Silent Speech for Human-Computer Interaction

João Freitas[1,2], António Teixeira[2] and Miguel Sales Dias[1,3]
[1]*Microsoft Language Development Center, Lisboa, Portugal*
[2]*Departamento de Electrónica Telecomunicações e Informática/IEETA, Universidade de Aveiro, Aveiro, Portugal*
[3]*ISCTE-University Institute of Lisbon, Lisboa, Portugal*

1 STAGE OF THE RESEARCH

Speech communication has been and still is the dominant mode of human-human communication and information exchange. Therefore, an interface based on speech allows people to interact with machines in a more natural and effective way (Teixeira et al., 2009) and, for this reason, spoken language technology has suffered a significant evolution in the last years. However, conventional automatic speech recognition (ASR) systems use only a single source of information – the audio signal. When this audio signal becomes corrupted in the presence of environmental noise or assumes particular patterns, like the ones verified in elderly speech, speech recognition performance degrades, leading users to opt by a different modality or to not use the system at all. This type of systems have also revealed to be inadequate in situations where privacy is required, for users without the ability to produce an audible acoustic signal (e.g. users who have undergone a laryngectomy) and users with speaking difficulties and speech impairments.

To tackle these problems and being speech a privileged interface for Human-Computer Interaction (HCI), a novel Silent Speech Interface (SSI) based on multiple modalities is envisioned. We propose an SSI for European Portuguese (EP), a language for which no SSI has yet been developed. In our in depth state-of-the-art critical assessment, we have identified several modalities to convey silent speech data and address the issues raised by adapting existing work on SSIs to EP such as, the recognition of nasal vowels. From this analysis several modalities with low-invasiveness were selected and a set of preliminary experiments based on Video, Depth, Surface Electromyography (sEMG) and Ultrasonic Doppler Sensing (UDS) were conducted. Taking in consideration the results from the literature review and the experiments, we have decided to develop a multimodal SSI for EP. Results have also show recognition problems between minimal pairs of words that only differ on a nasal sound using the visual and the sEMG approach, supporting our planned development of an SSI based on the fusion of multiple modalities and motivating the investigation of the detection of nasal sounds using less invasive approaches. We are presently collecting the necessary *corpora* for developing a prototype and analysing the use of an additional sEMG sensor to capture the myoelectric signal coming from the muscles related with the nasality phenomena.

2 OUTLINE OF OBJECTIVES

The objectives defined for this PhD thesis are the following:

European Portuguese Adoption – The adaptation of SSIs to a new language and the procedures involved constitute by itself an extension to the current scientific knowledge in this area. With this work we will address the challenges of developing a SSI for EP, the first approach for this language in the Portuguese and international academia. Using the techniques described in literature and adapting them to a new language will provide novel information towards language independence and language adoption techniques.

Identify and Address Problems caused by Nasality – Motivated by the EP adoption, one of the areas of research to address is the problem of recognizing nasal sounds, as pointed out in (Denby et al., 2010). Considering the particular nasal characteristics associated with EP, we have noticed performance deterioration in terms of recognition rates and accuracy using existent approaches. When this occurs, the root of the system performance deterioration cause needs to be identified and new techniques based on that information need to be thought. For example, adding a sensor that that can provide complementary information. This will allow concluding particular aspects that influence language expansion, language independency and limitations

Freitas J., Teixeira A. and Dias M..
Silent Speech for Human-Computer Interaction.

of SSIs for the EP case.

Multi-sensor Analysis - An SSI can be implemented using several types of sensors working separately or a multimodal combination of them in order to achieve better results. For this work we will preferably adopt the less invasive approaches and sensors that are able to work both in silent and noisy environments. Further investigation will also be conducted on silent speech processing, respectively on data acquisition, feature extraction, and classification, as well as, on combining techniques through multiple sensor devices, data fusion and solving asynchrony issues verified in different signals (Srinivasan et al., 2010) in order to complement and overcome the inherent shortcomings of some approaches without decreasing the usability of the system.

User Requirements and Scenarios Definition - After determining the different possibilities for each type of SSI, a hybrid and minimally invasive solution will be envisioned, specified, developed and tested, including existing hardware components and new software solutions, and targeting a universal interface that includes elderly people. The specific limitations and requirements imposed by an elderly speaker need to be stipulated based on a pre-defined user profile in order to provide an efficient use of the interface. During the full span of the project duration, close contact with end-users will be sought, starting from user requirements' capture to the adoption of a full usability evaluation methodology, which will collect feedback and draw conclusions based on real subjects while interacting (using SSI) with computing systems and smartphones, respectively, in real case indoor home scenarios and in mobility environments.

