Medical Engineering & Physics 32 (2010) 1189–1197

Contents lists available at ScienceDirect

Medical Engineering & Physics

journal homepage: www.elsevier.com/locate/medengphy

Isolated word recognition of silent speech using magnetic implants and sensors

J.M. Gilbert[a,*], S.I. Rybchenko[a], R. Hofe[b], S.R. Ell[c], M.J. Fagan[a], R.K. Moore[b], P. Green[b]

[a] Department of Engineering, University of Hull, Hull, HU6 7RX, UK
[b] Department of Computer Science, University of Sheffield, Sheffield, UK
[c] Department of Otolaryngology, Hull Royal Infirmary, Hull and East Yorkshire Hospitals NHS Trust, UK

ARTICLE INFO

Article history:
Received 10 May 2010
Received in revised form 20 August 2010
Accepted 20 August 2010

Keywords:
Laryngectomy
Speech restoration
Speech recognition
Silent speech interface

ABSTRACT

There are a number of situations where individuals wish to communicate verbally but are unable to use conventional means—so called 'silent speech'. These include speakers in noisy and covert situations as well as patients who have lost their voice as a result of a laryngectomy or similar procedure. This paper focuses on those who are unable to speak following a laryngectomy and assesses the possibility of speech recognition based on a magnetic implant/sensors system. Permanent magnets are placed on the tongue and lips and the changes in magnetic field resulting from movement during speech are monitored using a set of magnetic sensors. The sensor signals are compared to sets of pre-recorded templates using the dynamic time warping (DTW) method, and the best match is identified. Experimental trials are reported for subjects with intact larynx, typically using 500–1000 utterances used for speaker dependant training and testing. It is shown that recognition rates of over 90% are achievable for vocabularies of at least 57 isolated words: sufficient to drive command-and-control applications.

1. Introduction

Silent speech interfaces have attracted growing research interest over the past few years with a range of techniques and applications considered. The potential applications include those where the speaker is unable to generate normal speech due to disease or trauma; those where speech is inaudible, for instance due to ambient noise, and those where audible speech is undesirable such as in quiet environments or covert situations. While similar silent interfaces could be applied to all of these situations, the acceptable impact on the user differs with each scenario. Thus a person who is physiologically incapable of speech is likely to be more willing to accept an implanted device which restores their speech than a person with normal speech who wishes to communicate in a noisy environment. The work described in this paper is aimed at those who are unable to speak, in the first instance due to laryngectomy. In this situation a long-term solution is desirable and users are likely to be willing to undergo minor surgical procedures and undertake training in order to use an interface which restores speech communication.

A laryngectomy is surgical removal of the larynx, which may be required for the treatment of laryngeal carcinoma or other destructive diseases of the larynx, or following extensive trauma to the throat. The removal of the larynx inevitably results in the loss of the patient's voice and although a number of methods are available to restore speech, they all have limitations. Sound can be created by swallowing air and belching, forming the sound into words. This 'oesophageal speech' 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, sounds electronic, and can be difficult to understand. In addition, a hand held device is typically required which some users find inconvenient and socially unacceptable since it draws attention to their laryngectomised condition. The current preferred 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. These silicone valves work very well initially but rapidly become coated in a biofilm, causing them to fail after an average of only 3–4 months [2–5]. Various modifications have been tried over the years to discourage the biofilm growth (e.g. [6–8]), but to date none of these approaches appears to provide a long-term solution to the biofilm problem. The use of ceramic materials, which are resistant to biofilm growth, appears promising [9] but has not yet been tested clinically. Furthermore, the use of speech valves is not possible in all cases, particularly if there is poor healing post radiotherapy, requiring surgical procedures to close the tracheo-oesophageal fistula.

The recognition of speech in the absence of audio cues is by no means new, with lip reading being the longest established and most widely used method by which humans can understand the

[*] Corresponding author. Tel.: +44 1482 466379.
E-mail address: j.m.gilbert@hull.ac.uk (J.M. Gilbert).

1350-4533/$ – see front matter © 2010 IPEM. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.medengphy.2010.08.011

