Analysis of Phonetic Similarity in a Silent Speech Interface based on Permanent Magnetic Articulography

Jose A. Gonzalez[1], Lam A. Cheah[2], Jie Bai[2], Stephen R. Ell[3], James M. Gilbert[2], Roger K. Moore[1], Phil D. Green[1]

(1)Dept. Computer Science, University of Sheffield, UK; (2) School of Engineering, University of Hull, UK; (3) Hull and East Yorkshire Hospitals Trust, Castle Hill Hospital, UK

1. Introduction

Silent Speech Interface (SSI)

- Motivation
  - Patients with laryngeal cancer often lose their voice after *laryngectomy*.
  - Existing methods for voice restoration are still unsatisfactory.
  - SSIs enable speech communication to take place when the audible acoustic signal is unavailable.
  - Cameras, ultrasound sensors, and surface electrodes, among others, can be used to capture the movements of the remaining articulators.
- SSI approaches

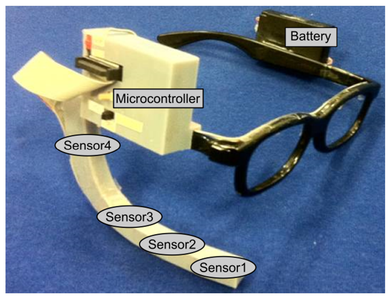
a) Speech is recognized from captured articulator movement and, then, TTS can be used to synthesise the text.

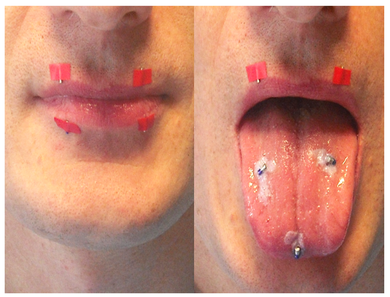
b) Direct transformation of the articulator movement to audio.

The DiSArM Project

- Summary
  - Our goal is to restore the ability to communicate to patients who have undergone laryngectomy.
  - In previous work we have shown that it is possible to recognize speech from articulator movement captured using **PMA**.
- Aim of this work
  - We investigate the suitability of PMA for applications involving unconstrained, phonetically rich speech (i.e. recognition & synthesis).
  - In particular, we investigate to what extent information about the speech production process can be recovered from PMA data.

2. Permanent Magnetic Articulography (PMA)

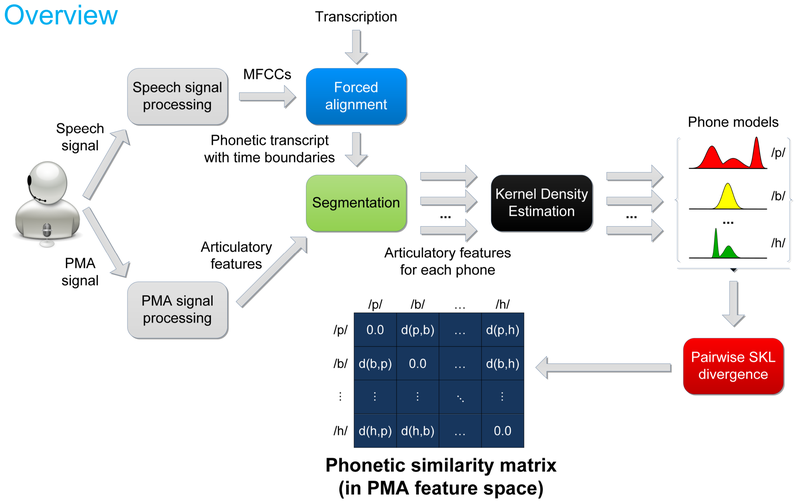




- How PMA works
  - Small magnets are attached to the lips and tongue of the patient.
  - The magnetic field generated when the patient 'speaks' is captured by triaxial magnetic sensors mounted on a headframe.
  - PMA does not provide the exact position of the magnets. Rather, articulator movement patterns are learned using machine learning.

