Evaluation of a Silent Speech Interface Based on Magnetic Sensing

Robin Hofe1, Stephen R. Ell2, Michael J. Fagan2, James M. Gilbert2, Phil D. Green1, Roger K. Moore1, Sergey I. Rybchenko2

1Department of Computer Science, University of Sheffield, UK
2Department of Engineering, University of Hull, UK
r.hofe@sheffield.ac.uk

Abstract

This paper reports on isolated word recognition experiments using a novel silent speech interface. The interface consists of magnetic pellets that are fixed to relevant speech articulators, and a set of magnetic field sensors that measure changes in the overall magnetic field created by these pellets during speech. The reported experiments demonstrate the effectiveness of this technique and show the suitability of the system, even at its early stages of development, for small vocabulary speech recognition.

Index Terms: silent speech interfaces, clinical speech technology, articulography, multimodal speech recognition

1. Introduction

The “Recognition and Reconstruction of Speech following Laryngectomy” (REdRESS) project aims to provide a speech interface for patients who have lost their ability to vocalise, for example after laryngectomy or diseases affecting the vocal cords. However, the technology could also be applied in other areas where silent, non-acoustic or multimodal speech recognition is of interest, for example in communication equipment used in extreme situations by emergency units [1].

This paper reports on the performance of the REdRESS silent speech interface in isolated word recognition experiments on a vocabulary of 57 words, using hidden-Markov-models (HMM) [2]. The word error rates obtained from the silent speech system are compared to those that result from using Mel frequency cepstral coefficients (MFCC) extracted from an acoustic reference signal. The paper is structured as follows.

A brief description of the interface itself is given in the following section (more detail can be found in [4]). The experimental methodology is explained in section 3, followed by the experiments themselves and their results in sections 4 and 5, respectively. The penultimate section discusses the results and gives an overview of the future development of the silent speech interface; the last section concludes.

2. The REdRESS Silent Speech Interface

The idea behind the REdRESS silent speech interface [3] is a variation of common articulography. State-of-the-art articulographs employ sensors that can be located in a three-dimensional, time-dependent magnetic field produced by a set of transmitters [5]. The transmitters are arranged around a test subject’s head and the sensors are fixed to relevant points on surfaces of speech articulators. The movements of the articulators during speech can then be reconstructed from the recorded locations and orientations of the sensors [6].

In order to provide a more convenient silent speech interface suitable for everyday use, this technique was altered in two ways. Firstly, the roles of magnetic field sources and sensors were exchanged. This means that in REdRESS, small magnetic pellets are placed on speech articulators, inducing a magnetic field that is picked up by sensors which are mounted on a wearable frame, similar to a microphone headset (see figure 1). Secondly, the magnetic field is not translated into discrete positions of the pellets at any given moment in time. Rather, changes in the magnetic field that occur during speech are used to train a speech recogniser directly by replacing the usual acoustic features by vectors of measured magnetic field components. Recognised utterances can then be used either to control technical devices, or to drive a speech synthesiser.

Figure 2 illustrates the placement of the magnetic pellets in the current system. There are a total of seven magnets, four on the lips, one on the tongue tip and two on the sides of the tongue. This configuration follows from the necessity to capture sufficient data from important articulators while deploying a small number of magnets in an arrangement which might realistically be implanted permanently in users in a practical system [7]. In the prototype system, the magnetic pellets are glued onto the skin and the tongue surface, respectively.

The magnetic field is measured at strategic points around the speaker’s face. The current system uses six magnetic field sensors that measure two magnetic field components each, i.e. 12 data channels can be used for speech recognition.
3. Methodology

3.1. Test subject

The data used in the experiments presented herein were collected from a single male native English speaker (age 46) who has normal speaking ability. Since the REDRESS interface is designed as a personal device that will be calibrated to the individual user, an evaluation of inter-speaker performance was not carried out. However, the system will be trained and tested on other individual speakers in the future, including medical trials with patients who are about to undergo laryngectomy or who have already had the operation.