Usability Evaluation - Usability evaluation will be conducted, considering different groups of users. The usability evaluation will be focused on real case indoor home scenarios. This evaluation will also include a comparison study, similar to the ones described in (Freitas et al., 2009) towards traditional interfaces such as, mouse and keyboard.

Fulfilling these objectives, even partially, will contribute to expanding knowledge in different areas of research. The used methodology will be based in state-of-the-art assessment, analytical modelling of the proposed solutions, specification, development, test and concrete deployment of algorithms and software systems, including also external sourcing of hardware components in the specified use cases and usability evaluation of such cases with end-users.

3 RESEARCH PROBLEM

An SSI performs ASR in the absence of an intelligible acoustic signal and can be used as a human-computer input modality in high-background-noise environments such as, living rooms, or in aiding speech-impaired individuals which are unable to benefit from the current HCI systems based on speech. By acquiring sensor data from elements of the human speech production process – from glottal and articulators activity, their neural pathways or the brain itself – an SSI produces an alternative digital representation of speech, which can be recognized and interpreted as data, synthesized directly or routed into a communications network. Informally, one can say that a SSI extends the human speech production model by the signal data of electrodes, ultrasonic receivers, cameras and other sources. This provides a more natural approach than currently available speech pathology solutions like, electrolarynx, tracheo-oesophageal speech, and cursor-based text-to-speech systems (Denby et al., 2010).

Since they are still at an early stage SSI systems aimed at HCI present several problems:

Currently, and to our knowledge, no SSI system exists for European Portuguese, leaving European Portuguese users with speech impairments unable to interact with HCI systems based on speech. Furthermore, no study or analysis has been made regarding the adoption of a new language with distinctive characteristics to this kind of systems, and the problems that may arise from applying existent work to EP are unknown. A particularly relevant characteristic of EP are the nasal sounds, which may pose problems to several SSI modalities.

Another problem with the current SSI modalities is how to achieve satisfactory accuracy rates without a high degree of invasiveness. The notion of a SSI system entails that no audible acoustic signal is available, requiring speech information to be extracted from articulators, facial muscle movement and brain activity. Considering a real world scenario, this often leads to unpractical and invasive solutions due to the difficulty in extracting silent speech information using current technologies.

SSI systems are also not directed for all types of users, especially the elderly, which impose several limitations and requirements. Elderly population individuals have developed resistance to conventional forms of human-computer interaction (Phang et al., 2006) like the keyboard and mouse, therefore making it necessary to test new natural forms of interaction such as silent speech. In

addition, elder people often have difficulties with motor skills due to health problems such as arthritis, so the absence of small and difficult to handle equipment may be presented as an advantage over current solutions. It is also known that due to ageing, senses like vision become less accurate, hence difficulties in the perception of details or important information in conventional graphical interfaces may arise, since current interfaces, most notably in the mobility area, are not designed with these difficulties in mind.

In summary, our research problem addresses SSIs aimed at HCI and four concrete hypothesis can be extracted, as follows:

1. Is it possible to extend/adapt the work on SSI for languages such as English to European Portuguese?
2. Do nasal sounds, particularly relevant in EP, poses problems to most, if not all, of the modalities, and is their detection possible using less invasive SSIs?
3. Does a multimodal approach has the potential to improve state-of-the-art results using several less invasive modalities?
4. Can an SSI be used in a real world scenario, robust enough to be usable, with sufficient user satisfaction, by all users including the elderly?

4 STATE OF THE ART

Several SSI based on different sensory types of data have been proposed in the literature and a detailed overviews can be found in Denby et al., (2009) and Freitas et al., (2011). In this section we summarize the existent approaches grouped according to the human speech production model.

The speech production model can be divided into several stages. According to Levelt (1989), the communicative intention is the first phase of each speech act and consists in converting patterns of goals into messages followed by the grammatical encoding of the preverbal message to surface structure. The next phase of the speech production is the passage from the surface structure to the phonetic plan, which, informally speaking is the sequence of phones that are fed to the articulators. This can be divided between the electrical impulse fed into the articulators and the actual process of articulating. The final phase consists on the consequent effects of the previous phases.

