Project Telepathy: Targeted Verbal Communication using 3D Beamforming Speakers and Facial Electromyography

Anne-Claire Bourland
University of Bristol
Bristol, BS8 1TH, UK.
ab14188.2014@my.bristol.ac.uk

Asier Marzo
University of Bristol
Bristol, BS8 1TH, UK.
amarzo@hotmail.com

Peter Gorman
University of Bristol
Bristol, BS8 1TH, UK.
pg14214.2014@my.bristol.ac.uk

Jess McIntosh
University of Bristol
Bristol, BS8 1TH, UK.
jm0152@my.bristol.ac.uk

CHI'17 Extended Abstracts, May 06-11, 2017, Denver, CO, USA.
 ISBN 978-1-4503-4656-6/17/05.
DOI: http://dx.doi.org/10.1145/3027063.3053129

Abstract
Speech is our innate way of communication. However, it has limitation such as being a broadcast process, it has limited reach and it only works in air. Here, we explore the combination of two technologies for realizing natural targeted communication with potential applications in coordination of tasks or private conversations. For detecting words, we measure the bioelectric signals produced by facial muscles during speech. An electromyographic system composed of 4 surface electrodes had an accuracy of 80% when discriminating between 10 words. More importantly, the system was equally effective in discriminating spoken and silently mouthed words. For transferring the words, we used the sound-through-ultrasound phenomenon to generate audible sound within a narrow beam. We built a phased array of ultrasonic emitters, capable of emitting sound that can be steered electronically without physically moving the array. Two prototypes that combine detection and transfer of words are presented and their limitations analysed.

Author Keywords
Targeted Communication; Electromyography; Ultrasound; Directive Speakers.

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

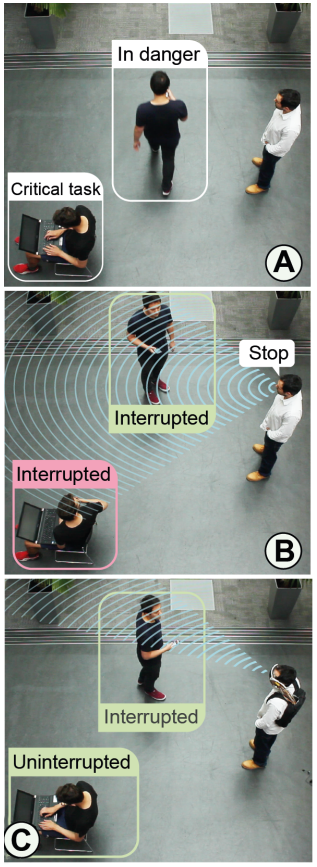

Figure 1: A) An observer sees a person that is going to trip over. B) Using speech: the observer says stop and undesirably interrupts a nearby person. C) With Telepathy: the observer silently mouths and the system emitted by a directional speaker only towards the target.

Introduction

Speech is our intrinsic way of communicating information. Normally it is a broadcast process meaning that when we speak, the information is received by all the people in the environment. However, in several human interactions directing the message towards specific recipients is of paramount importance. Think of secret information that you want to communicate without others noticing in a tough business negotiation, or coordinating a team of firefighters. Also, our voice can only be propagated properly in air and it has limited reach.

In this paper, we enhanced and combine two technologies for bringing ubiquitous targeted communication a step closer to its realization. On the one hand, an electromyographic system is used to capture the bioelectrical signals of the face muscles and detect silently mouthed words; this system is simple, portable, affordable and gives functional accuracy. On the other hand, we use an ultrasonic directive speaker that can electronically steer the beam in 3D to accurately point towards the target without mechanical actuation. More importantly, we combine this technologies into a wearable device. (Figure 1)

Related Work

To achieve a wearable targeted communicator we think that two technologies are necessary. Firstly, to recognize the word that we want to transmit without the user having to pronounce it, otherwise it would be heard by other people. Although this can be achieved with button pads or similar devices, we wanted a natural way. Secondly, a method for projecting the sound in a directive manner so that only a targeted individual receives the message.

Silent speech recognition

When our muscles move, they generate electric signals that can be detected even with surface electrodes. This allows to detect gestures, movements and in our case, speech. A review of EMG silent speech recognition [1] reinforces the potential of this technology for underwater communication, noisy environments, discreet or secure communications, and for users with speech related disabilities.

NASA has a lot of interest in this method since it could facilitate natural communication during space exploration missions. In their study, surface electrodes around the larynx were able to differentiate between 6 whispered words with accuracy rates of 90% [8].