3.2. Vocabulary

The vocabulary consisted of 57 items (see table 1), composed of the digits from “zero” to “nine” and a word list that covers the complete ARPAbet phonetic inventory [8]. It is important to note that there are a number of instances of minimal pairs in this vocabulary (e.g. bond/pond and two/zoo), increasing the difficulty of the recognition task.

Table 1: The 57 word vocabulary used in the experiments.

<table>
<thead>
<tr>
<th>one</th>
<th>two</th>
<th>three</th>
<th>four</th>
<th>five</th>
<th>six</th>
<th>seven</th>
<th>eight</th>
<th>nine</th>
<th>zero</th>
</tr>
</thead>
<tbody>
<tr>
<td>ahead</td>
<td>batman</td>
<td>batter</td>
<td>battle</td>
<td>bed</td>
<td>bond</td>
<td>bottom</td>
<td>boy</td>
<td>bud</td>
<td>child</td>
</tr>
<tr>
<td>dug</td>
<td>file</td>
<td>got</td>
<td>how</td>
<td>how</td>
<td>how</td>
<td>how</td>
<td>how</td>
<td>how</td>
<td>how</td>
</tr>
<tr>
<td>hoot</td>
<td>rope</td>
<td>roses</td>
<td>see</td>
<td>see</td>
<td>see</td>
<td>see</td>
<td>see</td>
<td>see</td>
<td>see</td>
</tr>
<tr>
<td>rope</td>
<td>rope</td>
<td>rope</td>
<td>rope</td>
<td>rope</td>
<td>rope</td>
<td>rope</td>
<td>rope</td>
<td>rope</td>
<td>rope</td>
</tr>
</tbody>
</table>

3.3. Data recording

All magnetic sensor data were collected while the words were being spoken, and with simultaneous acoustic recording, so that we have the speech and the corresponding sensor signals for each utterance in the data sets. The acoustic signal was recorded with a shock-mounted AKG C1000S condenser microphone and a Lexicon Lambda USB sound card; the six Honeywell HMC1022 magnetic field sensors were connected to a 12-bit ADC PCI board. The two data streams were digitised with sampling frequencies of 16kHz (speech) and 500Hz (sensors), respectively, using separate hardware. This experimental set-up is illustrated in figure 3.

Synchronisation of the two data streams was necessary due to small deviations from the ideal sampling frequencies of the analog-to-digital converters (ADC). This was carried out semi-automatically using a manually adjusted time warping factor on the sensor data.

The test speaker’s head was not restrained, although he was asked to avoid large head movements during recording sessions. This was necessary to ensure, as far as possible, that movements relative to the earth’s magnetic field did not contribute significantly to changes in the magnetic field perceived at the sensors’ locations. In the future, a set of reference sensors will be used to remove influences of the earth’s magnetic field from the data computationally.

3.4. Signal processing

The recorded magnetic field data underwent pre-processing before HMM training and recognition in order to remove influences of the earth’s magnetic field and interference from the electronic equipment. Signal processing was performed for each of the 12 channels individually and consisted of:

- Low pass filtering - 10-point moving average filter (first spectral minimum at 50Hz for 500Hz sampling frequency) to counter mains interference and high frequency noise.
- Background cancellation - the earth’s magnetic field is present at each sensor location in the form of a large offset that changes with movements of the head. This offset was cancelled by median subtraction over 2s windows with a 50% overlap. With expected word lengths of ~0.5s, the median over any 2s window corresponds to the background.
- Normalisation - to a range of values between 0 and 200.
4. Experiments

4.1. Recording sessions

Four data sets were collected, each containing five instances of all 57 items of the vocabulary, i.e. 285 utterances per data set and 20 instances of each item over all four sets. The word items were presented in random order on a computer screen, with a gap of 2.5s between words.

All four sessions took place on the same day, with short breaks in-between. This allowed the magnetic pellets to stay in place throughout the four sessions, avoiding data inconsistencies arising from imprecision in pellet replacement. Each of the four recording sessions lasted twelve minutes.

