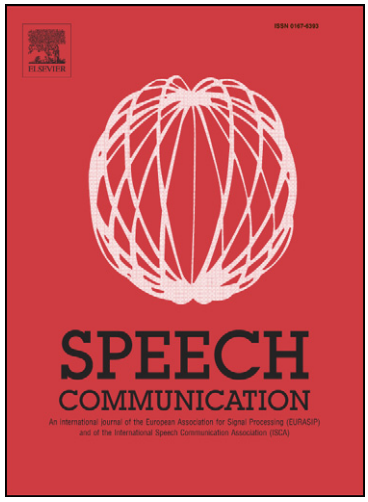
Silent Speech Interfaces

B. Denby, T. Schultz, K. Honda, Thomas Hueber, J.M. Gilbert, J.S. Brumberg



Silent Speech Interfaces

B. Denby, T. Schultz, K. Honda, T. Hueber, J.M. Gilbert, J.S. Brumberg

PII: S0167-6393(09)00130-7
DOI: 10.1016/j.specom.2009.08.002
Reference: SPECOM 1827

To appear in: *Speech Communication*

Received Date: 12 April 2009
Accepted Date: 20 August 2009

Please cite this article as: Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J.M., Brumberg, J.S., Silent Speech Interfaces, *Speech Communication* (2009), doi: 10.1016/j.specom.2009.08.002

Silent Speech Interfaces

B. Denby[1,4], T. Schultz[2], K. Honda[3], T. Hueber[4], J. M. Gilbert[5], J. S. Brumberg[6]

[1]Université Pierre et Marie Curie – Paris VI, 4 place Jussieu, 75005 Paris, France; denby@ieee.org
[2]Cognitive Systems Laboratory, Universität Karlsruhe, Am Fasanengarten 5, 76131 Karlsruhe, Germany; tanja@ira.uka.de
[3]ATR Cognitive Information Science Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan; honda@atr.jp
[4]Laboratoire d'Electronique de l'ESPCI-ParisTech, 10 rue Vauquelin, 75005 Paris, France; hueber@ieee.org
[5]Department of Engineering, University of Hull, Hull, HU6 7RX, UK; J.M.Gilbert@hull.ac.uk
[6]Department of Cognitive and Neural Systems, Boston University, 677 Beacon Street, Boston, MA, 02215 USA; brumberg@cns.bu.edu

Corresponding author: Bruce Denby
Université Pierre et Marie Curie – Paris VI
4 place Jussieu, 75005 Paris, France
telephone: +33 1 40794528
fax : +33 1 47041393
email: denby@ieee.org

Abstract. The possibility of speech processing in the absence of an intelligible acoustic signal has given rise to the idea of a '*silent speech' interface*, to be used as an aid for the speech handicapped, or as part of a communications system operating in silence-required or high-background-noise environments. The article first outlines the emergence of the silent speech interface from the fields of speech production, automatic speech processing, speech pathology research, and telecommunications privacy issues, and then follows with a presentation of demonstrator systems based on seven different types of technologies. A concluding section underlining some of the common challenges faced by silent speech interface researchers, and ideas for possible future directions, is also provided.

Keywords: silent speech; speech pathologies; cellular telephones; speech recognition; speech synthesis

I. Introduction

A silent speech interface (SSI) is a system enabling speech communication to take place when an audible acoustic signal is unavailable. By acquiring sensor data from elements of the human speech production process – from the articulators, their neural pathways, or the brain itself – an SSI produces a digital representation of speech which can be synthesized directly, interpreted as data, or routed into a communications network.

SSIs are still in the experimental stage, but a number of potential applications seem evident. Persons who have undergone a laryngectomy, or older citizens for whom speaking requires a substantial effort, would be able to mouth words rather than actually pronouncing them. Alternatively, those unable to move their articulators due to paralysis could produce speech or issue commands simply by concentrating on the words to be spoken. And because SSIs build upon the existing human speech production process, augmented with digital sensors and processing, they have the potential to be more natural-sounding, spontaneous, and intuitive to use than such currently available speech pathology solutions as the electrolarynx, tracheo-oesophageal speech (TES), and cursor-based text-to-speech systems.

While improving aids for the speech-handicapped has been an objective of biomedical engineering for many years, the recent increase of interest in SSI technology arises also from a second, quite different class of applications: providing privacy for cellular telephone conversations. It is widely agreed that cellphones can be an annoyance in meetings or quiet areas, and in many public places today their use is banned. Quite often the cellphone user, too, is uncomfortable having the content of his or her conversation become public. At the same time, the ability to field an urgent or important call at any location could in many instances be

a very useful service. An SSI, if non-invasive and small enough to be incorporated into a telephone handset, would resolve these issues by allowing users to communicate silently, without disturbing those around them. Given the numbers of cellphones in use today, the market for SSIs could potentially become very important if such a concept gained public acceptance.

Somewhat paradoxically, Silent Speech Interfaces also hold promise for speech processing in noisy environments. This is due to two principal observations:

1. Being based on non-acoustically acquired speech cues, SSIs are largely insensitive to ambient background noise;
2. In a noisy environment, vocalization is no longer restricted. Although an audible (i.e., intelligible) speech signal is not produced, the associated glottal activity creates signals which can be exploited via appropriate sensors incorporated into an SSI.

Speech communication in noisy environments is thus the third major application area of the Silent Speech Interface.

To date, experimental SSI systems based on seven different types of technology have been described in the literature:

1. Capture of the movement of fixed points on the articulators using Electromagnetic Articulography (EMA) sensors (Fagan et al. (2008);
2. Real-time characterization of the vocal tract using ultrasound (US) and optical imaging of the tongue and lips (Denby and Stone (2004); Denby et al. (2006); Hueber et al. (2007a-c); Hueber et al. (2008a-b); Hueber et al. (2009, this issue));
3. Digital transformation of signals from a Non-Audible Murmur (NAM) microphone (a type of stethoscopic microphone) (Nakajima et al. (2003a-b); Nakajima (2005);

Nakajima et al. (2006); Heracleous et al. (2007); Otani et al. (2008); Hirahara et al. (2009, this issue); Tran et al. (2008a-b); Tran et al. (2009, this issue));

4. Analysis of glottal activity using electromagnetic (Titze et al. (2000); Ng et al. (2000); Tardelli Ed. (2004); Preuss et al. (2006); Quatieri et al. (2006)), or vibration (Bos and Tack (2005); Patil and Hansen (2009, this issue)) sensors;
5. Surface electromyography (sEMG) of the articulator muscles or the larynx (Jorgensen et al. (2003); Maier-Hein et al. (2005); Jou et al. (2006); Hasegawa-Johnson (2008)); Jorgensen and Dusan (2009, this issue); Schultz and Wand (2009, this issue);
6. Interpretation of signals from electro-encephalographic (EEG) sensors (Porbadnigk et al. (2009));
7. Interpretation of signals from implants in the speech-motor cortex (Brumberg et al. (2009, this issue)).

The primary goal of this article is to provide a detailed, but concise introduction to each of these approaches. These summaries appear in section III. Our article would not be complete, however, without also outlining the historical context in which the SSI concept has evolved, starting from its roots in speech production research and biomedical engineering. That is the focus of section II, below. In the concluding section of the article, we first compare the different SSI technologies head-to-head, pointing out for each one its range of application, key advantages, potential drawbacks, and current state of development, and finally attempt to draw some general conclusions from the work carried out by the different groups, proposing possible avenues for future development in this exciting new interdisciplinary field.

II. Historical Framework

Humans are capable of producing and understanding whispered speech in quiet environments at remarkably low signal levels. Most people can also understand a few unspoken words by lip-reading, and many non-hearing individuals are quite proficient at this skill. The idea of interpreting silent speech electronically or with a computer has been around for a long time, and was popularized in the 1968 Stanley Kubrick science-fiction film "2001 - A Space Odyssey", where a "HAL 9000" computer was able to lip-read the conversations of astronauts who were plotting its destruction. Automatic visual lip-reading was initially proposed as an enhancement to speech recognition in noisy environments (Petajan 1984), and patents for lipreading equipment supposedly able to understand simple spoken commands began to be registered in the mid 1980's (Nakamura (1988)). What was perhaps the first "true" SSI system, although with very limited performance, originated in Japan. In 1985, scientists used signals from 3 electromyographic sensors mounted on the speaker's face to recognize 5 Japanese vowels with 71% accuracy, and output them to a loudspeaker in real-time (Sugie and Tsunoda (1985)). A few years later, an imaging-based system, in which lip and tongue features were extracted from video of the speaker's face, returned 91% recognition on a similar problem (Hasegawa and Ohtani (1992)).

While the possibility of robustness of silent speech devices to ambient noise was already appreciated in some of the earliest articles, the idea of also recovering glottal excitation cues from voiced speech in noisy environments was a somewhat later development. A major focal point was the DARPA Advanced Speech Encoding Program (ASE) of the early 2000's, which funded research on low bit rate speech synthesis "with acceptable intelligibility, quality, and aural speaker recognizability in acoustically harsh environments", thus spurring developments

in speech processing using a variety of mechanical and electromagnetic glottal activity sensors (Ng et al. (2000); Tardelli Ed. (2004); Preuss et al. (2006); Quatieri et al. (2006)).

It was not until the advent of cellular telephones, however, that SSIs in their current incarnation began to be discussed. Major deployment of GSM cellular telephone networks began around 1994. By 2004, there were more cellphones worldwide than fixed line phones, and the intervening years provided more than ample time for the issue of cellphone privacy to manifest itself. In Japan in 2002, an NTT DoCoMo press release announced a prototype silent cellphone using EMG and optical capture of lip movement (Fitzpatrick (2002)). “The spur to developing such a phone,” the company said, “was ridding public places of noise,” adding that, “the technology is also expected to help people who have permanently lost their voice.” The first SSI research papers explicitly mentioning cellphone privacy as a goal also began to appear around this time (Nakajima et al. (2003a); Denby and Stone (2004)).

The possibility of going further today than in some of the earlier SSI designs is due in large part to advances in instrumentation made by the speech production research community. Many of the sensing technologies proposed for use in SSIs have been developed over numerous years for extracting detailed, real time information about the human speech production process. There is thus today a wealth of resources available for applying ultrasound (Stone et al. (1983); Stone and Shawker (1986); Stone and Davis (1995); Wrench and Scobbie (2003); Stone (2005); Davidson (2005); Epstein (2005), Wrench and Scobbie (2007)), X-ray cineradiography (Arnal et al. (2000); Munhall et al. (1995)), fMRI (Gracco et al., (2005); NessAiver et al. (2006)), EMA (Perkell et al. (1992); Hoole and Nguyen (1999)), EMG (Tatham (1971); Sugie and Tsunoda (1985)), and EPG (Gibbon, F., (2005)) to speech-related research problems. Speech scientists interested in going back to the brain itself to find

exploitable SSI signals are able to profit from research experience on EEG and other BCI techniques (Epstein (1983); Wolpaw et al. (2002); IEEE (2008); Sajda et al. Eds. (2008)) as well.

The use of vocal tract imagery and other sensor information to help build speech synthesis systems, furthermore, is by no means a by-product of recent research on SSIs. It has been standard practice for many years in the fields of articulatory speech synthesis and multimodal speech processing, where, once again, the accent is on understanding speech production (Maeda (1990); Rubin and Vatikiotis-Bateson (1998); Schroeter et al. (2000); House and Granström (2002)). The goal of SSI research is less to further the understanding of the underlying speech production processes – though this is not ruled out should a breakthrough nonetheless occur – than to apply some of what has already been learned to perform new, useful functions, in particular: 1) providing speech of "acceptable intelligibility, quality, and aural speaker recognizability", as DARPA expressed it, to the speech handicapped; and 2) enabling speech processing in situations where an acoustic signal is either absent or is masked by background noise.

Finally, investigators in phonetics and speech pathologies, along with medical researchers and practitioners responsible for much of what is known about these handicaps today, and experts in biomedical engineering, have also laid much of the groundwork necessary for the development of successful SSI applications (Blom and Singer (1979); Baken et al. (1984); Marchal and Hardcastle (1993); Drummond et al., (1996); Nguyen et al. (1996) ; Crevier-Buchman (2002)).

