A Non-Parametric Articulatory-to-Acoustic Conversion System for Silent Speech using Shared Gaussian Process Dynamical Models

Jose A. Gonzalez[1], Phil D. Green[1] , Roger K. Moore[1], Lam A. Cheah[2], and James M. Gilbert[2]

(1)Dept. Computer Science, University of Sheffield, UK; (2) School of Engineering, University of Hull, UK

1. Introduction

Silent Speech Interfaces (SSIs)

- Motivation
  - Patients with larynx cancer often lose their voice after *laryngectomy*.
  - Existing methods for voice restoration are unsatisfactory.
  - SSIs enable speech communication when the audible acoustic signal is unavailable by exploiting other *speech-related biosignals*.
  - Devices for capturing articulator motion data: cameras, ultrasound, surface electrodes or **PMA**.
- SSI approaches

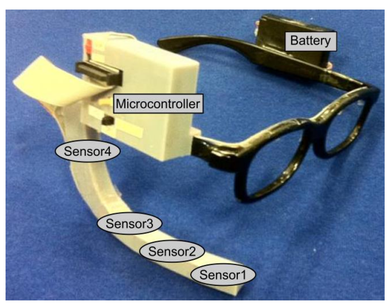
  a) ASR from articulator motion data + TTS synthesis.

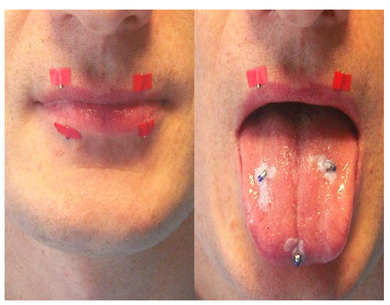
  b) Direct transformation of the articulator data to audible speech.

About this Work

- Summary
  - In previous work we have shown that it is possible to recognise speech from PMA data.
  - Here, we investigate the use of shared Gaussian process dynamical models (**SGPDMs**) for articulatory-to-acoustic conversion .
  - Results are reported in which audible speech is synthesised from PMA data for two speakers with no speech impairment.
  - Preliminary results are very promising, outperforming state-of-the-art GMM-based conversion, but further investigation is needed.
  - The ultimate goal is to **restore the ability to communicate to laryngectomees**.

2. Permanent Magnet Articulography (PMA)





- How PMA works
  - Small magnets are attached to the lips and tongue of the patient.
  - The magnetic field generated when the patient ‘speaks’ is captured by the magnetic sensors.
  - PMA does not provide the exact position of the magnets.

3. Articulatory-to-Acoustic Mapping

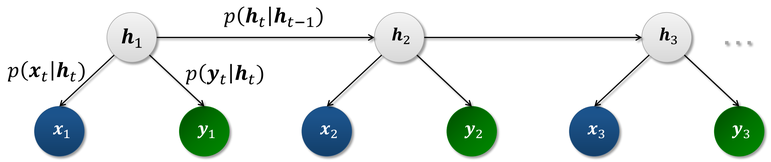
Problem Formulation

- Generative model
  - Mapping between PMA vectors $\boldsymbol{x}_t$ and speech parameter ones $\boldsymbol{y}_t$:

$$\boldsymbol{y}_t = \boldsymbol{f}(\boldsymbol{x}_t)$$

  - We assume that $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ are the outputs of an underlying stochastic process with hidden state $\boldsymbol{h}_t$:

$$\boldsymbol{x}_t = \boldsymbol{f}_x(\boldsymbol{h}_t) + \boldsymbol{\epsilon}_x$$
$$\boldsymbol{y}_t = \boldsymbol{f}_y(\boldsymbol{h}_t) + \boldsymbol{\epsilon}_y$$



  - Two problems
    - *Training:* estimation of $p(\boldsymbol{x}_t|\boldsymbol{h}_t)$, $p(\boldsymbol{y}_t|\boldsymbol{h}_t)$ and $p(\boldsymbol{h}_t|\boldsymbol{h}_{t-1})$.
    - *Conversion:* estimate the most likely sequence of speech parameter vectors for the source sequence $\mathbf{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots)$.

Shared Gaussian Process Dynamical Models

- Statistical modelling
  - **Gaussian processes**

$$p(\boldsymbol{z}|\boldsymbol{h}) = N(m(\boldsymbol{h}), k(\boldsymbol{h}, \boldsymbol{h}'))$$

    $m(\boldsymbol{h})$ and $k(\boldsymbol{h}, \boldsymbol{h}')$ are the mean and covariance (kernel) functions.
  - In a SGPDM we have several GPs sharing the same latent space + a dynamical model in the latent space.
  - Data modelling
    - $m(\boldsymbol{h}) = \mathbf{0}$.
    - An RBF kernel is used for the observation models:

$$k_x(\boldsymbol{h}, \boldsymbol{h}') = \beta_1 \exp\left(-\frac{\beta_2}{2}\|\boldsymbol{h} - \boldsymbol{h}'\|^2\right) + \frac{\delta_{\boldsymbol{h},\boldsymbol{h}'}}{\beta_3}$$