4.2. Endpointing and label correction

All data were endpointed in the audio domain, using an algorithm that was based on zero-crossing rate, total energy and the relatively stable interval of 2.5s between word starts. These automatically-derived endpoints were checked and corrected manually in a bespoke Matlab graphical user interface (GUI). Endpoints for the sensor data were then derived from the audio endpoints.

During the manual endpoint correction process, two items were rejected because of noise occurring during speech, caused by the speaker kicking the table. These items were instances of one and mother, respectively. Two further items were misread by the speaker. In one case, hide was replaced by hid and in the other, bottom by button.

After deletion of two items and label correction for two more, each word in the vocabulary was represented 20 times over all four data sets with the following exceptions. Bottom, hide, mother and one were represented by 19 instances, hid and button by 21. There were 1138 utterances in total.

4.3. HMM training and recognition

In order to cancel out drifts in the data over the four sessions, for example due to fatigue of the test speaker, all four data sets were combined into a single large set and then randomly split back into four subsets. Subsets 1 to 3 contain five randomly chosen instances of each word in the vocabulary, i.e. 285 utterances each. Subset 4 consists of the remaining items, i.e. four to six instances per word and 283 utterances in total (see deletion and label correction above).

Jack-knife training and testing was carried out in four cycles, using three subsets for HMM training and the fourth for testing the resulting models, respectively.

The pre-processed magnetic field sensor data (see section 3.4) was used directly in HMM training and the audio data was represented by 13 MFCC features (extracted according to the ETSI Aurora standard [9]). For both sensor and audio data, whole-word HMMs with eight emitting states (with single Gaussian distributions) were trained using the Viterbi training algorithm implemented in the Hidden Markov Model Toolkit (HTK) [10]. Recognition was performed using the endpointed data with start and end points of each utterance known to the recogniser a priori. It was a forced decision scenario, where one of the 57 labels of the vocabulary had to be chosen for each item in the respective test set. An overall word error rate was derived from the performances over all four jack-knife cycles.

4.4. Conditions

Five different conditions were tested, all of them based on the same data and the same random split into four subsets, as described above. The conditions were:

- **Sensor** - training and testing directly on the 12 channels of the sensor data;
- **SensorD** - as above, plus the first time derivatives of each channel’s data (i.e. articulator velocity, ‘delta features’);
- **SensorDD** - as above, plus the second time derivatives of each channel’s data (i.e. articulator acceleration, ‘delta-deltas’);
- **SensorDex** - first time derivatives of sensor data exclusively (12 channels);
- **Audio** - reference condition based on MFCCs extracted from the audio signal, without time derivatives.

An overview of the five conditions is presented in Table 2.

Table 2: Vector sizes for the different experimental conditions.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Original</th>
<th>1st deltas</th>
<th>2nd deltas</th>
<th>Vector size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensor</td>
<td>✓</td>
<td></td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>SensorD</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>SensorDD</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>36</td>
</tr>
<tr>
<td>SensorDex</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>12</td>
</tr>
<tr>
<td>Audio</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>13</td>
</tr>
</tbody>
</table>

5. Results

The recognition performances for each of the experimental conditions are shown in Table 3.

Table 3: Results for the different experimental conditions.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Errors over all 1138 items</th>
<th>Minimal pair confusions</th>
<th>Word accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensor</td>
<td>131</td>
<td>45</td>
<td>88.5%</td>
</tr>
<tr>
<td>SensorD</td>
<td>37</td>
<td>24</td>
<td>96.8%</td>
</tr>
<tr>
<td>SensorDD</td>
<td>37</td>
<td>28</td>
<td>96.8%</td>
</tr>
<tr>
<td>SensorDex</td>
<td>15</td>
<td>11</td>
<td>98.7%</td>
</tr>
<tr>
<td>Audio</td>
<td>3</td>
<td>3</td>
<td>99.7%</td>
</tr>
</tbody>
</table>

The first observation is that the silent speech interface is capable of delivering surprisingly accurate recognition on this task, especially in the SensorDex condition, where it approaches the reference MFCC performance. It is extremely interesting that the best results were obtained by using articulator velocities only, a result that will be followed up in future experiments. Adding articulator accelerations (condition SensorDD) does not affect overall performance. The performance increases from Sensor to SensorD, SensorD to SensorDex and SensorDex to Audio were all found to be statistically significant (McNemar’s test [11]: p<0.005).