III. Silent Speech Interface Technologies

Each of the following subsections describes a different technology which has been used to build an experimental SSI system reported in the literature. The order of presentation has been chosen to start with the "physical" techniques which characterize the vocal tract by measuring its configuration directly or by sounding it acoustically, before passing to an "electrical" domain, where articulation may be inferred from actuator muscle signals, or predicted using command signals obtained directly from the brain. An ad hoc comparison of the different methods, giving their range of application, advantages, drawbacks, and state of development, appears in section IV.

III.A Capture of the movement of fixed points on the articulators using Electromagnetic Articulography (EMA) sensors

As the shaping of the vocal tract is a vital part of speech production, a direct and attractive approach to creating a silent speech interfacing would be to monitor the movement of a set of fixed points within the vocal tract. Numerous authors have considered methods of tracking this motion using implanted coils which are electrically connected to external equipment and are electromagnetically coupled to external excitation coils (Carstens (2008); Schönle et al. (1987); Hummel et al. (2006)). These standard EMA systems aim to track the precise Cartesian coordinates, in two or three dimensions, of the implanted coils.

While it would be attractive to attempt to measure the Cartesian position of defined points in an SSI application, it is non-trivial to actually achieve in a convenient manner. However,

given that a nonlinear mapping already exists between the vocal tract shape and the resulting sounds, it appears worthwhile to consider a simpler monitoring system based not on Cartesian positions, but on some other, nonlinear, mapping. In Fagan et al. (2008), a system was investigated which consists of permanent magnets attached at a set of points in the vocal apparatus, coupled with magnetic sensors positioned around the user's head. The use of permanent magnets has the advantage that there is no necessity for an electrical connection to the implants, and so there is greater flexibility in terms of placement and use. In the test system developed, magnets were glued to the user's tongue, lips and teeth, and a set of six, dual axis magnetic sensors mounted on a pair of spectacles, as shown in Figure 1.

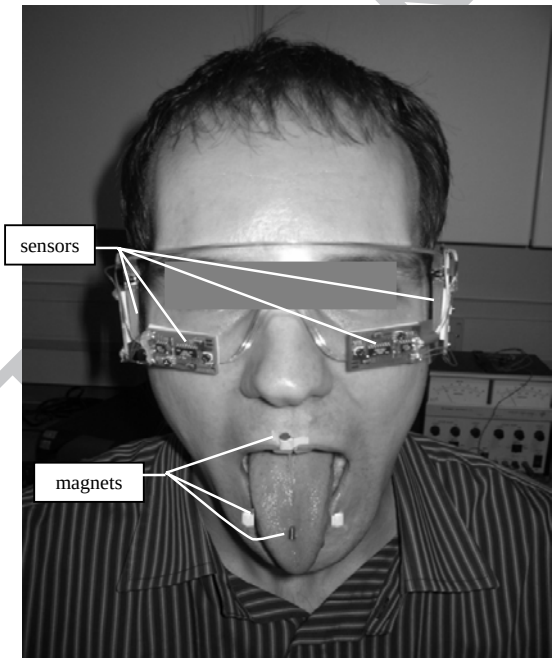


Figure 1. Placement of magnets and magnetic sensors for an EMA based SSI.

The aim of the experimental study was to establish whether it is possible to extract sufficient information from a set of sensors of this type to allow basic speech recognition. To this end, a simple recognition algorithm was adopted, based on an adaptation of the widely used

Dynamic Time Warping (DTW) algorithm, using Dynamic Programming (DP) (Holmes and Holmes (2001); Furui (2001)). In order to evaluate the behavior of the system, the subject was asked to repeat a set of 9 words and 13 phonemes (taken from the ARPAbet (Levinson (2005))) to provide training data. Ten repetitions of each word/phone were compared to the training set template. It was found that under laboratory conditions, with these very limited vocabularies, it was possible to achieve recognition rates of over 90%. It was noted that while the discrimination between, for instance, the labial phonemes (b-m-p-f) and between the velar phonemes (g-k) was less significant than for more distinct phonemes, the processing was still able to correctly identify the best fit. This was also found to be the case for voiced and unvoiced versions of the same phoneme (e.g. g-k and b-p). On the basis of these preliminary results it is believed that with further development of the sensing and processing systems it may be possible to achieve acceptable recognition for larger vocabularies in non-laboratory conditions.

III.B Real-time characterization of the vocal tract using ultrasound (US) and optical imaging of the tongue and lips

Another way to obtain direct information on the vocal tract configuration is via imaging techniques. Ultrasound imagery is a non-invasive and clinically safe procedure which makes possible the real-time visualization of one of the most important articulators of the speech production system – the tongue. Placed beneath the chin, an ultrasound transducer can provide a partial view of the tongue surface in the mid-sagittal plane. In the SSI developed in the *Ouisper* project (Ouisper (2006)), an ultrasound imaging system is coupled with a standard video camera placed in front of the speaker's lips. Non-acoustic features, derived exclusively

from visual observations of these two articulators, are used to drive a speech synthesizer, called in this case a “silent vocoder”, as illustrated in figure 2.

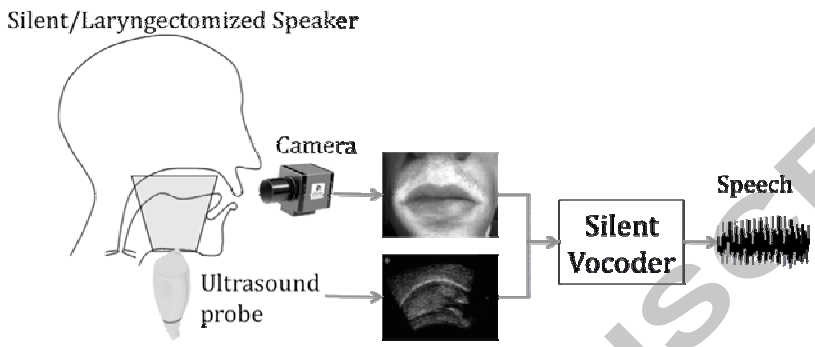


Figure 2: Ultrasound-based SSI (schematic).

Since neither glottal excitation nor airflow in the vocal tract is required, an ultrasound-based SSI is suitable for use by patients who have undergone a laryngectomy. And since laptop-based high performance ultrasound medical imaging systems are already available today, a wearable, real-time SSI, with an embedded ultrasound transducer and camera, appears to be quite realistic.

A number of solutions have been proposed and described in the literature to build a “silent vocoder” able to recover an acoustic speech signal from visual information only. In the first such attempt to achieve this “visuo-acoustic” mapping task, tongue contours and lip profiles extracted from a 2 minute ultrasound dataset were mapped either onto GSM codec parameters (Denby and Stone (2004)) or line spectral frequencies (Denby et al. (2006)) using multilayer perceptrons. In Hueber et al. (2007a), extraction and parameterization of the tongue contour was replaced by a more global coding technique called the *EigenTongues* decomposition. By projecting each ultrasound image into a representative space of “standard vocal tract configurations”, this technique encodes the maximum amount of information in the images –

predominantly tongue position, but also other structures, such as the hyoid bone and short tendon, as well as muscle and fat below the tongue. All these approaches, however, predict only spectral features, and thus permit only LPC-based speech synthesis, without any prescription for finding an appropriate excitation signal. One solution to this problem would be to make use of pre-recorded acoustic speech segments, as is done in state-of-the-art corpus-based speech synthesis systems. In that perspective, a new framework, combining a "visuo-phonetic decoding stage" and a subsequent concatenative synthesis procedure, was introduced in Hueber et al. (2007b), Hueber et al. (2008a), and Hueber et al. (2009, this issue).

The approach investigated is based on the construction of a large audio-visual unit dictionary which associates a visual realization with an acoustic one for each diphone. In the training stage, visual feature sequences are modeled for each phonetic class by a context-independent continuous Hidden Markov Model (HMM). In the test stage, the visuo-phonetic decoder "recognizes" a set of phonetic targets in the given sequence of visual features (Hueber et al. (2007c) and Hueber et al. (2009, this issue)). Evaluated on a one-hour continuous speech database, consisting of two speakers (one male, one female, native speakers of American English), this visuo-phonetic decoder is currently able to correctly predict about 60 % of phonetic target sequences, using video-only speech data. At synthesis time, given a phonetic prediction, a unit selection algorithm searches in the dictionary for the sequence of diphones that best matches the input test data, and a "reasonable" target prosodic pattern is also chosen. The speech waveform is then generated by concatenating the acoustic segments for all selected diphones, and prosodic transformations of the resulting speech signal are carried out using "Harmonic plus Noise Model" (HNM) synthesis techniques. An overview of the segmental approach to silent vocoding is given in figure 3.

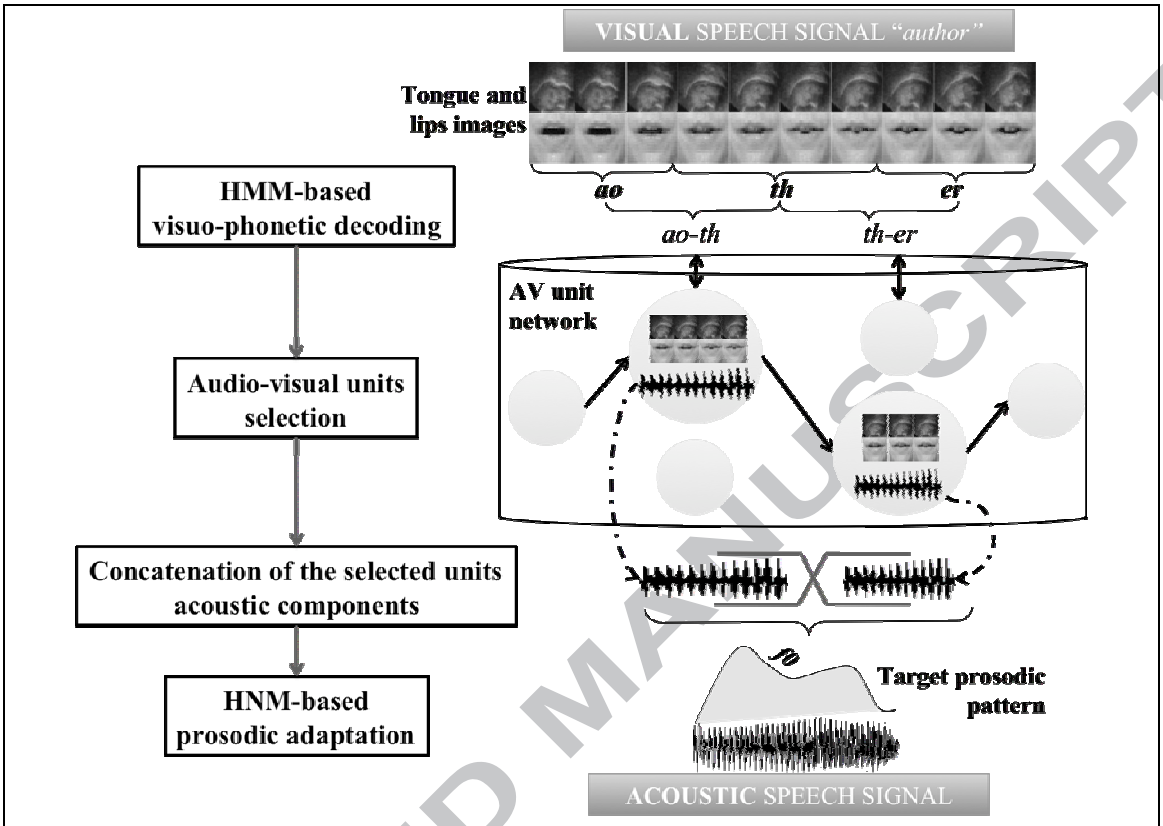


Figure 3: Overview of the segmental approach for a silent vocoder driven by video-only speech data.