The existent experimental SSI systems described in the literature, cover information extraction from all the stages of speech production, from intention to articulation effects, as depicted on Figure 1. The current approaches can then be divided as follows:

- Intention level (brain / Central Nerve System): Interpretation of signals from implants in the speech-motor cortex (Brumberg et al., 2010), Interpretation of signals from electro-encephalographic (EEG) sensors (Porbadnigk et al., 2009);
- Articulation control (muscles): Surface Electromyography of the articulator muscles (Schultz and Wand, 2010);
- Articulation (articulators): Capture of the movement of fixed points on the articulators using Electromagnetic Articulography (EMA) sensors (Fagan et al., 2008); Real-time characterization of the vocal tract using ultra-sound (US) and optical imaging of the tongue and lips (Florescu et al., 2010); Capture movements of a talker's face through ultrasonic Doppler sensing (Srinivasan et al., 2010).
- Articulation effects: Digital transformation of signals from a Non-Audible Murmur (NAM) microphone (Toda et al., 2009); Analysis of glottal activity using electromagnetic (Quatieri et al., 2006), or vibration (Patil and Hansen, 2010) sensors.

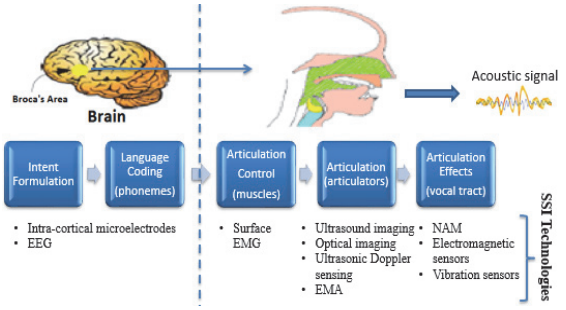

Figure 1: Phased speech production model with the correspondent SSI technologies.

4.1 SSIs for Portuguese

The existing SSI research has been mainly developed for English, with some exceptions for French (Tran et al., 2009) and Japanese (Toda et al., 2009). There was no published work prior to this thesis for European Portuguese in the area of SSIs, although there are previous research on related areas, such as the use of EMA (Teixeira and Vaz, 2001), Electroglotograph and MRI (Martins et al., 2008) for speech production studies, articulatory synthesis (Teixeira and Vaz, 2000) and multimodal

interfaces involving speech (Ferreira et al., 2013). There are also some studies on lip reading systems for EP that aim at robust speech recognition based on audio and visual streams (Pêra et al., 2004); (Sá et al., 2003). However, none of these addresses EP distinctive characteristics, such as nasality.

4.2 Multimodal SSIs

In 2004, Denby and Stone (2004), presented a first experiment where 2 input modalities, in addition to speech audio, were used to develop an SSI. Denby and Stone employed ultrasound imaging of the tongue area, lip profile video and acoustic speech data with the goal of developing an SSI. More recently, Florescu et al., (2010), using these same modalities achieved a 65.3% recognition rate only considering silent word articulation in an isolated word recognition scenario with a 50-word vocabulary using a DTW-based classifier. The reported approach also attributes substantially more importance to the tongue information, only considering a 30% weight during classification for the lip information. In 2008, Tran et al. (2008), also reported a preliminary approach using information from 2 modalities: whispered speech acquired using a NAM and visual information of the face using the 3D position of 142 coloured beads glued to the speakers face. Later, using the same modalities, the same author, achieved an absolute improvement of 13.2% when adding the visual information to the NAM data stream. The use of visual facial information combined with sEMG signals has also been proposed by Yau et al., (2008). In this study Yau et al. presents an SSI that analyses the possibility of using sEMG for unvoiced vowels recognition and a vision-based technique for consonant recognition. When looking at the chosen modalities, recent work using video plus depth information has been presented by Galatas et al., (2012), showing that the depth facial information can improve the system performance over audio-only and traditional audio-visual systems. In the area of sEMG-based SSIs, recent research on has been focused on the differences between audible and silent speech and how to decrease the impact of different speaking modes (Wand and Schultz, 2011a); the importance of acoustic feedback (Herff et al., 2011); EMG-based phone classification (Wand and Schultz, 2011b); and session-independent training methods (Wand and Schultz, 2011c). For what UDS is concerned, it has been applied to several areas (e.g. voice activity detection (Kalgaonkar et al., 2007), speaker identification (Kalgaonkar et al., 2008), synthesis (Toth et al., 2010) and speech recognition with promising results (Srinivasan et al., 2010); (Freitas et al., 2012).

4.3 Nasality Detection

The production of a nasal sound involves air flow through the oral and nasal cavities. This air passage for the nasal cavity is essentially controlled by the velum that, when lowered, allows for the velopharyngeal port to be open, enabling resonance in the nasal cavity and the sound to be perceived nasal. The production of oral sounds occurs when the velum is raised and the access to the nasal cavity is closed (Beddor, 1993). The process of moving the soft palate involves the several muscles (Fritzell, 1969); (Hardcastle, 1976); (Seikel et al., 2010), as depicted in Figure 2.

