Development of a (silent) speech recognition system for patients following laryngectomy

M.J. Fagan[a],
S.R. Ell[b],
J.M. Gilbert[a],
E. Sarrazin[a],
P.M. Chapman[c]
[a] Department of Engineering, University of Hull, UK
[b] Department of Otolaryngology, Hull Royal Infirmary, Hull and East Yorkshire Hospitals NHS Trust, UK
[c] Department of Computer Science, University of Hull, UK

Abstract

Surgical voice restoration post-laryngectomy has a number of limitations and drawbacks. The present gold standard involves the use of a tracheo-oesophageal fistula (TOF) valve to divert air from the lungs into the throat, which vibrates, and from this, speech can be formed. Not all patients can use these valves and those who do are susceptible to complications associated with valve failure. Thus there is still a place for other voice restoration options.

With advances in electronic miniaturization and portable computing power a computing-intensive solution has been investigated. Magnets were placed on the lips, teeth and tongue of a volunteer causing a change in the surrounding magnetic field when the individual mouthed words. These changes were detected by 6 dual axis magnetic sensors, which were incorporated into a pair of special glasses. The resulting signals were compared to training data recorded previously by means of a dynamic time warping algorithm using dynamic programming. When compared to a small vocabulary database, the patterns were found to be recognised with an accuracy of 97% for words and 94% for phonemes. On this basis we plan to develop a speech system for patients who have lost laryngeal function.

Keywords

- Speech recognition;
- Rehabilitation;
- Laryngectomy;
- Magnetic sensor;
- Speech system

1. Introduction

Patients with laryngeal cancer, whose larynx must be removed, inevitably lose their voice. Also, as a result of surgery, the viscera involved in swallowing and breathing are separated so that the patient must breathe through their neck via a permanent tracheostomy. The three main methods used currently to restore vocal function may encounter a number of problems and limitations. Sound can be created by swallowing air and belching, forming the sound into words. This is known as 'oesophageal speech' and is difficult to learn, and fluent speech is impossible. Vibrating the soft tissues of the throat by an electrolarynx creates sound, which can be articulated into speech, but the voice is monotonic, 'Dalek-like', and can be difficult to understand. The current 'gold-standard' method is to use a small silicone tracheo-oesophageal fistula speech valve that connects the trachea and the oesophagus [1]. Air, powered by the lungs, is diverted through the fistula into the throat which vibrates, and this is formed into speech. However, although these valves work very well initially, they rapidly become colonised by biofilm in many patients and fail after an average of only 3–4 months [2], [3], [4] and [5]. Various modifications have been tried over the years to discourage biofilm growth (e.g. [6], [7] and [8]), but to date none of these approaches appears to provide a long-term solution to this problem.

Thus there is a need for a fundamental improvement in the current methods for the restoration of speech after laryngectomy. Digital (voiced) speech recognition systems have been the subject of research for a number of years, based on measurement of sound emitted by the speaker [9] and a variety of methods exist for identifying the output (e.g. [10], [11] and [12]). While these techniques give good recognition rates (typically 90% for continuous speech), they are not perfect, particularly in noisy environments and so consideration has been given to the augmentation of acoustic signals by other measurements [13], [14], [15] and [16].

A new approach has recently been proposed by the authors [17] that tracks how an individual's mouth and tongue move when they 'speak' (i.e. mouth the words). Analysis of this information allows a computer to decide what the individual had 'said' and then plays back those words. Eventually this computer generated speech could be in the individual's original voice. If developed, such a system could restore speech to patients who have lost their voice as a result of throat cancer, trauma, destructive throat infections or damage to the laryngeal nerves.