Using Brain Computer Interfaces (BCI) that directly pick the signals from the neuron activity it was possible to recognize between two different syllabus with greater-than-chance accuracy [5]. Despite being promising, this technology is still not sufficient for recognizing a small vocabulary.

It was suggested that EMG signals can be captured directly from subvocalizing actions (i.e. just thinking loud about the words) [4]. However, later research has reported no existing EMG signals while subvocalizing [11].

On the contrary, EMG signals can be used to recognize the individual mouthed English phonemes with an accuracy better than chance [3]. Later, 90% accuracy in recognizing five English vowels was obtained [12]. With a dense array of electrodes over the larynx it was possible to differentiate Japanes vowels with an accuracy of 80% [9]. With 11 electrodes, it was possible to differentiate 65 words with 80% accuracy [11]. This result is very promising but the amount of required electrodes and bulky EMG equipment hinders its use in a wearable system.

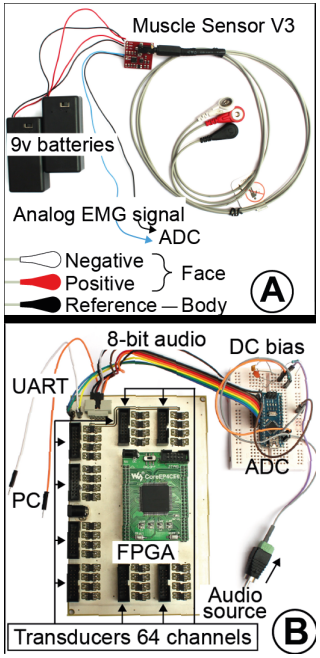

Figure 2: An electromyography set of electrodes with its batteries and amplifier. B) Directional Speaker driver board.

Directive Sound

An ultrasonic wave with a specific modulation generates audible sound. The sound is generated in a narrow beam due to the high frequency of the carrier and demodulated due to the non-linearity of air. This is the principle behind the Audio Spotlight [18] that was proposed back in the 80s. Several commercial systems for directional speakers exist today Holosonic, SoundLazer or Ultrasonic for applications in exhibitions, public displays or super markets. The military forces also has their own system LRAD capable of producing a narrow beam of sound that reaches a couple of kilometers.

Not only air, but the surface of objects can act as a demodulator, when the ultrasound beam is pointed towards an object, audible sound seems to be emitted from it. This has been used to direct the attention of users towards particular objects [7] with the directional speaker mounted on a mechanical rotational stage to steer the beam. Similarly, a hand-held directional speaker was pointed towards an object, a computer vision system recognized the object and projected into it associated sound clips [16].

The directivity of ultrasound speakers have been used to create sound sculptures [13] in which the audience received sound stimuli from specific locations forming a 3D landscape of sound, or as a way of transferring targeted audible information to the users and their devices [17].

Sound through ultrasound is not the only way of generating localised sound. Acoustic field synthesis is another alternative that uses traditional speakers to simulate localised sound sources [2]. This provides interesting possibilities in authoring 3D sound with virtual environments [10] or tangible control of sound sources [15]. However, this method lacks the ability to create directional focused beams of sound and the systems have to be large.

In interactive scenarios the directive nature of the speakers have not been explored as much as its ability to generate localised sources. More importantly, the steering of the beam has to be done mechanically being this slow, inaccurate and hindering the development of a wearable system.

Portable Facial Electromyography

The objective of this section is to present the portable system developed to detect mouthed words using EMG. Muscle tissue generates a difference in electrical potential between two points when the muscles contract or expand. This means that each word will have a unique electric pattern as a combination of the different muscle action potentials used to generate it.

Electrodes Location

A set of EMG surface electrodes consists of two detection electrodes and a reference electrode. The muscle action potential is measured between the detection electrodes and the reference electrode provides a common reference. The detection electrodes should not be placed too far apart and at carefully selected positions to minimize cross-talk between different muscles. The reference electrode should be placed in an electrically neutral tissue.

The employed electrodes were too large for some muscles, hence Levator Anguli Oris and Depressor Labii were discarded since they generated too much cross-talk. After testing different combinations, the chosen set of locations was: Depressor anguli oris, orbicularis oris upper and lower, and mylihoid. In a pilot study, these locations obtained the highest accuracy.