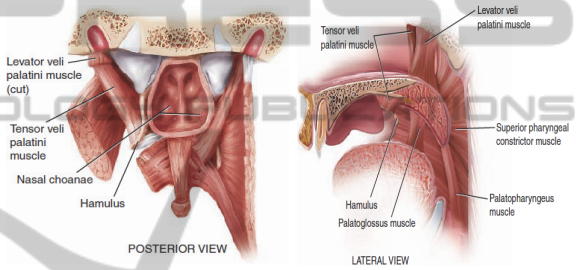

Figure 2: Muscles of the soft palate from posterior (left), and the side (right) view (Seikel et al., 2010).

In previous studies, the application of EMG to measure the level of activity of these muscles has been performed by means of intramuscular electrodes (Fritzell, 1969); (Bell-Berti, 1976) and surface electrodes positioned directly on the oral surface of the soft palate (Lubker, 1968); (Kuehn, 1982). Our work differs from the cited papers, since none of them uses surface electrodes placed in the face and neck regions, a significantly less invasive approach and quite more realistic and representative of the SSIs case scenarios. Also, although intramuscular electrodes may offer more reliable myoelectric signals, they also require considerable medical skills and, for both reasons, intramuscular electrodes were discarded for this study.

No literature exists in terms of detecting the muscles involved in the velopharyngeal function with surface EMG electrodes placed on the face and neck. Previous studies in the lumbar spine region have shown that if proper electrode positioning is considered a representation of deeper muscles can be acquired (McGill et al., 1996) thus raising a question that is currently unanswered: is surface EMG positioned in the face and neck regions able to detect

activity of the muscles related to nasal port opening/closing and consequently detect the nasality phenomena? Another related question that can be raised is how we can show, with some confidence, that the signal we are seeing is in fact the myoelectric signal generated by the velum movement and not spurious movements caused by neighbouring muscles unrelated to the velopharyngeal function.

5 METHODOLOGY

The chosen approach to the mentioned research problems can be divided into 4 main stages, as depicted in Figure 3. The following subsections describes each stage in more detail.

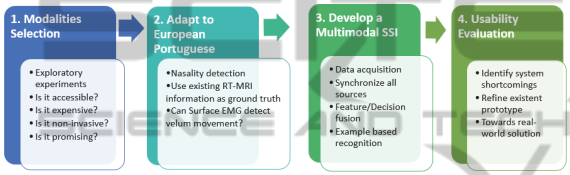

Figure 3: Stages of the chosen methodology.

5.1 Modalities Selection

In the initial stage of the PhD, an in depth study of related work and preliminary evaluation of different types of SSIs was made, the major problems were identified and the aim of the thesis was defined. This initial study contributed to determine which SSI or SSIs were more suited to the problem and what were the available resources. Results of this work include a state-of-the-art assessment as well as the main issues to be solved. Then by conducting preliminary experiments with non-invasive and recent modalities such as Ultrasonic Doppler (Freitas et al., 2012), we have selected several HCI technologies based on: the possibility of being used in a natural manner without complex medical procedures from the ethical and clinical perspectives, low cost, tolerance to noisy environments and be able to work with speech-handicapped users or elderly people, for whom speaking requires a substantial effort.

5.2 Adapt to European Portuguese

In the second stage of the PhD we chosen to address a known challenge in SSIs – the detection of nasality. This decision was motivated by the fact that nasality is an important characteristic of EP and an eventual solution for this problem would allow the development of a more adapted SSI for EP, and also because preliminary studies have shown current techniques for silent speech recognition based on sEMG will present a degraded performance when dealing with languages with nasal characteristics. Thus, we started by exploring the existence of useful information about the velum movement and also by assessing if deeper muscles could be sensed using surface electrodes in the regions of the face and neck and the best electrode location to do so. To accomplish these tasks, we have applied a procedure that uses Real-Time Magnetic Resonance Imaging (RT-MRI), collected from the same speakers, providing a method to interpret EMG data.

The main idea behind this approach consists in crossing two types of data containing information about the velum movement: (1) images collected using RT-MRI and (2) the myoelectric signal collected using surface EMG sensors. By combining these two sources, ensuring compatible scenario conditions and proper time alignment, we are able to accurately estimate the time when the velum moves and the type of movement (i.e. ascending or descending) under a nasality phenomenon, and establish the differences between nasal and oral vowels using surface EMG. Also, we need to know when the velum is moving, to avoid that signals coming from other muscles, artefacts and noise be misinterpreted as signals coming from the target muscles. To overcome this problem we take advantage of a previous data collection based on RT-MRI (Teixeira et al., 2012), which provides an excellent method to interpret EMG data and estimate when velum is moving.

