusion

Natural Speech Technology

- New adaptation approach based on the use of several AVM
- Factorisation approach is used for better voice **quality** and **control**
- Future work
  - use different factors (regional accent)
  - improve the adaptation procedure
  - controllability
  - update of the models

UNIVERSITY OF CAMBRIDGE