3. Methodology

Overview



Details

Data acquisition	– 420 utterances from the CMU Arctic corpus were uttered by a male native British English adult with normal speaking ability. – 22 minutes of material (audio & PMA) were simultaneously recorded.	
Feature extraction	**Audio**	**PMA**
	– Samp. frequency: 16 KHz. – Audio features: 25 MFCCs.	– Samp. frequency: 200 Hz. – 9 channels of data. – PCA is applied to reduce the dimensionality from 45 to 25.
	– Analysis window: 25-ms length & 10-ms overlap. – The (audio and PMA) features are normalized in mean and variance.	
Segmentation	– MAP was used to adapt a cross-word triphone-based SI model to the speaker's voice. – The SD model was used to obtain force-aligned phonetic transcripts of the audio. – The aligned transcripts were used to segment the PMA material.	

Phonetic Similarity

- Introduction
  - We want to determine to what extent the English phones can be differentiated using PMA.
  - The SKL divergence is used to compare the distributions estimated for each pair of phones in the PMA feature space.
  - The phone distributions are estimated using KDE.
- Kernel Density Estimation (KDE)
  - Let $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be the PMA feature vectors for a given phone and $K(\boldsymbol{x})$ an isotropic Gaussian kernel function. Then,

$$p(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} K(\boldsymbol{x} - \boldsymbol{x}_i)$$

- Symmetric Kullback-Leibler (SKL) divergence
  - The SKL divergence between two distributions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ is

$$d(p,q) = \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x} + \int q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} d\boldsymbol{x}$$

  - Properties: $d(p,q) \geq 0$ for all p, q and $d(p,q) = 0$ iff $p = q$.

4. Results

Details

Phonetic coverage	– The phonetic similarity is evaluated only for the consonants. – Rather than comparing all the phones individually, three different analysis are performed, one for each aspect of speech production: voicing, place and manner. – The phonetic categories are those in the IPA chart.
Similarity results	– The results obtained for both the audio & PMA features are compared below. – SKL divergences: small values indicate higher similarity between the phones since their distributions are more overlapped.

Phonetic Similarity Results

- Voicing

MFCC

	Voiced	Unvoiced
Voiced	0.00	0.90
Unvoiced	0.90	0.00

PMA

	Voiced	Unvoiced
Voiced	0.00	0.20
Unvoiced	0.20	0.00

- Manner of articulation (MoA)

MFCC

	Plosive	Nasal	Fricative	Affricate	Approximant
Plosive	0.00	0.80	0.40	0.52	0.95
Nasal	0.80	0.00	1.83	1.67	0.86
Fricative	0.40	1.83	0.00	0.72	1.59
Affricate	0.52	1.67	0.72	0.00	1.66
Approximant	0.95	0.86	1.59	1.66	0.00

PMA

	Plosive	Nasal	Fricative	Affricate	Approximant
Plosive	0.00	0.16	0.29	1.16	0.37
Nasal	0.16	0.00	0.50	1.53	0.47
Fricative	0.29	0.50	0.00	1.08	0.68
Affricate	1.16	1.53	1.08	0.00	1.60
Approximant	0.37	0.47	0.68	1.60	0.00

- Place of articulation (PoA)

MFCC (upper triangular matrix) / PMA (lower triangular matrix)

PMA \ MFCC	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Palatal	Velar	Glottal
Bilabial	0.00	0.88	0.59	0.61	1.80	1.88	0.38	1.96
Labiodental	0.53	0.00	0.44	0.47	0.99	2.16	0.59	0.95
Dental	1.80	2.20	0.00	0.41	1.16	1.90	0.54	1.39
Alveolar	0.53	0.70	1.51	0.00	1.01	1.62	0.46	1.19
Postalveolar	1.89	2.24	3.55	1.17	0.00	1.77	1.18	1.26
Palatal	1.70	2.09	2.10	1.66	1.40	0.00	1.36	2.31
Velar	1.09	0.83	2.62	1.28	1.85	1.29	0.00	1.19
Glottal	0.50	0.77	1.85	0.73	1.77	1.54	1.08	0.00

5. Discussion

- PMA achieves comparable accuracy to audio for discriminating the PoA.
- Other aspects of the speech production process (i.e. voicing & MoA) are less accurately modelled by PMA.
- Comparison with previous work shows that the articulator dynamics could play an essential role in phone identification.
- Future research will address the improvement of the current PMA prototype for better phone modelling.