With this configuration, synthesis quality depends strongly on the performance of the visual phone recognizer, and with about 60% of phones correctly identified, the system is not able to systematically provide an intelligible synthesis. Nevertheless, in those cases where the phonetic prediction is more nearly correct (above 90 %), initial intelligibility tests have shown that the system is able to synthesize an intelligible speech signal with acceptable prosody. Thus, improvement of the visuo-phonetic decoding stage remains a critical issue, and several solutions are envisioned. First, larger audiovisual speech databases are currently being recorded using a new acquisition system (Hueber et al. (2008b)), which is able to record both

video streams (US and camera), along with the acoustic signal, at more than 60 frames per second, as compared to 30 fps for the earlier baseline acquisition system. Because of the better temporal resolution, the observation, and therefore the modeling, of very short phones, such as /t/ or /d/, should now be more accurate. Also, in order to take into account the asynchrony between tongue and lip motions during speech, the use of multistream HMMs and the introduction of context-dependency are being tested. Finally, we remark that this approach to an ultrasound-based SSI has been evaluated on a difficult recognition task – the decoding of continuous speech without any vocabulary restrictions. Clearly, better performance could also be obtained either on a more limited vocabulary recognition task (less than 250 words, for instance), or on an isolated word "silent recognition" task, giving an ultrasound based SSI that could be used in more restricted, but nevertheless realistic situations.

III.C Digital transformation of signals from a Non-Audible Murmur (NAM) microphone

Non-audible murmur (NAM) is the term given to the low amplitude sounds generated by laryngeal airflow noise and its resonance in the vocal tract (Nakajima et al. (2003b); Otani et al. (2008)). NAM sound radiated from the mouth can barely be perceived by nearby listeners, but a signal is easily detected using a high-sensitivity contact microphone attached on the skin over the soft tissue in the orofacial region. The NAM microphone is designed for selective detection of tissue vibration during speech while being insensitive to environmental noise. It is thus expected to be a convenient input device for private telecommunications, noise-robust speech recognition, and communication enhancement for the vocally handicapped.

The idea of applying NAM for telecommunication purposes was first proposed by Nakajima (Nakajima et al. (2003a-b)), who discovered that NAM can be sensed by the ear alone using a stethoscope placed beneath the chin while whispering. A small stethoscope equipped with a microphone thus appeared to be a simple sensor for use in many situations where speaking aloud is not desirable. This early stethoscopic type of NAM microphone however displayed a problem of sound quality due to a sharply limited frequency range, up to only 2 kHz, which is certainly too narrow for speech transmission. Also, the structure of a stethoscope is very susceptible to noise due to friction against skin or clothing. Many improvements were made to resolve these problems via improved impedance matching between the microphone diaphragm and the skin. As shown in Fig. 4, the solution adopted was to encapsulate an entire microphone unit in a small enclosure filled with a soft silicone material, so that skin vibrations transmit to the microphone diaphragm via the layer of soft silicone without being affected by acoustic noise or external vibration. To reduce the size of the device, a miniature electret condenser microphone was used as the sensor unit, with its metal cover removed to expose the diaphragm and allow direct contact with the soft silicone (Nakajima (2005)). The best location for placing the NAM microphone was empirically determined to be on the skin below the mastoid process on a large neck muscle.

Even with these improvements, problems remain concerning the transmission characteristics. The frequency response of the silicone type NAM microphone exhibits a peak at 500-800 Hz, and a bandwidth of about 3 kHz. Tissue vibration at consonant bursts is conspicuous in amplitude in the detected signals. Spectral distortion of speech is small, but nonetheless present, due to the principle of signal detection, which differs from the natural acoustic propagation of speech sounds. Despite these problems, however, NAM microphones with

various types of construction have been applied as alternative speech input devices in a number of scenarios.

The NAM microphone has been applied to developing new systems for silent-speech telephony and recognition. For the purposes of telecommunications where privacy is required or in a high-noise environment, the microphone with an amplifier can be combined with a cellular phone headset to enable talking and listening. Simple amplification of NAM speech already produces an acceptable signal, and speech quality transformation techniques have been applied in order to produce more natural sounding speech. These successes have motivated speech engineers toward using "NAM recognition" as one of the acceptable solutions for robust speech recognition in noisy environments (Nakajima et al. (2006); Heracleous et al. (2007); Tran et al. (2008a-b); Tran et al. (2009, this issue)).

The NAM device is also useful for speakers with voice pathologies due to laryngeal disorders, who have difficulty producing voiced sounds that require vocal-fold vibration. Simple amplification of NAM speech is beneficial for the purposes of conversation, for lecturing, and for telephone calls. A more challenging task is the application of NAM as a talking aid for alaryngeal speakers. Removal of the larynx, with separation of the vocal tract from the upper airway, is clearly an unfavorable condition for the use of a NAM microphone because airflow from the lungs cannot produce the necessary vocal-tract resonance. In this case, an alternative sound source must be introduced externally to the vocal tract. Hirahara et al. (2009, this issue) employed a small vibration transducer attached on the neck surface to elicit vocal-tract resonance for this purpose, and also investigated speech transformation techniques to produce more natural sounding NAM-based speech synthesis. Adequately estimating voiced segments

for speech synthesis is of course a problem common to all vocal-tract sensing approaches to the SSI problem. We will return to this point in section IV.A.

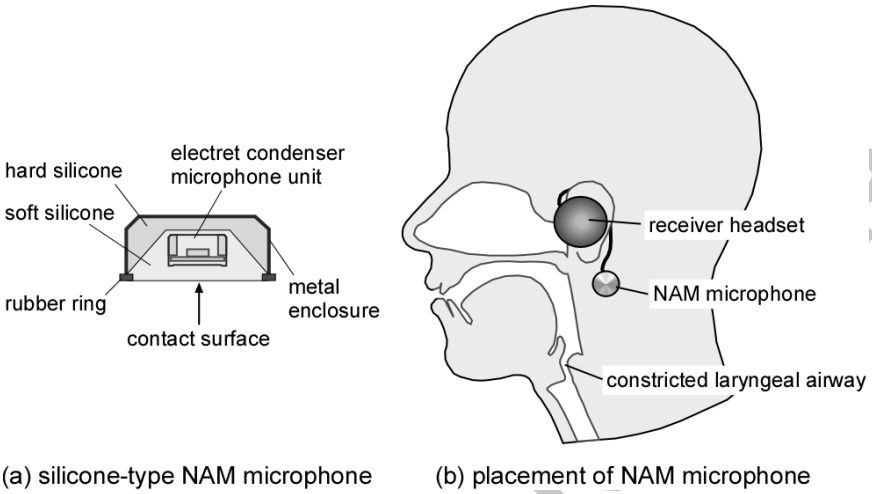


Figure 4. NAM microphone and its placement. (a) Typical silicone-type NAM microphone using a miniature electret condenser microphone unit that is embedded in a capsule filled with a soft silicone material. (b) Setup of NAM microphone for silent mobile telephony. The microphone is placed on the side of the neck to capture tissue-conducted vibration of vocal-tract resonance generated by airflow noise in the constricted laryngeal airway.

III.D Analysis of glottal activity using electromagnetic or vibration sensors

In the early 2000's, the United States Defense Department launched the DARPA Advanced Speech Encoding Program (ASE), which provided funding to develop non-acoustic sensors for low bit rate speech encoding in challenging acoustic environments, such as the interiors of fighting vehicles and aircraft, urban military terrains, etc. A number of projects were funded in Phase I of the program, usually to evaluate the efficacy of specific sensors used in conjunction with a standard close-talk microphone, and a special data base entitled "DARPA Advanced Speech Encoding Pilot Speech Corpus" was developed in order to benchmark the different proposed solutions (Tardelli Ed. (2004)). A program with similar goals was

undertaken by Defense Research and Development Canada (Bos and Tack (2005)), and in Europe, the EU project SAFIR (Speech Automatic Friendly Interface Research, IST-2002-507427) (Dekens et al. (2008)) contained work packages for developing a speech database with six types of hostile acoustic environments in order to evaluate non-acoustic sensors.

The basic principle in these studies is to obtain glottal waveforms which can be used for de-noising by correlation with the acoustic signal obtained from a standard close-talk microphone. The necessary waveforms may be obtained either via detectors which are directly sensitive to vibrations transmitted through tissue – throat microphones and the like – or from the interaction of glottal movement with an imposed electromagnetic field. In addition to the glottal closure information, spectrograms of the sensor signals in most cases exhibit vocal tract resonances as well, albeit in a modified form because of the way in which the information is captured. For vibration sensors, for example, the observed spectrum is modified by the acoustic transfer function of the path between the sound source and the capture device. The “non-acoustic” label is in fact a bit of a misnomer for such sensors; it refers to the fact that the captured acoustic signal, as was the case with NAM, propagates through tissue or bone, rather than through air.

Speech enhancement using such vibration or electromagnetic sensors has been shown to work very well, with gains of up to 20 dB reported ((Titze et al. (2000); Dupont and Ris (2004); Quatieri et al. (2006)). In Ng et al. (2000), perfectly intelligible speech was obtained using a GEMS device (described below) from a signal with an initial signal to noise ratio of only 3 dB, while excellent results on noise robust vocoding in three harsh military noise environments using GEMS and PMIC (description below) are reported in Preuss et al. (2006). Some of the devices studied in ASE and elsewhere are described in more detail in what

follows. The related area of speech enhancement in noise using audiovisual cues has been well covered elsewhere, and will not be addressed in this article.

Vibration sensors:

Throat Microphone. Throat microphones have been used by fighter pilots for decades. Based on the same technologies as standard microphones, these are usually double units attached to a neckband, with one bud worn on each side of the Adam's apple. The buds are simply pressed against the skin, without any special coupling, in order to capture that part of the speech signal which is transmitted through the flesh of the throat. Throat microphones are designed to have low response in the outward direction, so as to remain insensitive to background noise.

Bone microphone. Bone microphones are designed to pick up the speech signal which propagate through the bones of the skull, and as such are also resistant to contamination from ambient background noise. The best capture points are on the cheekbone, forehead or crown of the head. Bone microphones are sometimes incorporated directly into soldier's helmets for battlefield applications (see, for example, (Bos and Tack (2005)).

Physiological microphone, or PMIC. The PMIC (Bos and Tack (2005), Quatieri et al. (2006); Preuss et al. (2006); Patil and Hansen (2009, this issue)), is a more sophisticated form of the throat microphone. It is worn as a neckband, consisting of a piezoelectric sensor immersed in a gel bath inside a closed bladder, which is designed to have a stronger acoustic coupling to tissue than to air, thus assuring robustness to background noise. The PMIC, in addition to speech detection, is also intended to relay information on the wearer's heart and

respiration rates, etc., whence its name, “physiological”. A recent research article using the PMIC for speaker assessment appears in Patil and Hansen (2009, this issue).

In-ear microphone. In this device, noise-immunity is assure by inserting a small microphone into the ear canal (Bos and Tack (2005); (Dekens et al. (2008)), which is closed off by an earphone for incoming signals. The signal to noise ratio is very good, and even whispered speech can be detected. A disadvantage of the technique is that potentially important exterior sounds can no longer be detected by the instrumented ear.

Electromagnetic sensors:

EGG. The electroglottograph (Rothenberg (1992); Titze (2000); Quatieri (2006)), is a standard research tool designed to detect changes in electrical impedance across the throat during voiced speech. It consists of 2 gold-plated electrodes held in place on either side of the larynx by means of a collar, with an applied potential. When the vocal folds are closed, the electric impedance decreases, while when they are open, a larger value ensues. Glottal vibration in this way induces a signal of some 1 volt RMS on a 2-3 MHz carrier, which is quite readily detectable. A drawback of the technique is its sensitivity to the exact positioning of the electrodes.

GEMS. The General Electromagnetic Motion System is based on a relatively recent miniature micropower (< 1 mW) radar technology (Burnett (1997); (Titze (2000)), which can effectuate very high resolution reflectrometry using brief EM pulses in the 2.4 GHz ISM band. The GEMS antenna may be attached to the throat at the laryngeal notch, or in other positions. Articulator motion, particularly that of the glottis, can be accurately detected from the

Doppler frequency shifting of the reflected electromagnetic energy that such movement engenders. Vibrations from 20 Hz to 8 kHz are detectable. As with the EGG, antenna positioning is a crucial factor for GEMS.