The essence of the proposed system is that by monitoring the motion of the vocal apparatus, it is possible to determine the phonemes and words that an individual wishes to produce. A number of miniature magnets implanted into the appropriate parts of the patient's mouth (e.g. lips, tongue and teeth) will result in a variation in the magnetic field surrounding the mouth during 'speech'. By monitoring these variations in magnetic field and comparing these variations with a database of pre-recorded signals, it is proposed that the best matching word or phoneme may be identified. It is envisaged that the implants would be small enough that they would not be apparent when observing the patient and would not, affect the movement of the mouth, either for sound generation or for eating, etc. The implants used in the ideal system are expected to be 1–2 mm in size and coated with a biocompatible material, e.g. silicone. These may be hidden in the teeth of dentures, or implanted in, or bonded to, the posterior surface of the incisors using standard dental techniques, so that they are hidden from view. Implants may be placed into the lips and tongue by an injection technique under local anaesthetic. It is envisaged that the motion sensing system would be incorporated into the patient's normal attire, for example, in a pair of glasses or a necklace.

Fig. 1 provides an overview of a patient's status pre- and post-laryngectomy for laryngeal carcinoma. At the time of diagnosis, our patient will be able to speak (Fig. 1A), even though his/her voice may have been affected by the disease. Once it has been established that the curative procedure is a laryngectomy the patient may then enter the speech rehabilitation program. After radiographic imaging has been obtained for surgical planning, the patient undergoes minor surgery (Fig. 1B) to implant magnets as described previously. The patient will then undergo a sequence of clearly defined recording procedures (Fig. 1C) in order to generate a complete speech/magnetic field database unique to the patient (Fig. 1F). This process should not delay the patient's progress to curative surgery and would form part of the usual pre-operative work up. After this database has been successfully constructed and tested, the patient will undergo a laryngectomy (removing the patient's ability to speak, Fig. 1D). When the patient has recovered from this operation, they should be able to recalibrate (Fig. 1E) and use the speech system immediately conversing in their original pre-laryngectomy 'normal' voice.

Fig. 1.

At present, the research activity described here is concerned with the identification of the words and/or phonemes mouthed by the patient, with the aim of establishing whether such a system is feasible. There are a number of sophisticated speech generation systems already available that can produce voiced output from word or phoneme data; therefore, such a system will be used to recreate the individual's speech in this application.

2. Experimental investigation

2.1. Equipment

Up to 7 small magnets (typically 5 mm × 2 mm) were attached temporarily to the tongue, lips and teeth of members of the research team (see Fig. 2). The magnets were stuck to the tongue and lips with Histoacryl surgical tissue adhesive (Braun, Melsungen, Germany). Magnets were attached to the teeth using a cellulose veneer fitted over the teeth (not visible in Fig. 2). The sensing system was composed of 6 dual axis magnetic sensors [Honeywell HMC 1022, Plymouth, USA] which have been incorporated into a pair of glasses, as shown in Fig. 3. In each case the sensors were arranged with one sensitive axis vertical. The sensors mounted on the front of the glasses had the second axis aligned across the front of the face while those on the sides had their sensitive axes aligned along the side of the face. After amplification and removal of offset voltages, the 12 outputs from the sensors were captured on a PC for subsequent processing via a 16 channel, 12 bit data acquisition card [18], at a sample rate of 4 kHz. The sampling process, subsequent data analysis and visualisation were achieved using MATLAB.

Fig. 2.

Fig. 3.

2.2. Experimental procedure

In order to evaluate the performance of the proposed system and subsequent analysis methods, a number of experiments were conducted. The purpose of these experiments was to assess whether the sensor output contained sufficient information to allow word/phoneme recognition rather than to establish a complete speech recognition system since much of the processing required in a complete recognition and replay system would be identical to those used in conventional acoustically based speech recognition systems. A set of phonemes representing a cross section of phonetic categories and a set of simple words were selected for trials and the subject was asked to repeat each word or phoneme a number of times while the response of the magnetic sensors was recorded. The subject was encouraged to repeat the words in a consistent manner and to maintain the same head position throughout the trials.