    - For the dynamical model, we use an RBF+linear kernel:

$$k_H(\boldsymbol{h}, \boldsymbol{h}') = \alpha_1 \exp\left(-\frac{\alpha_2}{2}\|\boldsymbol{h} - \boldsymbol{h}'\|^2\right) + \alpha_3 \boldsymbol{h}^T \boldsymbol{h}' + \frac{\delta_{\boldsymbol{h},\boldsymbol{h}'}}{\alpha_4}$$

- Training and conversion phases

Training	Conversion

• **Model parameters:** kernels hyperparameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}\}$ and shared latent coordinates $\boldsymbol{H}$.
• **Loss function**: $\mathcal{L} \equiv p(\boldsymbol{H}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{X}, \boldsymbol{Y}) \propto p(\boldsymbol{X}|\boldsymbol{H}, \boldsymbol{\alpha}) p(\boldsymbol{Y}|\boldsymbol{H}, \boldsymbol{\beta}) p(\boldsymbol{H}|\boldsymbol{\gamma}) p(\boldsymbol{\alpha}) p(\boldsymbol{\beta}) p(\boldsymbol{\gamma})$
• Uninformative priors are chosen for $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}\}$.
• $\mathcal{L}$ is optimised using the SCG algorithm.
• $\boldsymbol{H}$ is initialised using canonical correlation analysis (CCA).
	• The latent sequence $\boldsymbol{H}^*$ for $\boldsymbol{X}$ is initialised using the Viterbi algorithm.
	  • Transition probabilities given by $p(\boldsymbol{h}_t|\boldsymbol{h}_{t-1})$.
	  • Observation probabilities given by $p(\boldsymbol{x}_t|\boldsymbol{h}_t)$.
	• Next, $\boldsymbol{H}^*$ is refined using the SCG algorithm.
	• Finally, $\hat{\boldsymbol{Y}}$ is just the mean of $p(\boldsymbol{Y}|\boldsymbol{H}^*)$.

4. Experiments

Conditions

Database	– Isolated digits. – PMA and speech data were recorded simultaneously. – Two native English speakers (with no speech impairment): male & female. – Amount of data: 7.2 minutes (male) & 8.46 minutes (female speaker).
Feature extraction	– Speech signal: 25 MFCCs computed every 10ms [Fs:16khz, window length:20ms]. – PMA signal: features extracted by Partial Least Squares [9 channels @ 100Hz]. – Speech is synthesised with no voicing (i.e. as whispered speech).

Objective evaluation	– The Mel-Cepstral distortion measure is used to evaluate reconstruction accuracy.
	– 10-fold cross-validation scheme is used.
	– SGPDM mapping is compared with GMM-based mapping proposed by Toda’2007.
	  – 32-component GMM is employed.
	  – Both MMSE and MLE estimation algorithms are evaluated.

Results

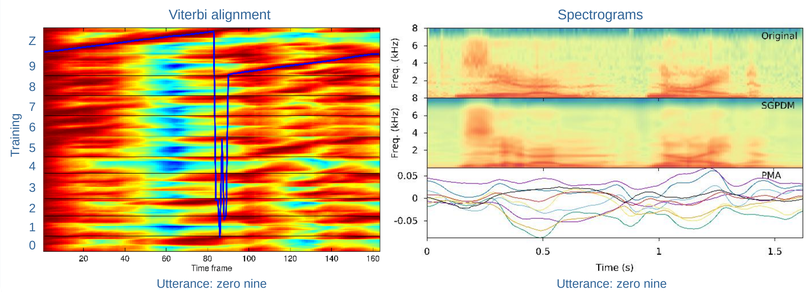
- Experiment 1
  - Conversion is performed using a model trained on the same digit.

Speaker	GMM		SGPDM		
	MMSE	MLE	$D_h = 3$	$D_h = 5$	$D_h = 7$
Male	5.71	5.04	4.37	4.64	4.72
Female	5.99	5.92	4.70	4.89	5.01
Average	5.85	5.48	**4.54**	4.77	4.87

- Experiment 2
  - The transformation is now estimated from sequences of isolated digits.

Speaker	GMM		SGPDM		
	MMSE	MLE	$D_h = 3$	$D_h = 5$	$D_h = 7$
Male	5.04	5.05	4.74	5.22	5.05
Female	5.57	5.64	4.82	5.71	5.97
Average	5.31	5.35	**4.78**	5.47	5.51

- Example: Digit sequence reconstruction



5. Conclusions

- We have presented a non-parametric approach for articulatory-to-acoustic conversion using shared Gaussian process dynamical models.
- Results demonstrate that the approach outperforms state-of-the-art mapping based on GMMs.
- Future research
  - Evaluation: more complex task & more speakers.
  - Model: introduce switching states & evaluate other kernel functions.