silent speech of others. Automatic visual lip reading has also been considered for machine interpretation of speech either alone [10] or to augment audio-based recognition [11]. While these systems perform acceptably in laboratory conditions, their performance can be sensitive to variations in lighting conditions and they are unlikely to be acceptable for use in public, owing to the need for a camera to image the mouth. A range of other techniques have been considered to extract signals which may be used to interpret speech. Contact between the tongue and the palate is important in many speech elements and recognition of speech using this contact information has been investigated in [12]. An electropalatograph (palatometer) consisting of 118 contacts was used, and trials involving 50 words were conducted using a range of recognition techniques, resulting in recognition rates of up to 78%. Imaging of the vocal tract, using either ultrasound [13] or radar [14], either alone or in conjunction with optical imaging provides reasonable performance but the sensors involved are not well suited to public use. Non-audible murmur (NAM) microphones, which are capable of detecting whispered speech, may be appropriate as a silent speech interface for users with intact larynx [15], but of course without a larynx there is no sound available to amplify. In this case, an additional vibration source is required in the form of an electrolarynx. The combination of electrolarynx and NAM has been shown to improve naturalness of the speech but reduce intelligibility when compared to the use of an electrolarynx alone [16]. It should also be noted that effective use of these devices requires extensive training. Similarly, measurement of glottal activity using vibration and electromagnetic sensing [17] has been investigated for speech enhancement and sub audible speech detection but, once again, this appears to be more suited to speakers with an intact larynx. Measurement of muscle stimulus signals using electromyography (EMG) techniques has been investigated for speech recognition (with an intact larynx) using small vocabularies and, generally, isolated words [18]. While recognition rates of 70–90% are reported under these circumstances, there are challenges in extending these results to larger vocabularies and continuous speech. In particular, it is noted that the EMG signal depends on the muscle size. This means that it may prove difficult to differentiate between the stimuli to some of the small muscles involved in speech and the larger, non speech-related muscles e.g. the strap muscle, particularly if the user moves during speech. Also, the needle electrodes required for EMG signal detection remain visible and are unsightly and prone to infection.

A further abstraction from acoustic recognition is to look for signals in the brain corresponding to speech or even thinking about speech. A number of approaches based on non-invasive EEG techniques or implanted electrodes are described in Ref. [19]. These techniques clearly have no requirement for functioning vocal apparatus, but so far results are only preliminary. A review and comparison of key silent speech interface technologies (including preliminary work on the system described here [20]) is presented in [21] where each method is evaluated against criteria such as cost, invasiveness, suitability for laryngectomees and market readiness.

The system described in the current paper aims to recognise speech based on measured movement of the articulators but using magnetic implants and sensors. This approach has been selected since it appears to be less obtrusive than some of the sensors described above, less invasive than others and of relatively low cost and providing a reasonable recognition rate, while being suitable for laryngectomees. The proposed approach, the signal processing and the proposed recognition algorithm are described in more detail in Section 2. The experimental trials and the key results are described in Section 3 and discussed, along with directions for future development, in Section 4.

2. System description

The concept underlying the proposed system is that multiple permanent magnets implanted into the articulators generate a varying three dimensional magnetic field during speech, and that by monitoring the vector magnetic field at a number of points around the face and head, it may be possible to identify patterns corresponding to particular elements of speech and thus recognise the intended speech in the absence of audio information. It should be noted that in our approach the aim is not to determine the Cartesian position of the implants, as is the case with some electromagnetic articulography (EMA) systems [22–24], but rather to recognise patterns arising from the nonlinear mapping between implant motion and sensed magnetic field, noting that there is also a nonlinear mapping between articulator position and generated speech.

Practical implementation of the proposed concept immediately raises a number of issues including:

- the optimum number of implants and sensors,
- the optimum locations and orientations for magnets and sensors,
- the required size of the magnets and the required sensitivity of the sensors and
- the signal processing and recognition techniques to be used.

These are addressed in the following subsections.

2.1. Implant and sensor configuration

2.1.1. General considerations

Selecting suitable magnetic implants and sensors and identifying appropriate locations involves a balance between many conflicting considerations.

From a clinical perspective, we would wish to keep the number and size of implants to a minimum, to locate them, if possible, under local anaesthetic, in soft tissue parts of the vocal apparatus that are large relative to magnet size and which heal readily. We would also wish to avoid the necessity of very precise placement and orientation of implants.

From the point of view of the appearance of the user, we would wish the sensors to be as small and discrete as possible and would prefer them to appear in some unobtrusive form such as a headset microphone or pair of spectacles.

In terms of extracting the most complete information about speech, we would want to detect motion in all of the key articulators (e.g. velum, hump and tip-of-tongue, teeth and lips). For a normal speaker, the behaviour of the larynx is also important in speech production but clearly cannot be part of a system for laryngectomees.

Finally, from an engineering perspective, it would be desirable to have a high signal-to-noise ratio, implying large magnets in close proximity to the sensors. Analysis and development of recognition systems might also be made easier if it were possible to decouple signals produced by different magnets (articulators). Such decoupling would not only improve the robustness of the subsequent recognition performance but might also allow extraction of information regarding the instantaneous state of individual articulators (such as their spatial position and orientation). This would require using a minimum number of spatially well-isolated magnets, which is actually in line with the clinical requirements.

Key articulators for speech generation are well established [25–27]; however, the choice of optimal implant location(s) and their number for any particular articulator is unclear. We have identified a limited number of the potentially most important locations for implants based on modelling and engineering/clinical judgement and followed the maximal variability criteria: maximising the