Evaluation of a Silent Speech Interface based on Magnetic Sensing and Deep Learning for a Phonetically Rich Vocabulary

Jose A. Gonzalez[1], Lam A. Cheah[2], Phil Green[1], James M. Gilbert[2], Stephen R. Ell[3], Roger K. Moore[1] and Ed Holdsworth[4]

[1] Department of Computer Science, University of Sheffield, Sheffield, UK
[2] School of Engineering, University of Hull, Kingston upon Hull, UK
[3] Hull and East Yorkshire Hospitals Trust, Castle Hill Hospital, Cottingham, UK
[4] Practical Control Ltd, Sheffield, UK

The University Of Sheffield.

1. Introduction

Background

- People diagnosed with larynx cancer often lose voice after laryngectomy.
- Existing methods for voice restoration are unsatisfactory.
- Alternative solution: synthesise speech from captured movement of the articulators.
- Two approaches:
  - ASR from articulator movement followed by TTS synthesis.
  - **Direct speech synthesis from the articulatory data.**

About this work

- In previous work we have shown that it is possible to synthesise speech from articulatory data.
- Here, we compare 3 machine learning techniques for modelling the articulatory-to-acoustic mapping: GMM, DNN & RNN.
- Techniques evaluated on phonetically-rich database with simultaneous PMA-and-speech recordings for 6 healthy subjects.
- **Best results are obtained by a RNN approach with fixed-latency, which is suitable for real-time processing.**

2. Permanent Magnet Articulography (PMA)

- Technique for capturing speech articulator movement.
- How it works:
  - Small magnets are attached to the lips and tongue.
  - Magnetic field generated by the magnets when the person 'speaks' captured by sensors close to the mouth.
- Main advantage w.r.t. other methods: potentially unobtrusive.

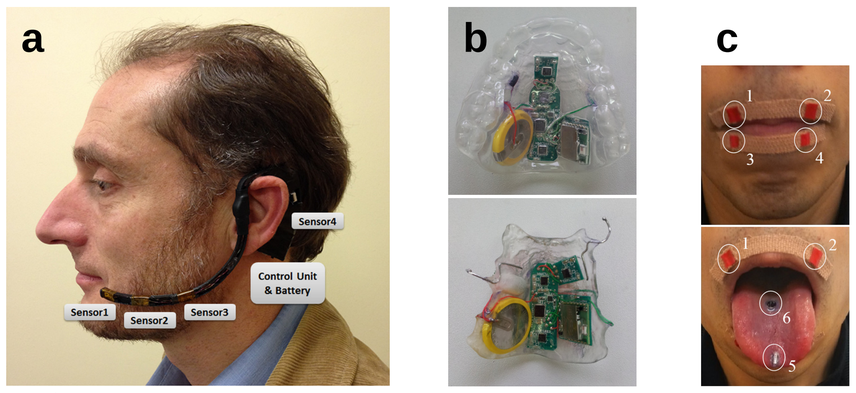


Figure 1: **(a)** External PMA headset; **(b)** Intra-oral versions; **(c)** Placement of magnets on the lips and tongue.