2.3. Analysis methodology

Prior to analysis the raw sensor signals are passed through a low-pass filter with a cut-off frequency of 40 Hz and an attenuation of at least 50 dB at 50 Hz and above. The analysis method used two stages. In the first stage, a template is generated from the training set for each word or phoneme. This template is composed of the sequence of samples of each sensor output averaged over the set of training repetitions. The variability between training repetitions is also stored. In the second stage a 'test' sample is compared to these words/phonemes to identify which template provides the 'best fit'. The comparison is based on an adaptation of the widely used Dynamic Time Warping (DTW) algorithm using Dynamic Programming (DP) [9] and [10]. In essence, this algorithm allows for the varying speed of speaking by applying a non-uniform time warping to incoming speech in order to obtain the minimum distance between the incoming speech and a set of templates. The template which gives the minimum overall distance from the incoming speech is considered the best match.

In these investigations, the Euclidean distance measure commonly used in the DTW algorithm has been augmented to take account of the variability in sensor signals between repetitions of the same word/phonemes in the training set. Thus, the distance measure between the *i*th frame of the incoming speech and the *j*th frame of the template is:

equation(1)

$$d_{i,j} = \sum_{k=1}^{N_k} \left(\frac{x_{i,k} - y_{j,k}}{\sigma_{j,k}} \right)^2$$ di,j=∑k=1Nkxi,k−yj,kσj,k2

where $x_{i,k}$ is the signal from the *k*th sensor for the *i*th frame of incoming speech, $y_{j,k}$ the signal from the *k*th sensor for the *j*th frame of the template, $\sigma_{j,k}$ the standard deviation of the *k*th sensor signal in the *j*th frame of the template and N_k is the number of sensors.

Note that the standard deviation may vary between subsequent frames of the template since there is typically more variability in certain parts of words/phonemes than others.

In the tests reported here the subject had a single magnet attached to the centre of the tongue (with the N–S axis aligned along the centreline of the tongue) and pairs of magnets attached to the upper and lower lips symmetrically positioned about the centre line of the face with north and south poles vertically opposing. No magnets were attached to the teeth. Two training sets were recorded, one consisting of 13 phonemes and the second consisting of 9 words. In each case the words were repeated 10 times. Each repetition occurred 3 s apart and

2 s of data was recorded following a signal for the subject to speak. Test samples were compared against this set of templates and the ‘best fit’ template identified using Eq. (1).

3. Results

A sample of typical responses from 2 of the 12 sensors for the words ‘cat’ and ‘dog’ are shown in Fig. 4. The waveforms shown in Fig. 4, which are taken from the vertical and horizontal sensors mounted on the front of the glasses, have been filtered as discussed above. It can be seen that the waveforms are distinctly different for the two words.

Fig. 4.

For each combination of test signal and training set the minimum distance was calculated and the best fit identified. The distances, given by Eq. (1), for a typical set of comparisons are given in Table 1 for the 13 phonemes tested. These figures have been normalised so that the diagonal terms are unity to allow easier comparison. Table 1 also contains the sounds, taken from the ARPAbet [11] corresponding to each phoneme. As may be seen, all of the off-diagonal terms in Table 1 are larger than the diagonal terms indicating that the phoneme has been correctly identified. In general, the off-diagonal terms are an order of magnitude greater than the diagonals but there are cases where the discrimination is less clear, for instance between labial phonemes (b-m-p-f) and velar (g-k), but the processing is still able to correctly identify the best fit, even where the difference is between voiced and unvoiced versions of the same phoneme (e.g. g-k and b-p). The same method has been applied to a series of nine words with typical results shown in Table 2. Once again, it can be seen that the system is able to correctly identify all of the words.

Table 1
Table 2.

Having conducted 10 trials of each word and each phoneme, it was found that the recognition accuracy was 97% for words and 94% for phonemes. It is believed that the higher recognition accuracy in the case of words is due to the fact that, being longer, there is more information available on which to make the comparison. However, clearly the number of possible words is far greater than the number of possible phonemes and so it is not possible to conclude, on the basis of the small sample considered here, whether it would be better to attempt recognition of whole words or sub-word units.