TERC. The Tuned Electromagnetic Resonating Collar (Brown et al. (2004); Brown et al. (2005); TERC (2009)) measures changes in the intrinsic electrical capacitance of the glottis. The device exploits the fact that when the glottis opens, the permittivity of a cross section of the neck through the larynx decreases. The device consists of a neckband composed of copper electrodes on an acrylic substrate, tuned to a sharp resonance at a particular frequency of several tens of MHz. The high-Q of the resonant circuit causes small movements of the glottis to lead to large deviations (~30 dB) from resonance, hence producing a readily detectable glottal vibration signal.

The goal of the ASE and other programs was to use non-acoustic sensors to enhance speech produced in acoustically challenging environments, for subsequent retransmission at low bit rates over limited-resource channels. As such, these projects share the SSI goal of enabling speech processing in noisy environments. Indeed, some of the sensors discussed exploit principles already evoked in our earlier discussion of the NAM microphone. These military/security applications, however, lack some of the “low acoustic profile” quality which is central to the SSI concept. Interestingly, when Phase 2 of ASE was launched in 2005 (only one bidder, BBN Corporation, was retained in Phase 2), a new, supplementary goal had appeared: to “explore and characterize the nature of sub-auditory (non-acoustic) speech and its potential utility as an alternative means of communication in acoustically harsh environments”. To the authors’ knowledge, results on sub-auditory processing in ASE have yet to be published.

III.E Surface electromyography (sEMG) based Speech Recognition

Surface ElectroMyoGraphy (sEMG) is the process of recording electrical muscle activity captured by surface (i.e., non-implanted) electrodes. When a muscle fiber is activated by the central nervous system, small electrical currents in the form of ion flows are generated. These electrical currents move through the body tissue, whose resistance creates potential differences which can be measured between different regions on the body surface, for example on the skin. Amplified electrical signals obtained from measuring these voltages over time can be fed into electronic devices for further processing.

As speech is produced by the activity of human articulatory muscles, the resulting myoelectric signal patterns measured at these muscles provides a means of recovering the speech corresponding to it. Since sEMG relies on muscle activity alone, speech can be recognized even if produced silently, i.e., without any vocal effort, and the signal furthermore cannot be corrupted or masked by ambient noise transmitted through air. As a result, sEMG-based speech recognition overcomes the major shortcomings of traditional speech recognition, namely preserving privacy of (silently) spoken conversations in public places, avoiding the disturbance of bystanders, and ensuring robust speech signal transmission in adverse environmental conditions.

The use of EMG for speech recognition dates back to the mid 1980's, when Sugie and Tsunoda in Japan, and Morse and colleagues in the United States published (almost simultaneously) their first studies. As mentioned in section II, Sugie and Tsunoda (1985) used three surface electrodes to discriminate Japanese vowels, and demonstrated a pilot system

which performed this task in real-time. Morse and O'Brien (1986) examined speech information from neck and head muscle activity to discriminate two spoken words, and in the following years, extended their approach to the recognition of ten words spoken in isolation (Morse et al. (1989); Morse et al. (1991)). Although initial results were promising, with accuracy rates of 70% on a ten word vocabulary, performance decreased dramatically for slightly larger vocabularies, achieving only 35% for 17 words, and thus did not compare favorably with conventional speech recognition standards. More competitive performance was first reported by Chan et al. (2001), who achieved an average word accuracy of 93% on a vocabulary of the English digits. Chan was also the first to combine an EMG-based recognizer with a conventional system, achieving a significant improvement in the presence of ambient noise (Chan (2003)). In Jorgensen et al. (2003), the authors proved the applicability of myoelectric signals for non-audible speech recognition, reporting 92% word accuracy on a set of six control words.

Recent research studies aim to overcome the major limitations of today's sEMG-based speech recognition systems and applications, to, for example:

- remove the restriction of words or commands spoken in isolation and evolve toward a less limited, more user-friendly continuous speaking style (Maier-Hein et al. (2005));
- allow for acoustic units smaller than words or phrases, enabling large vocabulary recognition systems (Walliczek et al. (2006); Schultz and Wand (2009, this issue));
- implement alternative modeling schemes such as articulatory phonetic features to enhance phoneme models (Jou et al. (2006); Schultz and Wand (2009, this issue));
- study the effects of electrode re-positioning (Maier-Hein et al. (2005)) and more robust signal preprocessing (Jorgensen and Binsted (2005); Jou et al. (2005));

- examine the impact of speaker dependencies on the myoelectric signal (Wand and Schultz (2009));
- investigate real-life applicability, by augmenting conventional speech recognition systems (Chan et al. (2001); Chan (2003)), and addressing size, attachment, and mobility of the capturing devices (Manabe et al. (2003), Manabe and Zhang (2004)).

Figure 5: Demonstration of the Silent Speech Recognizer developed at Carnegie Mellon and University of Karlsruhe (Maier-Hein et al. (2005)).

Most of the early studies furthermore reported results on a small number of words spoken in isolation (an example of an experimental setup appears in figure 5) (Chan et al. (2001); Jorgensen et al. (2003); Maier-Hein et al. (2005)), whereas recent work has shown, for example, that larger vocabularies of 100 words can be recognized with a word accuracy of around 70% in a single speaker setup (Jou et al. (2006). The training of reliable acoustic models for a larger vocabulary of course requires breaking words into sequences of sub-word units, such as syllables, phonemes, or even context dependent model units. Jorgensen and Binsted (2005) applied phonemes as units for vowel and consonant classification, and Walliczek et al. (2006) compared a variety of units on a 100-word vocabulary in continuously spoken speech. A successful application of articulatory features to augment phoneme based units was presented by Jou et al. (2007), and Schultz and Wand (2009, this issue) describe the

training of context dependent phonetic feature bundles, which further improved recognition performance on the same 100-word vocabulary, with up to 90% word accuracy in a speaker dependent setup. Finally, Wand and Schultz (2009) have presented initial experiments on speaker independent and speaker adaptive sEMG-based speech recognition, based on a large collection of EMG data recorded from 78 speakers reading sentences in both audible and silent speaking mode, in a collaboration between Carnegie Mellon and Pittsburgh University.

The applicability of EMG-based speech recognition in acoustically harsh environments, such as first responder tasks where sirens, engines, and firefighters breathing apparatus may interfere with reliable communication, has been investigated at NASA. For example, Jorgensen and colleagues (Betts et al. (2006)) achieved 74% accuracy on a 15-word classification task, in a real-time system which was applied to subjects exposed to a 95 dB noise level.

There has also been interesting work at Ambient Corporation in the United States, who report the development of a system whose inputs are surface EMG signals from one or more electrodes placed above the larynx. To operate the system, the user issues commands by speaking silently to himself, without opening the lips or uttering any sound. Increased activation in the laryngeal muscles is then detected by the system, and classified using a speaker-dependent HMM-based speech recognizer. A prototype deployed in late 2007 demonstrated a vocabulary of four to six directional commands, and could be used to steer a motorized wheelchair Hasegawa-Johnson (2008)).

Current EMG recording systems still lack practicability and user-friendliness. For example, the surface electrodes need to be firmly attached to the skin for the duration of the recording.

Manabe and colleagues have addressed these issues by developing ring-shaped electrodes wrapped around the thumb and two fingers (Manabe et al. (2003); Manabe and Zhang (2004)). To capture the EMG signals from facial muscles, the fingers are pressed against the face in a particular manner. It should be possible to perfect such a system for a mobile interface that can be used in both silent and noisy environments.

Electromyography thus captures electrical stimuli from the articulator muscles or the larynx, which can subsequently be exploitated in speech processing applications. One may also imagine, however, capturing viable speech biosignals directly from the brain, using electroencephalography (EEG) or implanted cortical electrodes. These possibilities are discussed in the following two sections. Although considerably further off in terms commercial application, these Brain Computer Interface (BCI) approaches – very much in vogue today – are fascinating, and hold enormous promise for speech, as well as for other types of applications.

III.F Interpretation of signals from electro-encephalographic (EEG) sensors

In addition to its well-known clinical applications, electroencephalography has also recently proven to be useful for a multitude of new methods of communication. EEG-based BCIs have consequently become an increasingly active field of research. Good overviews can be found in Dornhege et al. Eds. (2007) and in Wolpaw et al. (2002), while Lotte et al. (2007) provides a review of classification algorithms. Examples of some current BCIs include the “Thought Translation Device” (Birbaumer (2000)) and the “Berlin Brain Computer Interface” (Blankertz et al. (2006)). The aim of a BCI is to translate the thoughts or intentions of a subject into a control signal suitable for operating devices such as computers, wheelchairs or

prostheses. Suppes et al. (1997) were the first to show that isolated words can be recognized based on EEG and MEG (magnetoencephalography) recordings. Using a BCI usually requires the users to explicitly manipulate their brain activity, which is then transformed into a control signal for the device (Nijholt et al. (2008)). This typically involves a learning process which may last several months, as described, for example, in (Neuper et al. (2003)).

In order to circumvent this time consuming learning process, as well as develop a more intuitive communications interface based on silent speech, Wester and Schultz (2006) investigated a new approach which directly recognizes “unspoken speech” in brain activity measured by EEG signals (see figure 6). “Unspoken speech” here refers to the process in which a user imagines speaking a given word without actually producing any sound, indeed without performing any movement of the articulatory muscles at all. Such a method should be applicable in situations where silent speech input is preferable – telecommunications and the like – as well as for persons unable to speak because of physical disabilities, such as locked-in syndrome, and who consequently have very limited options for communicating with their environment. During the study, 16 channel EEG data were recorded using the International 10-20 system; results indicated that the motor cortex, Broca’s and Wernicke’s areas were the most relevant EEG recording regions for the task. The system was able to recognize unspoken speech from EEG signals at a promising recognition rate – giving word error rates on average 4 to 5 times higher than chance on vocabularies of up to ten words. In a followup study, Porbadnigk et al. (2009) discovered that temporally correlated brain activities tend to superimpose the signal of interest, and that cross-session training (within subjects) yields recognition rates only at chance level. These analyses also suggested several improvements for future investigations:

- using a vocabulary of words with semantic meaning to improve recognition results;

- increasing the number of repetitions of each word (20 were used in the study) and normalizing phrase lengths, in order to improve the model training;
- providing the subject with feedback on whether words were correctly recognized. Birbaumer (2000) showed that subjects can be trained to modify their brain waves when using an EEG-based BCI; subjects may thus be able to adapt their brain waves to enable words to be recognized more easily.

Figure 6: EEG-based recognition system for unspoken speech (Wester and Schultz (2006))

In another study, DaSalla and colleagues (DaSalla et al. (2009)) proposed a control scheme for a silent speech BCI using neural activities associated with vowel speech imagery. They recorded EEG signals in three healthy subjects performing three tasks: unspoken speech of the English vowels /a/ and /u/; and a no-action state as a control. Subjects performed 50 trials for each task, with each trial containing two seconds of task-specific activity. To discriminate between tasks, the authors designed spatial filters using the common spatial patterns (CSP) method. Taking 20 randomly selected trials from each of two tasks, the EEG time series data were decomposed into spatial patterns which were both common between, and optimally discriminative for, the two tasks. Applying these spatial filters to new EEG data produced new times series optimized for classification. Since the CSP method is limited to two-class discriminations, spatial filters for all pair-wise combinations of the three tasks were designed.

Resultant spatial patterns showed mostly symmetrical activations centered at the motor cortex region, specifically the Cz and Fz positions in the International 10-20 system. After spatially filtering the EEG data, the authors trained a nonlinear support vector machine using the previously selected 20 trials per task, and classified the remaining 30 trials per task. This randomized training and testing procedure was repeated 20 times to achieve a 20-fold cross validation. Accuracies and standard deviations (in %) obtained for the three subjects were 78±5, 68±7 and 68±12. The study thus shows that motor cortex activations associated with imaginary vowel speech can be classified, with accuracies significantly above chance, using CSP and a nonlinear classifier. The authors envision the proposed system providing a natural and intuitive control method for EEG-based silent speech interfaces.