3. Direct Speech Synthesis from PMA data

Overview

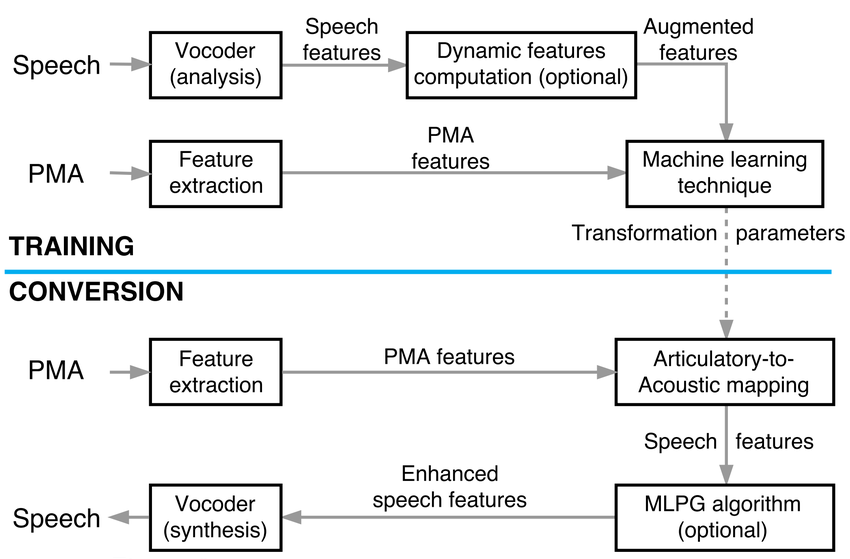


Figure 2: Training and conversion phases of our silent speech system.

Details

- **Speech vocoder**: STRAIGHT (25 MFCCs, 5 BAPs, log F0 and U/V).
- **PMA features**: PLS dim. reduction technique applied over sliding windows containing several PMA samples.
- **Dynamic features**: $\Delta_t = y_t - y_{t-1}$.
- **MLPG algorithm**: improves predictions by taking into account the dynamic speech features (only for GMM & DNN).

$$\hat{Y} = \arg\max_{Y} p(WY|X)$$

W is the matrix for computing the dynamic features.

4. Articulatory-to-Acoustic Mapping

Gaussian Mixture Model (GMM)

- **Training**: Joint pdf of source and augmented target vectors modelled as a GMM.

$$p(x, \overline{y}) = \sum_{k=1}^{K} \pi^{(k)} \mathcal{N}\left(\begin{bmatrix} x \\ \overline{y} \end{bmatrix} ; \begin{bmatrix} \mu_x^{(k)} \\ \mu_{\overline{y}}^{(k)} \end{bmatrix}, \begin{bmatrix} \Sigma_{xx}^{(k)} & \Sigma_{x\overline{y}}^{(k)} \\ \Sigma_{\overline{y}x}^{(k)} & \Sigma_{\overline{y}\overline{y}}^{(k)} \end{bmatrix} \right)$$

- **Conversion**: MLPG algorithm predicts the most likely sequence of speech parameters from the sequence of PMA features (under both static & dynamic constraints).

Deep Neural Network (DNN)

- **Training**: DNN directly models the mapping between PMA & speech feature vectors.
- **Conversion**: MLPG applied to post-process the DNN outputs.

Recurrent Neural Network (RNN)

- **Training**: evaluated the following GRU-RNN architectures:
  - Bi-directional RNN (BiRNN): $y_t = f(x_1, \ldots, x_t, \ldots, x_T)$
  - Fixed-lag RNN: $y_t = f(x_1, \ldots, x_t, \ldots, x_{t+\delta})$
- **Conversion**: no MLPG-based post-processing applied.