The results presented above made use of the outputs from all 12 of the available sensors. It was noted that some channels showed significantly more response than others. The four sensors placed at the front of the glasses, with sensing axes coplanar with the lens, produced the greatest outputs; the four mounted on the sides close to the lenses produced the next greatest and those mounted further back on the side arms of the glasses produced the least signal. This was as expected since the side mounted sensors were further from the magnets. The same analysis was performed on a subset of the sensor outputs to assess whether all 12 would be necessary. Sensor outputs were selected on the basis of those which produced the largest standard deviation in their responses to all of the training data. With the eight most active sensors only (those on the front of the glasses and those at the front of the side arms), the recognition rates fell to 93% for words and 87% for phonemes while with only 4 sensors (those mounted on the front of the glasses) the rates were 84% and 58%, respectively. With

these reduced numbers of sensors, it was found that a modified distance measure, in which the standard deviation term is removed from Eq. (1), provided better performance with recognition rates of 98% and 79% using eight sensors and 94% and 73% using four sensors. It is believed that the removal of the standard deviation term reduces the effect of very small signals which are dominated by noise upon the distance measure. This in turn suggests that further improvements to the algorithm, or the use of more sophisticated algorithms, are worthy of investigation.

4. Discussion

The aim of this work was to assess the potential of indirect measurement of movement of the vocal apparatus as a means to determine the intended speech of a subject, rather than to develop a complete speech recognition system. As such the results show considerable promise. Based on a simple template matching algorithm it has been shown that it is possible to classify a subset of phonemes and a small number of words with degree of accuracy which is similar to that achieved for speech recognition based on acoustic information [9] – albeit with a significantly smaller vocabulary data set.

Clearly, there is still considerable work to be carried out before the proposed system could be used in a normal environment. The testing of the system with a complete set of phonemes and/or a larger set of words would be necessary and this might well indicate that more sophisticated analysis algorithms would be required, particularly if continuous speech is to be recognised accurately. In this context, it is important to consider that much of the work carried out on acoustic speech recognition could be applied to this approach. For instance a typical Hidden Markov Model (HMM) based system for isolated word recognition consists of three stages with the information generated by the first stage being a series of Linear Predictive Coding (LPC) coefficients derived from a spectral analysis [12]. These coefficients, which vary from frame to frame of the incoming speech, may be considered analogous to the filtered sensor signals generated by the magnetic sensors and so it may be speculated that subsequent processes could be handled in a similar manner to the standard HMM based recogniser.

In the final system, speed of operation will obviously be critical and it is clear that significant processing power will be required to provide near real-time interpretation and articulation. Furthermore, it seems highly unlikely that the current system will be able to capture all the original voice quality, but it may be that through more complex signal analysis and, perhaps, the sensing of additional elements of the vocal apparatus, it will be possible to extract further voice characteristics, but this will require further research.

The results presented here are based on a system of 12 sensors mounted on a pair of spectacles and five magnets; two pairs on the upper and lower lips and one midway along the centreline of the tongue. The number, locations and orientation of sensors and magnets were based on a combination of convenience and conjecture and it seems likely that other combinations of magnets and sensors would allow improved discrimination between movements and hence more accurate recognition.

Since anatomy, speech pattern and the location of implants will differ between patients, it will be necessary for the system to be trained to recognise an individual's ‘speech’. The training process will build the database that is used to link the detected movements to the corresponding sounds. This will entail the patient repeating prescribed sounds while having

their mouth movements monitored. For patients due to have a laryngectomy, it would be possible to conduct the training while the patient's speech is still intact. It would thus be possible to record the patient's own voice and it would be this voice which could be replayed following the laryngectomy.