III.G Interpretation of signals from implants in the speech-motor cortex

The SSIs discussed thus far have been based on relatively non-invasive sensing techniques such as US, EMG and EEG, and others. Attempts have also recently been made to utilize intracortical microelectrode technology and neural decoding techniques to build an SSI which can restore speech communication to paralyzed individuals (Kennedy (2006); Brumberg et al. (2007); Brumberg et al. (2008); Guenther et al. (2008); Bartels et al. (2008)), or to restore written communication through development of mouse cursor control BCIs for use with virtual keyboards (Kennedy et al. (2000)) and other augmentative and alternative communication (AAC) devices (Hochberg et al. (2008)). A number of factors must be considered for an intracortical microelectrode SSI, though two stand out as the most important: choice of electrode and decoding modality.

Electrode choice

A successful intracortical SSI requires electrodes capable of chronic human implantation. These electrodes must be durable and provide consistent observations of neural signals. Early on in neurophysiological research, intracortical electrodes were simply not designed for long term use in a behaving animal. However, recent advances have yielded designs which have been used in human subjects and are capable of recording from dozens of isolated neurons over many years (Kennedy and Bakay (1998); Kennedy et al. (2000); Hochberg et al. (2006); Hochberg et al. (2008); Bartels et al. (2008)). Two electrode designs in particular have been implanted in human subjects for the purpose of brain computer interfacing: the Utah microelectrode array (Maynard et al. (1997); Hochberg et al. (2006)), and the Neurotrophic Electrode (Kennedy (1989); Bartels et al. (2008)).

The Utah array consists of a single silicon wafer with many (commonly ~96) recording electrodes and is implanted on the surface of the cerebral cortex. The recording electrodes penetrate the cortical surface and sample from neurons in close proximity to the recording tips. The Neurotrophic Electrode differs in fundamental design from the Utah array. Rather than utilizing many recording tips, the Neurotrophic Electrode utilizes few, low impedance wires, encased in a glass cone filled with a neurotrophic growth factor. The entire assembly is implanted into the cortex, as with the Utah array, but the growth factor then encourages nearby cells to send neurite projections (i.e. axons and dendrites) to the implanted cone. The result is that the neurites are “captured” by the glass cone, ensuring that the recording wires are recording from a viable neural source. Today, both types electrodes have been used in human volunteers with severe paralysis, and have remained operational for many years.

Decoding modality

The decoding modality is critically important for neural prosthesis development in general, but it is possible (as illustrated in the following) that many modalities can be possible for SSIs. Modality in this context refers to the nature of the signal decoded or interpreted from observed neural activity. A common decoding modality in neural prosthesis design is arm and hand kinematics. For instance, an electrode can be implanted in the hand area of a monkey or human motor cortex, and a device can be constructed to interpret the neural activity as related to the subject's movements or intended movements. In humans, this particular decoding modality has been used to provide mouse pointer control for BCIs in paralyzed individuals (Kennedy et al. (2000); Hochberg et al. (2006); Kim et al. (2007); Truccolo et al. (2008)). For an SSI, a hand kinematic decoding modality may only be used for communication via BCI pointer control, for example by choosing letters on a virtual keyboard, selecting words on a graphical interface or utilizing graphical Augmentative and Alternative Communication (AAC) devices with mouse pointer interfaces in general.

Though neural decoding of hand kinematics is grounded in decades of neurophysiological research (e.g., Georgopoulos et al. (1982)), the modality is not natural for speech production. Given this limitation, recent research has been aimed at decoding or predicting characteristics of speech directly from cortical areas mediating speech production (Kennedy (2006); Brumberg et al. (2007); Brumberg et al. (2008); Guenther et al. (2008)). Specifically, these investigations studied the relationship between neural activity in the speech motor cortex and production of discrete speech sound segments (i.e., phonemes) and treated speech production as a complex motor task rather than an abstract language problem.

In this speech motor control approach, intracortical electrodes are implanted into the speech motor cortex rather than the hand area. Early attempts focused primarily on the discrete prediction of individual phonemes based on the ensemble activity of a population of neural units (Miller et al. (2007); Wright et al. (2007)). More recent work has placed the problem of speech motor decoding within a framework analogous to arm and hand kinematics decoding. Within this framework, the most straightforward decoding modality for a speech motor cortical implant is vocal tract, or speech articulatory (i.e., jaw, lips, tongue, etc.) kinematics. A nearly equivalent alternative modality to vocal tract kinematics for speech decoding is an acoustic representation of sounds produced during the act of speaking. In particular, formant frequencies (the resonant frequencies of the vocal tract) are inherently linked to the movements of the speech articulators and provide an acoustic alternative to motor kinematics that has already been incorporated into speech neural prosthesis designs (Brumberg et al. (2007); Brumberg et al. (2008); Guenther et al. (2008)). Both speech articulatory and acoustic modalities are appropriate for decoding intended speech from an intracortical microelectrode implanted in the speech motor cortex; therefore, they are well suited for invasive silent speech interfaces.

IV. Conclusions and Perspectives

Seven current candidate SSI technologies have now been introduced. Further details on the majority of the methods, as well as some of recent results, may be found in the accompanying articles of this special issue. In this final section of the present article, we attempt to draw some overall conclusions on the current SSI situation. We begin with an outline of common challenges faced by SSI researchers, as a consequence of the unique nature of the SSI problem, which sets it apart from traditional speech processing systems. A second subsection

then makes a qualitative comparison of the different methods by highlighting the relative benefits and drawbacks of each approach according to a set of simple criteria. Finally, we propose some possible directions for future exploration, which may hopefully lead to useful new products and services emerging from the nascent, interdisciplinary field of SSIs.

IV.A Common challenges

Sensor positioning and robustness – In all of the technologies presented, the sensors used must be carefully positioned to obtain the best response. In a system sensitive to the orientation of the tongue surface in an ultrasound image, for instance, any movement of the probe consists of a change of image reference frame, which has to be taken into account. EMA, EMG, and EEG are also sensitive to variations in sensor positioning, and researchers using NAM and EM/vibration technology have reported the need to find "sweet spots" in which to place their devices for best results. A corollary to these observations is that unavoidable changes in sensor position or orientation introduced at the beginning of each new acquisition can give session-dependent results once the subsequent signal processing algorithms are applied. Systematic ways of ensuring optimal, repeatable sensor positioning have not yet been adequately addressed in the experimental SSI systems presented to date. Further research will be necessary to find ways of ensuring that the SSI sensors used remain attuned to the relevant articulatory information in robust ways.

Speaker independence – A related concern is speaker independence. While audio-based speech recognition has made excellent progress in speaker independence, the situation may be quite different when the features which feed the recognition system depend, for example, on the speaker's anatomy, or the exact synaptic coding inherent in movements of his or her

articulatory muscles. Most of the systems we have presented have only just begun to assess speaker independence. The extent to which it will influence the advancement of SSI development is thus as yet not known.

Lombard and silent speech effects – Speakers are known to articulate differently when deprived of auditory feedback of their own speech, for example in high-noise environments – the so-called Lombard effect. Lombard speech will thus be an issue for SSIs unless they are able to provide a high-quality, synchronous audio signal via an earphone, which is of course a very challenging task. Beyond the Lombard effect resides the additional question of whether speakers articulate differently when speaking silently, either in quiet or in noisy environments, and most indications are that silent and vocalised articulation are indeed not identical. In any case, the best practice would no doubt be to train SSI systems on silent speech, rather than audible speech, since this is the context in which they will ultimately operate. To do so is experimentally much more difficult, however, since the absence of an audio stream precludes using standard ASR tools for labelling and segmenting SSI sensor data, not to mention hindering the development of an output speech synthesizer for the SSI. Although some SSI researchers have already begun to address these issues, substantial further research, again, will be required in order to discover what the actual stumbling blocks will be here.

Prosody and nasality – For SSI applications in which a realistic output synthesis is envisaged, the production of a viable prosody is a critical issue, since the glottal signal necessary for pitch estimation is either completely absent or substantially modified. When a recognition step precedes synthesis, lexical and syntactical cues could in principal be used to alleviate this problem to a certain extent, but the task is quite difficult. The problem is similar

to that encountered by electrolarynx users, who are forced to contend with a constant, monotonous F0 value. Some of these products today provide an external thumbwheel for variable pitch control, and perhaps a similar solution could prove appropriate for SSIs. In addition to prosody, depending on the SSI technology used, information on nasality may also be absent. The possibility of recovering prosody and nasality, using context or artificial compensatory mechanisms, will be an additional topic for future research.

Dictionaries – Continuous speech ASR is a difficult task, particularly in real time interactive and portable systems. It seems likely that the first useful SSI applications will concentrate on the more easily realizable goal of limited vocabulary speech recognition. A common challenge for all of the potential technologies will then be the creation of dictionaries which are of limited size, but rich enough to be genuinely useful for the SSIs tasks and scenarios for which they are tailored, e.g., telephony, post-laryngectomy speech aids, verbal command recognition, and the like.

IV.B Comparison of the Technologies

It is difficult to compare SSI technologies directly in a meaningful way. Since many of the systems are still preliminary, it would not make sense, for example, to compare speech recognition scores or synthesis quality at this stage. With a few abstractions, however, it is possible to shed light on the range of applicability and the potential for future commercialization of the different methods. To carry out our analysis, we have chosen to "fast forward" to a situation in which all of technologies are "working". To be classified as such, an SSI should be able to genuinely enable useful silent speech processing tasks – essentially recognition and synthesis – as well as present a relatively portable, human-oriented

form factor. Clearly, we have left out the possibility that one or another of the technologies ultimately fails to meet these criteria, for example if the chosen sensors simply do not provide enough information about the speech production process to enable useful subsequent processing; but that need not concern us here. For the purposes of our assessment, we have defined a set of 6 criteria, ranked on a scale of 1 (worst) to 5 (best), as defined below:

- **Works in silence** – Can the device be operated silently?
- **Works in noise** – Is the operation of the device affected by background noise?
- **Works for laryngectomy** – Can the device be used by post-laryngectomy patients? It may be useful for other pathologies as well, but laryngectomy is used as a baseline.
- **Non-invasive** – Can the device be used in a natural fashion, without uncomfortable or unsightly wires, electrodes, etc.?
- **Ready for market** – Is the device close to being marketed commercially? This axis also takes into the account in a natural way the current technological advancement of the technique, responding, in essence, to the question, "How well is this technology working as of today?".
- **Low cost** – Can the final product be low cost? The answer will depend, among other factors, on whether any "exotic" technologies or procedures are required to make the device function.

The comparisons we make will clearly be qualitative, and should not be considered as being in any sense exact. In the next paragraphs, we first rank each of the seven technologies with a grade from 1 to 5 in each of the 6 categories, giving brief explanations for the marks given. This ranking is summarized in figure 7 on six-axis "spiderweb" plots of the different technologies. For the purposes of the comparison, we have adopted shortened labels for the seven technologies, for convenience: EMA markers; US/imaging; NAM; EM/vibration; EMG electrodes, EEG electrodes; and BCI cortical.

EMA markers

Works in silence: 5 – Silent articulation is possible.

Works in noise: 5 – Background noise does not affect the operation.

Works for laryngectomy: 5 – No glottal excitation is necessary.

Non-invasive: 2 – Magnetic beads need to be fixed permanently on the tongue and other articulators.

Ready for market: 2 – Published recognition results to date are promising but still preliminary.

Low cost: 4 – The magnetic beads, detectors, and associated electronics can probably be manufactured very cheaply.

US/imaging

Works in silence: 5 – Silent articulation is possible.

Works in noise: 5 – Background noise does not affect the operation.

Works for laryngectomy: 5 – No glottal activity is required.

Non-invasive: 4 – Miniaturisation of ultrasound and camera and gel-free coupling should eventually lead to a relatively portable and unobtrusive device.

Ready for market: 3 – Recognition results suggest a useful, limited vocabulary device should not be far off, but instrumental developments are still necessary.

Low cost: 3 – Although costs much below those of medical ultrasound devices should eventually be possible, ultrasound remains a non-trivial technology.

NAM

Works in silence: 4 – The device is nearly, but not totally silent, and could be inappropriate for the most demanding non-eavesdropping scenarios.

Works in noise: 4 – Researchers have reported problems with noise caused by clothing, hair, respiration, etc.

Works for laryngectomy: 2 – The device requires an external vibrator in order to work for laryngectomees; results on this so far seem preliminary.

Non-invasive: 4 – The device resembles 2 large buttons held behind the ears with a headband or collar.