System

We used 4 sets of EMG surface electrodes (Skintact Ag/AgCl), each set was connected to a Muscle Sensor v3 (SparkFun) designed to amplify, rectify and smooth the signal. It was

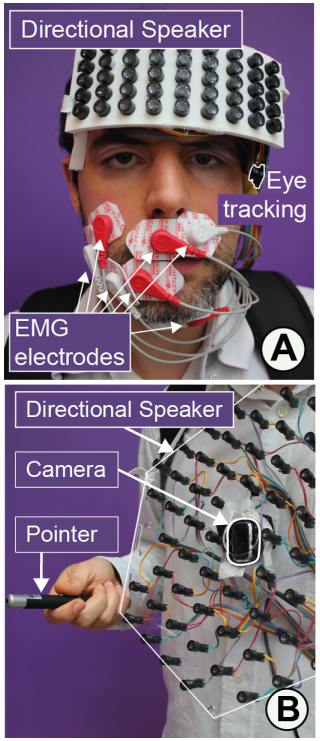

Figure 3: Prototypes for Telepathy systems. A) Ultra Band, B) Sonic Chest.

powered with a pair of 9V batteries. An Arduino Uno was used to sample the four analog signals and send them to a portable PC through Serial communication (Figure 2.A).

For each of the signals we computed four features: the average, standard deviation, wavelet and summation of values; these features provided the best results in the pilot study. With the features from the four channels we trained a machine learning algorithm to determine the pronounced word. We tested Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Neural Networks (NN) and a Gaussian classifier; the accuracies were 80%(SD=15), 55%(SD=9), 58%(SD=12) and 35%(SD=18) respectively. We chose the LDA since it gave the best result. SciKit was used for the machine learning algorithms.

User study

6 participants took part in the study (2 female, 4 male). The participants were asked to sit in front of the computer. After the electrodes were placed they were asked to follow the instructions displayed on the screen. The video showed the user which word to pronounce or mouth and an indicator gave visual cues for the timing.

The word set consisted of 10 commands and had variety in terms of phonemes: back, faster, forwards, left, no, right, stop, turn, up and yes. We consider that other sets words could be used.

The conditions for the experiment were mumbling (complete silent speech) and speaking. For each condition the participant had a total of 10 words repeating 10 times each (words changing every 5 pronunciations). The average session lasted 40 min, including the time to place the electrodes and explain the study to the participant. To summarize: 6 participants X 2 conditions X 10 words X 10 repetitions = 1200 words.

Results

The accuracy was calculated across a 10-fold cross validation classification results. It was 80.3% (SD=15.4) for the spoken condition and 83.3%(SD=7.9) for the mouthed condition; the difference was not significant (t(5)=-0.7, p=0.516).

The accuracy in word recognition ranged from 74% to 97%. The lowest accuracy obtained could have suffered from external factors, we hypothesize that some of the electrodes lost contact with the skin during the study. Furthermore, we think that differences between certain English accents had effects on how one forms words with their facial muscles, this could be a factor that influences the study as the participants displayed different accents. In summary, the results from the study suggest that there is no major difference between the speaking and mumbling conditions in classification and that the accuracy rates are sufficient for testing a prototype.

3D Beamformer Speaker

In this section we present a system that is capable of generating an ultrasonic beam that can be steered without mechanical actuation and modulated to generate audible sound along its narrow path. Previous systems have used ultrasound to generate a directive beam feeding the same signal into all the ultrasonic transducers. Here, we use the technique employed in phased arrays [14] to steer the beam electronically and therefore within milliseconds and millimetre accuracy.

An FPGA Cyclone IV was used to generate the ultrasonic signals. It had 64 digital outputs that were amplified by MOSFET drivers (TC4427) up to 18vpp. A half-square signal was generated for each channel and the narrow-band nature of the transducers make them output a sinusoidal

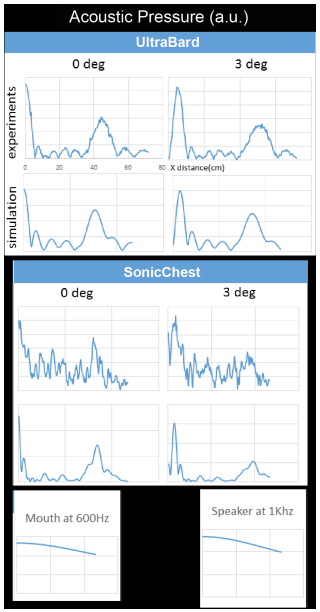

Figure 4: Amplitude scans and simulations for UltraBand and SonicChest. The beam is center at 0 degrees and electronically steered 3deg. At the right it can be seen the directivity pattern of a 10cm diameter speaker and a person speaking.

wave. The FPGA had 8 digital inputs to control the amplitude of the emitted waves and another UART input so that a computer could send the phase for each of the channels. The transducers had a central frequency of 40kHz and produced 120db at 30cm (MA40S4S). The whole system consumed 3W at 12vpp playing human voice commands.

