Building personalised VOCAs

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

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

**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
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)
“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)

**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

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

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

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

- 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

Interpolation weights can be adjusted manually by a SLT
Voice Reconstruction

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

KL Distances

Patient-adapted voice models

Average voice models

outliers

KL-Distance
Voice Reconstruction

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

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

- 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

**Listening tests (38 listeners):**

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

**Similarity Test**

**Intelligibility Test**
Voice Reconstruction

Selected approach: model interpolation with KLD-based confidence measure

- **Short recording**
- **Voice clone**
- **Repaired voice**

**speaker adaptation**

- **Voice catalogue**
- **Healthy reference** (AVM or selected donor)

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

- **patient**
- **accent specific AVM**
- **voice donors**
The need of accent-specific AVMs

**Implements 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

**ABX Test** ($X = \text{target speaker}$)

- 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

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

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

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

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

**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

More than 900 voice donors already
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:

  ![Diagram showing speaker clustering hierarchy]

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

  - Female average:
    - Scottish
    - Irish
    - Welsh
    - English
    - CentralEast
    - CentralWest
    - North

  - Male average:
    - Scottish
    - Irish
    - Welsh
    - English
    - SouthEast
    - North
    - SouthWest
    - Midlands
Speaker clustering

Voice corpus → Accent Distance Measure (ACCDIST) → Hybrid hierarchical clustering → Accent dependent AVMs

Prior Accent Knowledge (metadata)

Hierarchical clustering based on metadata

Hierarchical clustering based on acoustic distance

gender

country
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)

<table>
<thead>
<tr>
<th></th>
<th>SouthEast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowel Distance</td>
<td>father</td>
</tr>
<tr>
<td>after</td>
<td>2.27</td>
</tr>
<tr>
<td>father</td>
<td>0</td>
</tr>
</tbody>
</table>

- 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 metadata versus the best AVM derived from acoustic data (hierarchical clusters)
- Reference is the target speaker voice

![Similarity (MOS)](image)
Tools for clinical trial

- Healthy volunteers
  - Voice donation and banking

- Voice catalogue
  - Voice repair
  - Short recording
  - Voice clone
  - Voice banking

- Patients with speech disorders
  - Short recording
- Patients who can speak well

- 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
Speech Recorder

- 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)

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

- iOS application
- Automatic download of the repaired voice model
- Offline synthesis
- Feedback form
Delivering Voices: Speak Unique
Voice evaluation

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

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

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
Conclusions

- 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