Ready for market: 5 – Commercial systems are already available in Japan.

Low cost: 5 – The devices can be mass produced very cheaply.

EM/vibration

Works in silence: 1 – Glottal activity is required.

Works in noise: 5 – The devices are designed to work well in noisy environments.

Works for laryngectomy: 1 – Glottal activity is required.

Non-invasive: 4 – The devices are relatively small and unobtrusive, often resembling a neckband.

Ready for market: 4 – Some of these devices are already commercially available.

Low cost: 3 – Some of the technologies, such as GEMS and pulse radar, are not completely trivial.

EMG electrodes

Works in silence: 5 – Silent articulation is possible.

Works in noise: 5 – Background noise does not affect the operation.

Works for laryngectomy: 5 – No glottal activity is required.

Non-invasive: 4 – A facemask-like implementation should eventually be possible, thus eliminating unsightly glued electrodes.

Ready for market: 3 – EMG sensors and their associated electronics are already widely available.

Low cost: 4 – The sensors and the data processing system are relatively manageable.

EEG electrodes

Works in silence: 5 – Silent articulation is possible.

Works in noise: 5 – Background noise does not affect the operation.

Works for laryngectomy: 5 – No glottal activity is required.

Non-invasive: 3 – Today's systems require a skull cap and conductive gel under each electrode. Ultimately, an articulated, gel-free helmet such as those proposed for some video games may be possible.

Ready for market: 1 – Tests to date are promising, but still very preliminary.

Low cost: 4 – The electrodes and signal processing electronics are relatively standard; commercial EEG systems (although not for speech) exist today.

BCI cortical

Works in silence: 5 – Silent articulation is possible.

Works in noise: 5 – Background noise does not affect the operation.

Works for laryngectomy: 5 – No glottal activity is required.

Non-invasive: 1 – Cortical electrodes must be implanted.

Ready for market: 1 – The results to date are interesting but quite preliminary.

Low cost: 1 – Brain surgery is required to implant the electrodes in the cortex.

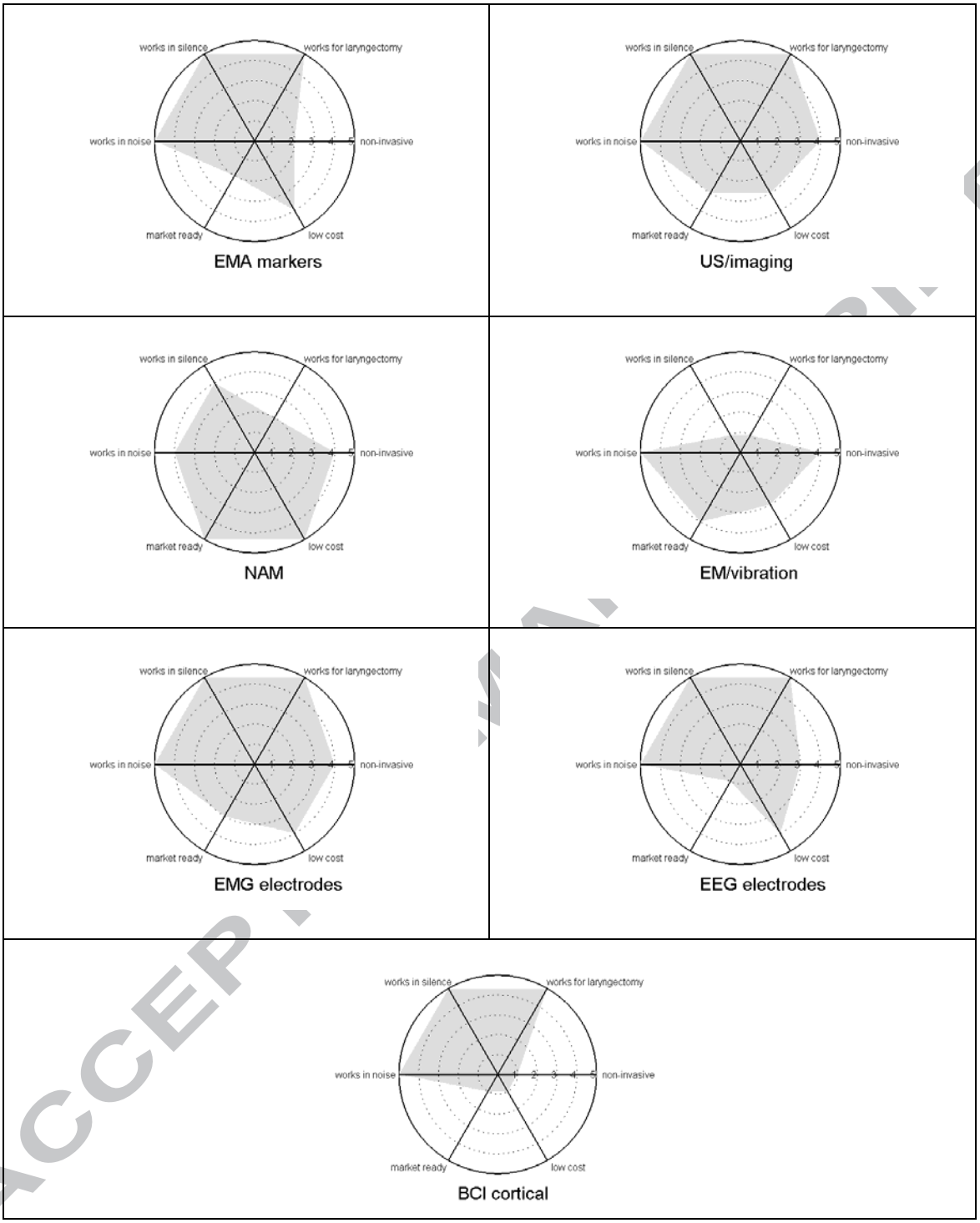


Figure 7. “Spiderweb” plots of the 7 SSI technologies described in the article. Axes are described in the text.

IV.C Future Directions:

Figure 7 shows that no SSI technology today proposes a device responding fully to the demands of all six axes of evaluation. Reaching this goal will likely require advances both in instrumentation and in signal processing. On the instrumentation side, three main lines of attack seem pertinent:

- First, the efforts being carried out on the individual technologies are in many cases still in the early stages. These projects need to continue and expand in order to extract the maximum potential from each technique.
- The systems being developed for SSIs today make use of technologies borrowed from other domains, such as general-purpose medical imaging or diagnostics. It would be interesting to develop dedicated instruments explicitly for use in the SSI field – for example, a customized ultrasound probe designed to highlight features which are the most pertinent for speech recognition and synthesis.
- Finally, it will be interesting to combine technologies in a multi-sensor device and perform data fusion, in hopes that the complementarity of the acquired streams can overcome the shortcomings of some of the devices on their own.

Signal processing and modelling advances in the SSI field are likely to come from one of two possible directions:

- techniques facilitating speech recognition and synthesis from incomplete representations of speech production mechanisms; and,
- a continued and enhanced symbiosis between the more “expert” type methods popular in articulatory inversion research, and the machine learning oriented approaches being employed in the majority of the current SSI investigations.

Finally, as this paper was going to press, Alcatel-Lucent Corporation issued a press release claiming an experimental **eighth** SSI technology based on low frequency ultrasound reflectrometry (Moeller (2008)). It will likely turn out, quite fittingly, that the last word on SSIs technologies – has not been spoken!

Acknowledgements

The authors acknowledge support from the French Department of Defense (DGA); the "Centre de Microélectronique de Paris Ile-de-France" (CEMIP); the French National Research Agency (ANR) under the contract number ANR-06-BLAN-0166; the ENT Consultants' Fund, Hull and East Yorkshire Hospitals NHS Trust; the National Institute on Deafness and other Communication Disorders (R01 DC07683; R44 DC007050-02); and the National Science Foundation (SBE-0354378). They also wish to thank Gérard Chollet; Maureen Stone; Laurent Benaroya; Gérard Dreyfus; Pierre Roussel; and Szu-Chen (Stan) Jou for their help in preparing this article.

References

Arnal et al. (2000) : Arnal, A., Badin, P., Brock, G., Connan, P.-Y., Florig, E., Perez, N., Perrier, P. Simon, P., Sock, R., Varin, L., Vaxelaire, B., Zerling, J.-P., 2000. Une base de données cinéradiographiques du français, XXIIIèmes Journées d'Etude sur la Parole, Aussois, France, p. 425-428.

Baken et al. (1984): Baken, R.J., Robbins, J., Fisher, H., Blom, E., Singer, M., 1984. A comparative acoustic study of normal, esophageal and tracheoesophageal speech production. Journal of Speech and Hearing Disorders, 49, pp. 202-210.

Bartels et al. (2008): Bartels, J.L., Andreasen, D., Ehirim, P., Mao, H., Seibert, S., Wright, E. J., and Kennedy, P.R., 2008. Neurotrophic electrode: method of assembly and implantation into human motor speech cortex, Journal of Neuroscience Methods, 174(2), pp. 168-176.

Betts et al. (2006): Betts, B.J., Binsted, K., Jorgensen, C., 2006. Small-vocabulary speech recognition using surface electromyography, Interacting with Computers: The Interdisciplinary Journal of Human-Computer Interaction, 18, pp. 1242-1259.

Birbaumer (2000): Birbaumer, N. (2000). The thought translation device (TTD) for completely paralyzed patients, IEEE Transactions on Rehabilitation Engineering, 8(2), pp.190–193.

Blankertz et al. (2006): Blankertz, B., Dornhege, G., Krauledat, M., Müller, K.-R., Kunzmann, V., Losch, F., Curio, G., 2006. The Berlin brain-computer interface: EEG-based

communication without subject training, IEEE Transactions on Neural Systems and Rehabilitation Engineering, 14(2), pp. 147–152.

Blom and Singer (1979): Blom, E.D., Singer, M.I., 1979. Surgical prosthetic approaches for postlaryngectomy voice restoration. In: Keith, R.L., Darley, F.C., Eds., Laryngectomy rehabilitation, Houston, Texas College Hill Press, pp. 251-76.

Bos and Tack (2005): Bos, J.C., Tack, D.W., Speech input hardware investigation for future dismounted soldier computer systems, DRCD Toronto CR 2005-064.

Brown et al. (2004): Brown, D.R., Ludwig, R., Pelteku, A., Bogdanov, G., Keenaghan, K., 2004. A novel non-acoustic voiced speech sensor, Measurement Science and Technology, 15, pp. 1291-1302.

Brown et al. (2005): Brown, D.R., Keenaghan, K., Desimini, S., 2005. Measuring glottal activity during voiced speech using a tuned electromagnetic resonating collar sensor, Measurement Science and Technology, 16, pp. 2381-2390.

Brumberg et al. (2007): Brumberg, J.S., Andreasen, D.S., Bartels, J.L., Guenther, F.H., Kennedy, P.R., Siebert, S.A., Schwartz, A.B., Velliste, M., Wright, E.J., 2007. Human speech cortex long-term recordings: formant frequency analyses, in Neuroscience Meeting Planner 2007, Program No. 517.17, San Diego, USA.

Brumberg et al. (2008): Brumberg, J.S., Nieto-Castanon, A., Guenther, F.H., Bartels, J.L., Wright, E.J., Siebert, S.A., Andreasen, D.S., and Kennedy, P.R., 2008. Methods for

construction of a long-term human brain machine interface with the Neurotrophic Electrode, in Neuroscience Meeting Planner 2007, Program No. 779.5, Washington, DC.

Brumberg et al. (2009, this issue): Brumberg, J.S, Nieto-Castanon, A., Kennedy, P.R., Guenther, F.H., 2009. Brain-Computer Interfaces for Speech Communication, Speech Communication, this issue.

Burnett et al. (1997): Burnett, G.C., Gable, T.G., Holzrichter, J.F., Ng, L.C., 1997. Voiced excitation functions calculated from micro-power impulse radar information, Journal of the Acoustical Society of America, 102, p. 3168(A).

Carstens (2008): Carstens Medizinelektronik, http://www.articulograph.de/.

Calliess and Schultz (2006): Calliess, J.-P., Schultz, T., 2006. Further investigations on unspoken speech, Studienarbeit, Universität Karlsruhe (TH), Karlsruhe, Germany.