The confusion matrices of the five experiments reveal that over a third of the errors in the Sensor condition were between minimal pairs, as were all three errors of the reference condition Audio. In the conditions that include sensor data derivatives, between two thirds and three quarters of the errors were minimal pair confusions. Table 4 lists the most prominent of these.
Table 4: Examples of minimal pair confusion. The table gives the total number of confusions (out of 39 or 40 possible) between specific minimal pairs during recognition.

<table>
<thead>
<tr>
<th>Pair</th>
<th>Sensor</th>
<th>SensorD</th>
<th>SensorDD</th>
<th>SensorDex</th>
</tr>
</thead>
<tbody>
<tr>
<td>bond/pond</td>
<td>14</td>
<td>3</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>four/paw</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>heard/had</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>one/won</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>dug/tug</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>heed/heat</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>hid/heed</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

6. Discussion

6.1. General observations

As mentioned previously, one of the potential applications of the REdRESS silent speech interface is the recognition of small command vocabularies to control technical devices, for example in situations where it is difficult or impossible to acquire a clean audio signal. Since these vocabularies can be designed to avoid minimal pair confusion, it is clear that the interface does have a potential for this purpose.

One factor that might have influenced the current system’s performance is the definition of start and end points of utterances in the audio domain. It is reasonable to expect relevant articulatory activity to take place shortly before the acoustic speech signal starts and probably for some time after it has ended. However, this was not considered in these early experiments as the audio signal is currently more amenable to endpointing; the fluctuations of the magnetic field during speech are a completely new representation of speech that has yet to be explored in detail.

The performance of the REdRESS silent speech interface at this early stage of development is very promising. Considering that there are numerous possibilities for further improvement, there is realistic potential to extend the system for large vocabulary speech recognition in the future.

It has to be remarked that, in the experiments presented, auditory feedback was available to the test speaker. Future trials will have to evaluate the system’s performance when there is no such feedback available, as will be the case in patients after laryngectomy.

6.2. Minimal pairs

One of most difficult minimal pairs to distinguish on basis of the sensor data was bond/pond over all conditions. It is, however, noteworthy that discrimination improved when articulator velocities were taken into account. Initially, the distinction between /b/ and /p/ was not expected to be feasible using a technique that does not have access to any information about voicing. The same is true for other voiced/unvoiced minimal pairs, especially in the SensorDex condition.

Another surprising observation is that replacing a representation of articulatory positions (Sensor) with velocities improved the distinction between most tested minimal pairs, with the exception of hid/heed. To verify and to further investigate this effect will be a task for future work.

6.3. Future improvements

The interface is still in its early stages of development and there are a number of ways in which the performances will be improved in future versions. Amongst these are:

- cancellation of the earth’s magnetic field by a set of reference sensors;
- optimisation of magnetic pellet placements;
- adding further pellets, for example to capture movements of the velum;
- deriving feature vectors from the raw sensor data that are more suitable for statistical modelling.

7. Conclusion

A novel silent speech interface based on measuring the magnetic field produced by magnetic pellets on a speaker’s articulators has been introduced. Experiments have shown the system’s effectiveness in small vocabulary, isolated word recognition, delivering a word accuracy of 98% on a set of 1138 utterances.

The technique is still in its early stage of development and it can be argued that, given the current performance and the obvious potential for improvement, useful performance levels on large vocabularies could be well within reach.

Please visit www.redress-project.org for further information about the REdRESS project.

8. Acknowledgements

REdRESS is funded by Action Medical Research, UK, grant number AP 1164.

9. References