Other practical issues related to the implantation of magnets also need further investigation – for example, their effect on the field of an MRI scanner, should the patient require an MRI investigation. This is also an issue with the Provox ActiValve which includes a small magnet to improve valve closure, and was found to cause field distortion around the valve [8]. The system described here will suffer from the same problem, and the magnets would need to be removed if they affected the scan in an area that was diagnostically important.

5. Conclusions

We estimate that there are 16,500 tracheo-oesophageal speech valve changes every year in the UK (based on knowledge of the number of valve changes in our locality). Thus any alternative method of restoring vocal function would be welcome to both patients and health providers.

The aim of this research is to develop a long-term solution, not just for individuals who have had a laryngectomy, but for patients who have lost their voice for other reasons. For laryngectomy patients a fistula, valve and their associated problems would be unnecessary and the patient's original voice could be restored. Such a digital speech rehabilitation system would be suitable for all laryngectomy patients. Some would still elect for valved speech, but for those patients unsuitable for valved speech, who currently can only choose oesophageal speech or external devices to vibrate the soft tissues of their throat, it should be very attractive. In addition, patients who have had extensive resections, repairs or fistula complications, and patients who have lost their voice as a result of trauma, chronic inflammatory disease or laryngeal nerve dysfunction, would be expected to achieve superior speech with the final system.

References

1. E.D. Blom, M.I. Singer
Surgical prosthetic approaches for post-laryngectomy voice restoration
RL. Keith, F.C. Darley (Eds.), Laryngectomy rehabilitation, Texas College Hill Press, Houston (1979), pp. 251–276

2. J.M. Heaton, A.J. Parker
Indwelling tracheo-oesophageal voice prostheses post-laryngectomy in Sheffield, UK: a 6-year review
Acta Oto-Laryngologica, 114 (1994), pp. 675–678

3. S.R. Ell, A.J. Mitchell, A.J. Parker
Microbial colonisation of the Gröningen speaking valve and its relationship to valve failure
Clin Otolaryngol, 20 (1995), pp. 555–556

4. S.R. Ell, A.J. Mitchell, R.T. Clegg, A.J. Parker
Candida: the cancer of silastic
J Laryngol Otol, 110 (3) (1996), pp. 240–242

5. Ell SR. A retrieval study to investigate the failure of silastic speaking valves used post-laryngectomy. M.D. thesis. Leeds; 2000.

6. S.E.J. Eerenstein, P.F. Schouwenburg, L.A. van der Velden, M.F. de Boer
First results of the VoiceMaster prosthesis in three centres in the Netherlands
Clin Otolaryngol, 26 (2001), pp. 99–103

7. E.P.J.M. Everaert, H.F. Mahieu, R.P. Wong Chung, G.J. Verkerke, H.C. van der Mei, H.J. Busscher
A new method for in vivo evaluation on surface-modified silicone rubber voice prostheses
Eur Arch Otorhinolaryngol, 254 (1997), pp. 261–263

8. F.J.M. Hilgers, A.H. Ackerstaff, A.J.M. Balm, M.W.M. van den Brekel, I.B. Tan, J.O. Persson
A new problem solving indwelling voice prosthesis, eliminating frequent candida- and ‘under-pressure’-related replacements: Provox ActiValve
Acta Otolaryngol, 123 (2003), pp. 972–979

9. J. Holmes, W. Holmes
Speech synthesis and recognition
Taylor and Francis (2001)

10. S. Furui
Digital speech processing, synthesis and recognition
(2nd ed.)Marcel Dekker (2001)

11. S.E. Levinson
Mathematical models for speech technology
John Wiley (2005)

12. L. Rabiner
A tutorial on hidden Markov models and selected applications in speech recognition
Proceedings of IEEE, vol. 77, no. 2 (1989), pp. 257–289

13. J. Huang, G. Potamianos, J. Connell, C. Neti
Visual speech recognition using an infrared headset
Speech Commun, 44 (1) (2004), pp. 83–96