Chan et al. (2001): Chan, A.D.C., Englehart, K., Hudgins, B., Lovely, D.F., 2001. Myo-electric signals to augment speech recognition, Medical and Biological Engineering and Computing, vol. 39, pp.500-504.

Chan (2003): Chan, A.D.C., 2003. Multi-expert automatic speech recognition system using myoelectric signals, Ph.D. Dissertation, Department of Electrical and Computer Engineering, University of New Brunswick (Canada).

Crevier-Buchman (2002): Crevier-Buchman, L., 2002. Laryngectomy patients and the psychological aspects of their tracheostomy, Review of Laryngology Otolaryngology and Rhinology, 123, pp. 137-139

Davidson (2005): Davidson, L., 2005. Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance, Journal of the Acoustical Society of America, 120:1, pp. 407-415.

DaSalla et al. (2009): DaSalla, C.S., Kambara, H., Sato, M., Koike, Y., 2009. Spatial filtering and single-trial classification of EEG during vowel speech imagery, Proceedings of the 3rd International Convention on Rehabilitation Engineering and Assistive Technology (i-CREATe 2009), Singapore, in press.

Dekens et al. (2008): Dekens, T., Patsis, Y., Verhelst, W., Beaugendre, F., Capman, F., 2008. A multi-sensor speech database with applications towards robust speech processing in hostile environments, Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco, 28-30 May 2008.

Denby and Stone (2004): Denby, B., Stone, M., 2004. Speech synthesis from real time ultrasound images of the tongue, Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'04), Montréal, Canada, 17-21 May 2004, volume 1, pp. I685-I688.

Denby et al. (2006): Denby, B., Oussar, Y., Dreyfus, G., Stone, M. 2006. Prospects for a Silent Speech Interface Using Ultrasound Imaging. IEEE ICASSP, Toulouse, France, pp. I365-I368.

Dornhege et al. Eds. (2007): Dornhege, G., del R. Millan, J., Hinterberger, T., McFarland, D., Müller, K.-R., Eds., 2007. Towards brain-computer interfacing, MIT Press.

Drummond et al., (1996): Drummond, S., Dancer, J., Krueger, K., Spring, G., 1996. Perceptual and acoustical analysis of alaryngeal speech: determinants of intelligibility, Perceptual Motor Skills, 83, pp. 801-802.

Dupont and Ris (2004): Dupont, S., Ris, C., 2004. Combined use of close-talk and throat microphones for improved speech recognition under non-stationary background noise, Proceedings of Robust 2004, Workshop (ITRW) on Robustness Issues in Conversational Interaction, Norwich, UK, August 2004.

Epstein (1983): Epstein, C.M., 1983. Introduction to EEG and evoked potentials, J. B. Lippincot Co.

Epstein (2005): Epstein, M.A., 2005. Ultrasound and the IRB, Clinical Linguistics and Phonetics, 16:6, pp. 567-572.

Fagan et al. (2008): Fagan, M.J., Ell, S.R., Gilbert, J.M., Sarrazin, E., Chapman, P.M., 2008. Development of a (silent) speech recognition system for patients following laryngectomy, Medical Engineering and Physics, 30:4, pp. 419-425.

Fitzpatrick (2002): Fitzpatrick, M., 2002. Lip-reading cellphone silences loudmouths, New Scientist, edition of 03 April 2002.

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

Georgeopoulos et al. (1982): Georgopoulos, A.P., Kalaska, J.F., Caminiti, R., Massey, J.T., 1982. On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex, Journal of Neuroscience, 2(11), pp. 1527-1537.

Gibbon, F., (2005): Bibliography of electropalatographic (EPG) studies in English (1957-2005), Queen Margaret University, Edinburgh, UK, September 2005, http://www.qmu.ac.uk/ssrc/cleftnet/EPG_biblio_2005_september.pdf.

Gracco et al. (2005): Gracco, V.L., Tremblay, P., Pike, B., Imaging speech production using fMRI, NeuroImage, volume 26, issue 1, 15 May 2005, pp. 294-301.

Guenther et al. (2008): Guenther, F.H., Brumberg, J.S., Nieto-Castanon, A., Bartels, J.L., Siebert, S.A., Wright, E.J., Tourville, J.A., Andreasen, D.S., and Kennedy, P.R., 2008. A brain-computer interface for real-time speech synthesis by a locked-in individual implanted with a Neurotrophic Electrode, in Neuroscience Meeting Planner 2008, Program No. 712.1, Washington, DC.

Hasegawa and Ohtani (1992): Hasegawa, T. and Ohtani, K., 1992. Oral image to voice converter, image input microphone, Proceedings of IEEE ICCS/ISITA 1992 Singapore, vol. 20, no. 1, pp. 617-620.

Hasegawa-Johnson (2008): Hasegawa-Johnson, M., 2008. Private communication.

Heracleous et al. (2007): Heracleous, P., Kaino, T., Saruwatari, H., and Shikano, K., 2007. Unvoiced speech recognition using tissue-conductive acoustic sensor, EURASIP Journal on Advances in Signal Processing, volume 2007, issue 1, pp. 1-11.

Hirahara et al. (2009, this issue): Hirahara, T., Otani, M., Shimizu, S., Toda, M., Nakamura, K., Nakajima, Y., Shikano, K., 2009. Silent-speech enhancement system utilizing body-conducted vocal-tract resonance signals, Speech Communication, this issue.

Hochberg et al. (2006): Hochberg, L. R., Serruya, M. D., Friehs, G. M., Mukand, J. A., Saleh, M., Caplan, A. H., Branner, A., Chen, D., Penn, R. D., and Donoghue, J. P. (2006). Neuronal ensemble control of prosthetic devices by a human with tetraplegia, Nature, 442(7099), 164-171.

Hochberg et al. (2008): Hochberg, L.R., Simeral, J.D., Kim, S., Stein, J., Friehs, G.M., Black, M.J., and Donoghue, J.P., 2008. More than two years of intracortically-based cursor control via a neural interface system, in Neurosicence Meeting Planner 2008, Program No. 673.15, Washington, DC.

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

Hoole and Nguyen (1999): Hoole, P., Nguyen, N., 1999. Electromagnetic articulography in coarticulation research, in Hardcastle, W.H., Hewlitt, N. Eds., Coarticulation: Theory, Data and Techniques, pp. 260-269, Cambridge University Press, 1999.

House and Granström (2002): House, D., Granström, B., 2002. Multimodal speech synthesis: Improving information flow in dialogue systems using 3D talking heads, in Artificial Intelligence: Methodology, Systems, and Applications, Lecture Notes in Computer Science, volume 2443/2002, Springer Berlin/Heidelberg, pp. 65-84.

Hueber et al. (2007a): Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., Stone, M., 2007. Eigentongue feature extraction for an ultrasound-based silent speech interface, IEEE International Conference on Acoustic, Speech, and Signal Processing, ICASSP07, Honolulu, vol. 1, pp. 1245-1248.

Hueber et al. (2007b): Hueber, T., Chollet, G., Denby, B., Stone, M., Zouari, L, 2007. Ouisper: Corpus based synthesis driven by articulatory data, International Congress of Phonetic Sciences, Saarbrücken, Germany, pp. 2193-2196.

Hueber et al. (2007c): Hueber, T., Chollet, G., Denby, B., Dreyfus, G., Stone, M., 2007. Continuous-speech phone recognition from ultrasound and optical images of the tongue and lips, Interspeech, Antwerp, Belgium, pp. 658-661.

Hueber et al. (2008a): Hueber, T., Chollet, G., Denby, B., Dreyfus, G., Stone, M., 2008. Phone recognition from ultrasound and optical video sequences for a silent speech interface. Interspeech, Brisbane, Australia, pp. 2032-2035.

Hueber et al. (2008b): Hueber, T., Chollet, G., Denby, B., Stone, M., 2008. Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application. International Seminar on Speech Production, Strasbourg, France, pp. 365-369.

Hueber et al. (2009, this issue): Hueber, T., Benaroya, E.-L., Chollet, G., Denby, B., Dreyfus, G., Stone, M., 2009. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips, Speech Communication, this issue.

Hummel et al. (2006): Hummel, J., Figl, M., Birkfellner,W., Bax, M.R., Shahidi, R., Maurer, C.R., Bergmann, H., 2006. Evaluation of a new electromagnetic tracking system using a standardized assessment protocol, Physics in Medicine and Biology, 51, pp. N205–N210.

IEEE (2008): Brain Computer Interfaces, IEEE Computer, volume 41, number 10, October 2008.

Jorgensen et al. (2003): Jorgensen, C., Lee, D.D., Agabon, S., 2003. Sub auditory speech recognition based on EMG signals, Proceedings of the International Joint Conference on Neural Networks (IJCNN), pp. 3128–3133.

Jorgensen and Binsted (2005): Jorgensen, C., Binsted, K., 2005. Web browser control using EMG based sub vocal speech recognition, Proceedings of the 38th Annual Hawaii International Conference on System Sciences, IEEE, pp. 294c.1 – 294c.8.

Jorgensen and Dusan (2009, this issue): Jorgensen, C., Dusan, S., 2009. Speech interfaces based upon surface electromyography, Speech Communication, this issue

Jou et al. (2006): Jou, S., Schultz, T., Walliczek, M., Kraft, F., 2006. Towards continuous speech recognition using surface electromyography, INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing, vol. 2, pp. 573–576.

Jou et al. (2007): Jou, S., Schultz, T., Waibel, A., 2007. Multi-stream articulatory feature classifiers for surface electromyographic continuous speech recognition, International Conference on Acoustics, Speech, and Signal Processing, IEEE, Honolulu, Hawaii.

Kennedy (1989): Kennedy, P.R., 1989. The cone electrode: a long-term electrode that records from neurites grown onto its recording surface, Journal of Neuroscience Methods, 29, pp. 181-193.

Kennedy (2006): Kennedy, P.R., 2006. Comparing electrodes for use as cortical control signals: Tiny tines, tiny wires or tiny cones on wires: which is best?, in The Biomedical Engineering Handbook, The Electrical Engineering Handbook Series, 3rd edition, vol. 1, Boca Raton, CRS/Taylor and Francis.

Kennedy et al. (2000): Kennedy, P.R., Bakay, R.A.E., Moore, M.M., Adams, K., Goldwaithe, J., 2000. Direct control of a computer from the human central nervous system, IEEE Transactions on Rehabilitation Engineering, 8(2), pp. 198-202.

Kennedy and Bakay (1998): Kennedy, P.R., and Bakay, R.A.E., 1998. Restoration of neural output from a paralyzed patient by direct brain connection, NeuroReport, 9, pp. 1707-1711.

Kim et al. (2007): Kim, S., Simeral, J.D., Hochberg, L.R., Donoghue, J.P., Friehs, G.M., Black, M.J., 2007. Multi-state decoding of point-and-click control signals from motor cortical activity in a human with tetraplegia, Neural Engineering, 2007, CNE'07 3rd International IEEE/EMBS Conference, pp. 486-489.

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

Lotte et al. (2007): Lotte, F., Congedo, M., Lecuyer, A., Lamarche, F., and Arnaldi, B., 2007. A review of classification algorithms for EEG-based brain computer interfaces, Journal of Neural Engineering, 4, pp. R1–R13.

Maier-Hein et al. (2005): Maier-Hein, L., Metze, F., Schultz, T., Waibel, A., 2005. Session independent non-audible speech recognition using surface electromyography, IEEE Workshop on Automatic Speech Recognition and Understanding, San Juan, Puerto Rico, pp. 331-336.

Manabe et al. (2003): Manabe, H., Hiraiwa, A., Sugimura, T., 2003. Unvoiced speech recognition using EMG-mime speech recognition, Proceedings of CHI, Human Factors in Computing Systems, Ft. Lauderdale, Florida, pp. 794-795.

Manabe and Zhang (2004): Manabe, H., Zhang, Z., 2004. Multi-stream HMM for EMG-based speech recognition, Proceedings of 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 1-5 Sept. 2004, San Francisco, California, volume 2, pp. 4389-4392.

Marchal and Hardcastle (1993): Marchal, A., Hardcastle, W.J., 1993. Instrumentation and database for the cross-language study of coarticulation, Language and Speech, 1993, 36, 1, pp. 3-20.