The phase necessary to focus or steer the beam was calculated by time-reversal [6]. That is, to focus at one point, the phases of the transducers are set as kd where k is the wavenumber (2pi / lambda), lambda is the wavelength (8.6mm in our case) and d is the distance between the transducer and the target point. Analog audio was generated with the computer, PureData was used to band-pass filter between 400Hz and 4kHz (human speech range), compress and limit the audio to increase its playback quality. An Arduino Nano used the internal ADC to generate the 8bits digital signal from the audio that was fed into the FPGA. The audio input was biased with 3.3v coming from the Arduino. The driver board is presented in Figure 2.B.

Telepathy Systems

In this section we describe two different prototypes that were used to combine silent speech detection with 3D beamforming directional speaker. In all the instances the facial electromyographic system was composed of 4 pair of electrodes that were on the face and a directional speaker. Different sizes and locations were picked for the speaker as well as ways of electronically finely steer the beam towards a target. The ability of finely steering the sound beam electronically is a must in this systems since a 1 degree offset could imply a 60% reduction of the pressure level received by the target.

Ultra Band

A bent surface composed of 12x5 ultrasonic speakers is wore on the forehead. At the same time it has camera that performs eye-tracking (PyGaze). The user can broadly point his or her head towards the target and used the gaze to finely target the sound beam. Ultra band is shown in Figure 3.A.

Sonic Chest

A flat hexagonal pattern of 64 speakers was wore on the chest, the emitters were separated by 4 wavelengths. The user could point towards the target by orientating the torso. For finely controlling the beam, a laser pointer was used to indicate the exact target point. The array had a camera in the center that could locate the position of the pointer. The beamforming system was aligned with the camera. Sonic Chest is depicted in Figure 3.B.

Comparison

Both systems were usable for targeting commands at different individuals. However, UltraBand had to be small therefore making the sound not as directive as in Sonic Chest. Gaze was a natural way of pointing but the accuracy was better using a target pointer. Sonic Chest was more cumbersome to wear and to perform maneuvers while wearing. Incidentally, the sound emitted with UltraBand was clearly discernible for the wearer due to bone conduction and thus provided some sort of feedback mechanisms.

The directivity of the beam is determined by the frequency and the aperture of the array. The highest the frequency the more directive the beam is, but commercially available transducers work at 40kHz. Besides, the higher the frequency the faster it attenuates on air. Consequently, the only way maximizing the directivity of the beam was to increase the aperture, but this has a limit since the system has to be wearable. In Figure 4 it is shown the ampli-

tude levels (arbitrary units) for the two systems electronically steering the beam 3 degrees, for comparison we also shown the patterns for a speaker and a person speaking. The scans were taken with an electronic stage with a microphone mounted on it, the distance from the speaker to the array was 90cm and it moved across 74cm horizontally.

Future Work

Other means of capturing words need to be tested. Using contact microphones a certain amount of sound needs to be produced by the user. We tried to detect the resonant frequency of the mouth cavity but the results were not promising. Recognizing words by lip reading with a camera seems an interesting input modality although it would be sensitive to light conditions. We could also use something simpler as a keypad. However, we wanted to use a system as natural and close to speech as possible, mumbling seems the closest since Brain Computer Interfaces or the speculated subvocalization signals are not feasible.

It will be very interesting to run a user study given the diverse scenarios in which Telepathy can be tested (task coordination, secret information, underwater tasks). However, in a first development of the technology that was unfeasible.

The steering performance of the directional speaker needs to be tested more deeply. Sound can be focused and steered at any point of space. Although as the angles get stepper, the sidelobes become a bigger problem. We just show the case for 3 degrees because in the prototypes the body was use to point broadly and then the electronic steering was used for fine steering of the sound beam (directed by gaze or light pointing).

In this paper, we have only tried to surpass the limitation of speech being a broadcast process. But there is further investigation in increasing its reach with more powerful directional speakers and make it work underwater with water-coupled ultrasonic transducers.

The accuracy in discriminating words still has to be improved if these systems had to be used in real scenarios, we hope that further progress in wearable transparent electrodes such as Second Skin will improve the signals and make the system less cumbersome. The advances in smart textiles and wearable computing will also improve this aspect.

On the other hand, the directivity of the speakers could be improved but not by increasing their size since they have to be wearable, one possible solution would be manufacturing custom transducers that operate in air at higher frequencies.

Conclusion

We have presented two systems aimed at targeted natural communication that combine and enhanced both silent speech recognition through facial electromiographic signal processing, and a 3D beamformer ultrasonic directional speaker.