14. J.F. Holzrichter, W.A. Lea
Speech articulatory measurements using low power EM-wave sensors
J Acoustic Soc Am, 103 (1) (1998), pp. 622–625

15. D.R. Brown, R. Ludwig, A. Pelteku, G. Bogdanov, K. Keenaghan
A novel non-acoustic voiced speech sensor
Measure Sci Technol, 15 (2004), pp. 1291–1302

16. Betts BJ, Jorgensen C. Small vocabulary recognition using surface electromyography in an acoustically harsh environment. Tech. memo TM-2005-213471, NASA; 2005.

17. Fagan MJ, Chapman PM, Ell SR, Gilbert JM. Generation of data from speech or voiceless mouthed speech. Patent application P40152GB (2005).

18. ADlink PCI9118DG available at http://www.adlinktech.com . Accessed 7 July 2006.

Table 1.														
Modified Euclidean distances for a typical set of comparisons for the 13 phonemes tested														
Phoneme	Template													
	a	o	b	f	g	k	M	u	p	r	s	t	T	
	hAda	cOda	Bonda	Filea	Gota	Kita	Mothera	hOOta	Ponda	Ropea	Seea	Tuga	THicka	
Test input														
a	1	5.2	18.2	4.4	7.6	22.1	7.4	20.1	26.7	10.3	15.2	17	14.2	
o	2.9	1	28.4	6.2	5.6	19.8	11	22.2	38.5	13.6	20.7	16.3	24.4	
b	15.5	25.6	1	2.9	6.7	14.4	1.7	17.4	2	4.2	18.5	11.1	5.1	
f	13.7	22.7	5.3	1	8.1	17.2	2.6	22.8	9.2	7.3	21.7	13.1	10.7	
g	10.8	10.5	26.9	4.7	1	12.1	12.5	13.7	29.6	9	31.6	10.1	43.6	
k	7.6	12.2	7.1	2.8	2.1	1	2.4	8.3	7.3	2	6.3	1.6	12.1	
m	14	21.1	1.6	2.2	7.7	11.2	1	17	3.7	5.2	14.7	10	5.3	
u	7.8	11.6	14.4	3.6	2.8	8.5	5.2	1	15.2	5.5	16.7	5.4	14.9	
p	17.2	27.7	2.1	3.8	7.4	13.3	1.5	17.3	1	6.1	20.2	10.3	10.8	
r	10.8	15.6	8.2	3.8	3.1	4.7	2.5	9.4	10.4	1	9.8	4.2	10.8	
s	8.3	12.6	8.4	3.4	3.4	3.7	2	13.6	12.1	4.1	1	3.7	7.2	
t	9.6	12.8	12.1	3.9	2.3	2.9	3.9	13.1	13	3.1	8.5	1	18.5	
T	16.7	28.7	17.9	7.7	15.1	30.6	5.9	29	28.6	10	18.8	32.1	1	

Table 2.									
Modified Euclidean distances for a typical set of comparisons for the 9 words tested									
	Training set								
	Cat	Dog	On	Off	Up	Down	Yes	No	Bag
Test input									
Cat	1	25.5	22.1	20.6	13.6	16.6	4.8	13.9	8
Dog	15.6	1	6.1	17.6	13.6	22.9	12.3	5	16.1
On	21.3	8.4	1	14.8	16.7	27.8	11.4	6.2	20.2
n Off	5.7	5.1	5.1	1	6.1	11	4.2	4.4	5.1
Up	6.2	11.5	13.6	13.6	1	8	5.7	8.1	6.7
Down	4.1	9.1	9.4	10.8	4.4	1	3.6	5.1	3.2
Yes	2.9	23.4	18.2	23.7	16.6	20.5	1	15.5	10.7
No	9	4.4	6.3	15.9	7	8.2	6.7	1	7.3
Bag	4	10.3	13	11.8	6	5.6	3.8	6.3	1