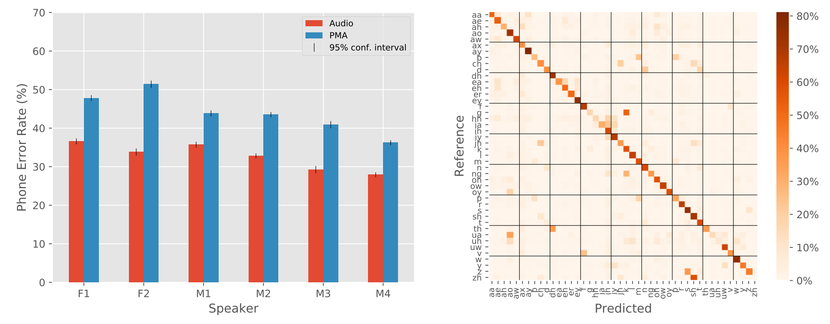
5. Results

Conditions

Database	• CMU Arctic sentences • PMA & speech recorded simultaneously (Fs: 100 Hz & 48 kHz) • 6 healthy British subjects: 4 males & 2 females • Amount of data: 20 to 35 min. recorded per subject (Avg. 25.5 min)
Feature extraction	• Audio: STRAIGHT (win. len: 25 ms, shift: 5 ms) • PMA: PLS dim. reduction applied to inputs • GMM & DNN inputs: windows with 21 frames • RNN inputs: individual PMA frames
Model training	• Speaker-dependent models trained for each speech feature • GMM: 128 mixtures & full covariances • DNN: 4 hidden ReLU layers • RNN: 4 hidden layers & 150 GRUs in each layer (look-ahead 50 ms)
Evaluation	• 10-fold cross-validation • Objective speech quality metrics • ABX listening test on speech quality (18 listeners)

Experiment 1

- Frame-wise phone classification from audio & PMA using RNNs.



(a) (b)

Figure 3: **(a)** Phone classification results. **(b)** Normalized confusion matrix for PMA phone classification.

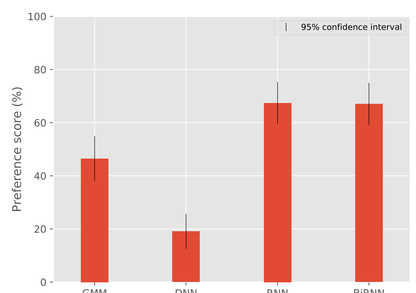
- Average results: 32.68% audio vs. 43.95% PMA.
- For PMA most confusions are for voicing, manner of articulation & phones at back of the mouth.
- Those areas are not well captured by current PMA device.

Experiment 2

- Comparison of techniques for articulatory-to-acoustic mapping.

	MCD (dB)	BAP (dB)	F_0 RMSE (Hz)	U/V error rate (%)
GMM	5.84	4.56	26.30	18.71
DNN	5.74	4.68	26.13	15.88
RNN	5.56	**4.50**	**23.77**	13.23
BiRNN	**5.52**	4.52	24.87	**13.12**

(a)



(b)

Figure 4: **(a)** Objective distortion metrics. **(b)** Results of the ABX listening test on speech quality.

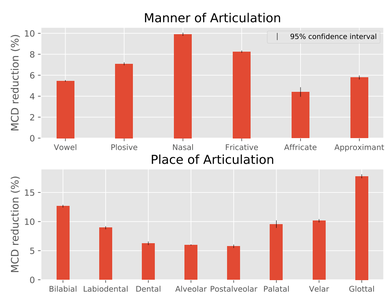
- Best mapping method: RNNs.
- RNNs better capture long-term correlations of F_0 and voicing.
- Fixed-lag RNN approaches BiRNN performance but with less latency.
- DNNs obtains lower subjective scores than GMMs (samples sound 'buzzy').
- Might be related with the poor performance for BAP estimation.

Experiment 3

- Direct synthesis using linguistic information.

Input features	MCD (dB)	BAP (dB)	F_0 RMSE (Hz)	U/V error rate (%)
PMA	5.56	4.50	23.77	13.23
Phones	4.95	**3.72**	**22.70**	**8.84**
Senones	5.15	3.85	22.97	9.28
PMA+Phones	**4.75**	3.79	22.81	9.18
PMA+Senones	4.90	3.88	22.88	9.65

(a)



(b)

Figure 5: **(a)** Objective distortion metrics for the fixed-lag RNN mapping. **(b)** Relative improvements of PMA+Phones w.r.t. PMA.

- Are PMA data and linguistic information complementary?.
- Ling. features derived from the text transcripts (i.e. cheating).
- Ling. features provide more information than PMA data alone.
- Particularly for the excitation parameters.
- Best results obtained when PMA & ling. features combined.
- Biggest improvements achieved for the phones that PMA models more poorly.

6. Conclusions

- Described a technique for synthesising audible speech from articulator movement data.
- Speech generated has a reasonable quality and is fairly intelligible.
- Further improvements could be achieved in future by exploiting ling. information.