Despite the limitations, the systems are a step closer towards natural targeted verbal communication. This systems would have a significant impact in underwater exploration, space missions or military scenarios.

Acknowledgments

Anne and Peters were funded by University of Bristol summer internships. Jess is supported by EPSRC Doctoral Training (grant EP/M507994/1). Asier Marzo is funded by UK EPSRC (EP/N014197/1). We thank Matt Sutton for assisting with the elaboration of figures and videos.

References

[1] Amean Shareaf Al_safi and Liqaa Alhafadhi. 2015. Review of EMG-based Speech Recognition. *IJRECE* 3, 3 (2015), 56–60.

[2] Augustinus J Berkhout, Diemer de Vries, and Peter Vogel. 1993. Acoustic control by wave field synthesis. *The Journal of the Acoustical Society of America* 93, 5 (1993), 2764–2778.

[3] Kim Binsted, Charles Jorgensen, and IC Code. 2006. Sub-auditory speech recognition. Citeseer.

[4] Michael Braukus and John Bluck. 2004. NASA Develops System To Computerize Silent Subvocal Speech. *Found on NASA website* (2004).

[5] Michael D'Zmura, Siyi Deng, Tom Lappas, Samuel Thorpe, and Ramesh Srinivasan. 2009. Toward EEG sensing of imagined speech. In *International Conference on Human-Computer Interaction*. Springer, 40–48.

[6] Emad S Ebbini and Charles A Cain. 1989. Multiple-focus ultrasound phased-array pattern synthesis: optimal driving-signal distributions for hyperthermia. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 36, 5 (1989), 540–548.

[7] Kentaro Ishii, Yukiko Yamamoto, Michita Imai, and Kazuhiro Nakadai. 2007. A navigation system using ultrasonic directional speaker with rotating base. In *Symposium on Human Interface and the Management of Information*. Springer, 526–535.

[8] Chuck Jorgensen, Diana D Lee, and Shane Agabont. 2003. Sub auditory speech recognition based on EMG signals. In *Neural Networks, 2003. Proceedings of the International Joint Conference on*, Vol. 4. IEEE, 3128–3133.

[9] Takatomi Kubo, Masaki Yoshida, Takumu Hattori, and Kazushi Ikeda. 2014. Towards excluding redundancy in electrode grid for automatic speech recognition based on surface EMG. *Neurocomputing* 134 (2014), 15–19.

[10] Frank Melchior, Tobias Laubach, and Diemer De Vries. 2005. Authoring and user interaction for the production of wave field synthesis content in an augmented reality system. In *Proceedings of the 4th IEEE/ACM International Symposium on Mixed and Augmented Reality*. IEEE Computer Society, 48–51.

[11] Geoffrey S Meltzner, Jason J Sroka, James T Heaton, L Donald Gilmore, Glen Colby, Serge H Roy, Nancy Chen, and Carlo J De Luca. 2008. Speech recognition for vocalized and subvocal modes of production using surface EMG signals from the neck and face.. In *INTERSPEECH*. 2667–2670.

[12] José AG Mendes, Ricardo R Robson, Sofiane Labidi, and Allan Kardec Barros. 2008. Subvocal speech recognition based on EMG signal using independent component analysis and neural network MLP. In *Image and Signal Processing, 2008. CISP'08. Congress on*, Vol. 1. IEEE, 221–224.

[13] Daichi Misawa. 2013. Transparent sculpture: an embodied auditory interface for sound sculpture. In *Proceedings of the 7th International Conference on Tangible, Embedded and Embodied Interaction*. ACM, 389–390.

[14] Michael Moles. 2012. Ultrasonic Phased Array. *The NDT Technician* 11, 3 (2012), 1–5.

[15] Jörg Müller, Matthias Geier, Christina Dicke, and Sascha Spors. 2014. The boomRoom: mid-air direct interaction with virtual sound sources. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 247–256.

[16] Ken Nakagaki and Yasuaki Kakehi. 2011. SonalShooter: a spatial augmented reality system using handheld directional speaker with camera. In *ACM SIGGRAPH 2011 Posters*. ACM, 82.

[17] Hitomi Tanaka and Yasuaki Kakehi. 2013. SteganoSonic: a locally information overlay system using parametric speakers. In *ACM SIGGRAPH 2013 Posters*. ACM, 95.

[18] Masahide Yoneyama, Jun-ichiroh Fujimoto, Yu Kawamo, and Shoichi Sasabe. 1983. The audio spotlight: An application of nonlinear interaction of sound waves to a new type of loudspeaker design. *The Journal of the Acoustical Society of America* 73, 5 (1983), 1532–1536.