Maynard et al. (1997): Maynard, E.M., Nordhausen, C.T., and Normann, R.A., 1997. The Utah intracortical electrode array: a recording structure for potential brain-computer interfaces, Electroencephalography and Clinical Neurophysiology, 102(3), pp. 228-239.

Miller et al. (2007): Miller, L.E., Andreasen, D.S., Bartels, J.L., Kennedy, P.R., Robesco, J., Siebert, S.A., and Wright, E.J., 2007. Human speech cortex long-term recordings : Bayesian analyses, in Neuroscience Meeting Planner 2007, Program No. 517.20, San Diego, USA.

Moeller (2008): available online:
http://innovationdays.alcatel-lucent.com/2008/documents/Talking Beyond Hearing.pdf

Morse and O'Brien (1986): Morse, M.S., O'Brien, E.M., 1986. Research summary of a scheme to ascertain the availability of speech information in the myoelectric signals of neck and head muscles using surface electrodes, Computers in Biology and Medicine, vol. 16, no. 6, pp. 399-410.

Morse et al. (1989): Morse, M.S., Day, S.H., Trull, B., Morse, H., 1989. Use of myoelectric signals to recognize speech, in Images of the Twenty-First Century - Proceedings of the 11th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, part 2, vol.11. Alliance for Engineering in Medicine and Biology, pp. 1793-1794.

Morse et al. (1991): Morse, M.S., Gopalan, Y.N., Wright, M., 1991. Speech recognition using myoelectric signals with neural networks, in Proceedings of the 13th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 13., no 4, Piscataway, NJ, United States, IEEE, pp. 1877-1878.

Munhall et al. (1995): Munhall, K.G., Vatikiotis-Bateson, E., Tohkura, Y., 1995. X-ray film database for speech research, Journal of the Acoustical Society of America, 98, pp. 1222-1224.

Nakajima et al. (2003a): Nakajima, Y., Kashioka, H., Shikano, K., Campbell, N., 2003. Non-audible murmur recognition, Proceedings of Eurospeech 2003, pp. 2601-2604.

Nakajima et al. (2003b): Nakajima, Y., Kashioka, H., Shikano, K., Campbell, N., 2003. Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin, Proceedings of IEEE ICASSP, pp. 708-711.

Nakajima (2005): Nakajima, Y., 2005. Development and evaluation of soft silicone NAM microphone, Technical Report IEICE, SP2005-7, pp. 7-12, in Japanese.

Nakajima et al. (2006): Nakajima, Y., Kashioka, H., Campbell, N., Shikano, K., 2006. Non-audible murmur (NAM) recognition, IEICE Transactions on Information and Systems, E89-D, 1, pp. 1-8.

Nakamura (1988): Nakamura, H., 1988. Method of recognizing speech using a lip image, Unites States patent 4769845, September 06, 1988.

NessAiver et al. (2006): NessAiver, M.S., Stone, M., Parthasarathy, V., Kahana, Y., Paritsky, A., 2006. Recording high quality speech during tagged cine-MRI studies using a fiber optic microphone, Journal of Magnetic Resonance Imaging 23, pp. 92–97.

Neuper et al. (2003): Neuper, C., Müller, G. R., Kübler, A., Birbaumer, N., Pfurtscheller, G., 2003. Clinical application of an EEG-based brain computer interface: a case study in a patient with severe motor impairment, Clinical Neurophysiology, 114, pp. 399-409.

Ng et al. (2000): Ng, L., Burnett, G., Holzrichter, J., Gable, T., 2000. Denoising of human speech using combined acoustic and EM sensor signal processing, International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Istanbul, Turkey, 5-9 June 2000, vol. 1, pp. 229-232.

Nguyen et al. (1996): Nguyen, N., Marchal, A., Content, A., 1996. Modeling tongue-palate contact patterns in the production of speech, Journal of Phonetics, 1996, pp. 77-98.

Nijholt et al. (2008): Nijholt, A., Tan, D., Pfurtscheller, G., Brunner, C., J.d.R. Millan, Allison, B., Graimann, B., Popescu, F., Blankertz, B., Fraunhofer, Müller, K.R., Brain-Computer Interfacing for Intelligent Systems, Intelligent Systems, vol. 23, no. 3, pp. 72-79.

Otani et al. (2008): Otani, M., Shimizu, S. & Hirahara, T., 2008. Vocal tract shapes of non-audible murmur production, Acoustical Science and Technology, 29, pp. 195-198.

Ouisper (2006): Oral Ultrasound synthetIc SpEech SouRce, “Projet Blanc”, National Research Agency (ANR), France, contract number ANR-06-BLAN-0166, 2006-2009.

Patil and Hansen (2009, this issue): Patil, S.A., Hansen, J.H.L., 2009. A competitive alternative for speaker assessment: Physiological Microphone (PMIC), Speech Communication, this issue.

Perkell et al. (1992): Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I., and Jackson, M., 1992. Electro-magnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements, J. Acoust. Soc. Am. 92, pp. 3078-3096.

Petajan (1984): Petajan, E.D., 1984. Automatic lipreading to enhance speech recognition, in IEEE Communications Society Global Telecommunications Conference, Atlanta, USA.

Porbadnigk (2009): Porbadnigk, A., Wester, M., Calliess, J., Schultz, T., 2009. EEG-based speech recognition – impact of temporal effects, Biosignals 2009, Porto, Portugal, January 2009, pp. 376-381.

Preuss et al. (2006): Preuss, R.D., Fabbri, D.R., Cruthirds, D.R., 2006. Noise robust vocoding at 2400 bps, 8th International Conference on Signal Processing, ICSP 2006, Guilin, China, November 16-20, 2006, vol. 1, pp. 16-20.

Quatieri et al. (2006): Quatieri, T.F., Messing, D., Brady, K, Campbell, W.B., Campbell, J.P., Brandstein, M., Weinstein, C.J., Tardelli, J.D., and Gatewood, P.D., 2006. Exploiting nonacoustic sensors for speech enhancement, IEEE Transactions on Audio, Speech, and Language Processing, March 2006, volume: 14, issue: 2, pp. 533- 544.

Rothenberg (1992): Rothenberg, M., A multichannel electroglottograph, Journal of Voice, vol. 6, no. 1, pp. 36-43, 1992.

Rubin and Vatikiotis-Bateson (1998): Rubin, P., Vatikiotis-Bateson, E. 1998. Talking heads, Proceedings of Audio Visual Speech Processing 1998, pp. 233-238.

Sajda et al. Eds. (2008): Sajda, P., Mueller, K.-R., Shenoy, K.V., Eds., special issue, Brain Computer Interfaces, IEEE Signal Processing Magazine, Jan. 2008.

Schönle et al. (1987): Schönle, P. W., Gräbe, K., Wenig, P., Höhne, J., Schrader, J., Conrad, B. 1987. Electromagnetic articulography: Use of alternating magnetic fields for tracking

movements of multiple points inside and outside the vocal tract, Brain and Language, 31, pp. 26–35.

Schroeter et al. (2000): Schroeter, J., Ostermann, J., Graf, H.P., Beutnagel, M., Cosatto, E., Syrdal, A., Conkie, A., Stylianou,Y., 2000. Multimodal speech synthesis, IEEE International Conference on Multimedia and Expo 2000, pp. 571-574.

Schultz and Wand (2009, this issue): Schultz, T., Wand, M., 2009. Modeling coarticulation in large vocabulary EMG-based speech recognition, Speech Communication, this issue.

Stone et al. (1983): Stone, M., Sonies, B., Shawker, T., Weiss, G., Nadel, L., 1983. Analysis of real-time ultrasound images of tongue configuration using a grid-digitizing system, Journal of Phonetics, 11, pp. 207-218.

Stone and Shawker (1986): Stone, M., Shawker, T., 1986. An ultrasound examination of tongue movement during swallowing, Dysphagia, l, pp. 78-83.

Stone and Davis (1995): Stone, M., Davis, E., 1995. A head and transducer support (HATS) system for use in ultrasound imaging of the tongue during speech, Journal of the Acoustical Society of America, 98, pp. 3107-3112.

Stone (2005): Stone, M., 2005. A guide to analyzing tongue motion from ultrasound images, Clinical Linguistics and Phonetics. 19(6-7), pp. 455–502.

Suppes et al. (1997): Suppes, P., Lu, Z.-L., Han, B., 1997. Brain wave recognition of words, Proceedings of the National Academy of Scientists of the USA, 94, pp.14965–14969.

Sugie and Tsunoda (1985): Sugie, N., Tsunoda, K., 1985. A speech prosthesis employing a speech synthesizer-vowel discrimination from perioral muscle activities and vowel production, IEEE Transactions on Biomedical Engineering, vol. BME-32, no. 7, pp. 485-490.

Tardelli Ed. (2004): Tardelli, J.D., Ed., 2003. MIT Lincoln Labs report ESC-TR-2004-084, Pilot corpus for multisensor speech processing.

Tatham (1971): Tatham, M., 1971. The place of electromyography in speech research, Behavioural Technology, no. 6.

TERC (2009): online, http://spinlab.wpi.edu/projects/terc/terc.html

Titze et al. (2000): Titze I. R., Story, B.H., Burnett, G.C., Holzrichter, J.F., NG, L.C., Lea W.A., 2000. Comparison between electroglottography and electromagnetic glottography, Journal of the Acoustical Society of America, January 2000, volume 107, issue 1, pp. 581-588.

Tran et al. (2008a): Tran, V.-A., Bailly, G., Loevenbruck, H., Jutten C., 2008. Improvement to a NAM captured whisper-to-speech system, Interspeech 2008, Brisbane, Australia, pp. 1465-1498.

Tran et al. (2008b); Tran, V.-A., Bailly, G., Loevenbruck, H., Toda, T., Predicting F0 and voicing from NAM-captured whispered speech, Proceedings of Speech Prosody, Campinas, Brazil.

Tran et al. (2009, this issue): Tran, V.-A., Bailly, G., Loevenbruck, H., Toda, T., 2009. Improvement to a NAM-captured whisper-to-speech system, Speech Communication, this issue.

Trucculo et al. (2008): Truccolo, W., Friehs, G.M., Donoghue, J.P., and Hochberg, L.R., 2008. Primary motor cortex tuning to intended movement kinematics in humans with tetraplegia, Journal of Neuroscience, 28(5), pp. 1163-1178.

Walliczek et al. (2006): Walliczek, M., Kraft, F., Jou, S.-C., Schultz, T., Waibel, A., 2006. Sub-word unit based non-audible speech recognition using surface electromyography, Proceedings of Interspeech, Pittsburgh, USA, pp. 1487-1490.

Wand and Schultz (2009): Wand, M., Schultz, T., 2009. Towards speaker-adaptive speech recognition based on surface electromyography, Biosignals, Porto, Portugal, in press.

Wester and Schultz (2006): Wester, M. and Schultz, T., 2006. Unspoken speech - speech recognition based on electroencephalography, Master's thesis, Universität Karlsruhe (TH), Karlsruhe, Germany.

Wolpaw et al. (2002): Wolpaw, J. R., Birbaumer, N., McFarland, D., Pfurtscheller, G., Vaughan, T., 2002. Brain-computer interfaces for communication and control, Clinical Neurophysiology, 113(6), pp. 767-791.

Wrench and Scobbie (2003): Wrench, A.A., and Scobbie, J.M., 2003. Categorising vocalisation of English / l / using EPG, EMA and ultrasound, 6th International Seminar on Speech Production, Manly, Sydney, Australia, 7-10 December 2003, pp. 314–319.

Wrench and Scobbie (2007): Wrench, A., Scobbie, J., Linden, M., 2007. Evaluation of a helmet to hold an ultrasound probe, Ultrafest IV, New York, USA.

Wright et al. (2007): Wright, E.J., Andreasen, D.S., Bartels, J.L., Brumberg, J.S., Guenther, F. H., Kennedy, P.R., Miller, L.E., Robesco, J., Schwartz, A.B., Siebert, S.A., and Velliste, M., 2007. Human speech cortex long-term recordings: neural net analyses, in Neuroscience Meeting Planner 2007, Program No. 517.18, San Diego, USA.