Recent advances in MRI technology allow real-time visualization of the vocal tract with an acceptable spatial and temporal resolution. This sensing technology enables us to have access to real time images with relevant articulatory information for our study, including velum raising and lowering. In order to make the correlation between the two signals, audio recordings were performed in both data collections by the same speakers. Notice that EMG and RT-MRI data can't be collected together, so the best option is to collect the same corpus for the same set of speakers, at different times, reading the same prompts in EMG and RT-MRI.

For the EMG and RT-MRI signals synchronization we start by aligning both EMG and the information extracted from the RT-MRI with the corresponding audio recordings. Next, we apply Dynamic Time Warping (DTW) to the signals, finding the optimal match between the two sequences. Based on the DTW result we map the

information extracted from RT-MRI from the original production to the EMG time axis, establishing the needed correspondence between the EMG and the RT-MRI information, as depicted in Figure 4.

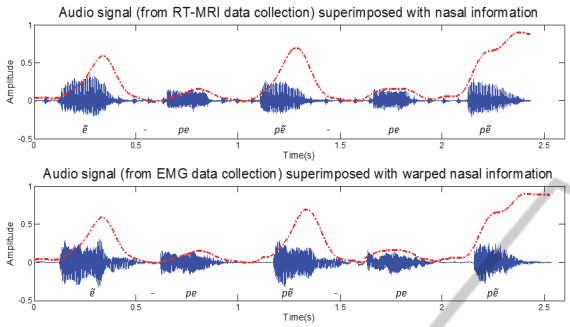

Figure 4: Exemplification of the warped signal representing the nasal information extracted from RT-MRI (dashed red line) superimposed on the speech recorded during the corresponding RT-MRI and EMG acquisition, for the sentence [ɐ̃pɐ, pɐ̃pɐ, pɐ̃].

5.3 Multimodal Ssi

The third stage consists in the development of a multimodal SSI prototype based on the conclusions of the previous stages. Since no SSI of this kind exists the first step is to collect data from the selected modalities in a synchronized way. Having collected the necessary data a feature selection analysis must be conducted due to the high number of information streams (5 if we consider audio) in order to avoid an excessively high dimensionality space. During this analysis it is also necessary to understand how we are going to combine all these streams and what type of fusion should be used (i.e. feature or decision fusion, or both) in our pipeline. Finally, it is necessary to understand what classifier (or classifiers) is most appropriate for this case. Since not much data exists for this novel interface, our aim is to explore example-based methods such as Dynamic Time Warping.

To the current time we have collected data from four modalities with the following specifications: (1) video input, which captures the RGB colour of each image pixel of the speakers' mouth region and its surroundings, including chin and cheeks; (2) depth input, which captures depth information of each pixel for the same areas, providing useful information about the mouth opening and tongue position, in the sensor reference frame, in some cases; (3) surface EMG sensory data, which provides information about the myoelectric signal produced by the targeted facial muscles during speech movements; (4) Ultrasonic Doppler Sensing, a technique which is based on the emission of a pure tone in the ultrasound range towards the speaker's face, that is received by an ultrasound sensor tuned to the transmitted frequency. The reflected signal then contains Doppler frequency shifts that correlate with the movements of the speaker's face (Srinivasan et al., 2010). To the best of our knowledge, this is the first silent speech corpus that combines more than two input data types and the first to synchronously combine the corresponding four modalities, thus, providing the necessary information for future studies on multimodal SSIs.

After assembling all the necessary data collection equipment which, in the case of ultrasound, led us to the development of custom built equipment based on the work of Zhu (2008), we needed to create the necessary conditions to record all signals with adequate synchronization. The challenge of synchronizing all signals resided in the fact that a potential synchronization event would need to be captured simultaneously by all (four) input modalities. To that purpose, we have selected the EMG recording device, which had an available I/O channel, as the source that generates the alignment pulse for all the remaining modalities. After the data collection system setup was ready, a proof-of-concept database, was collected for further analysis.

The devices employed in this data collection were: (1) a Microsoft Kinect (2013) that acquires visual and depth information; (2) an sEMG sensor acquisition system from Plux (2013), that captures the myoelectric signal from the facial muscles; (3) a custom built dedicated circuit board (referred to as UDS device), that includes: 2 ultrasound transducers (400ST and 400SR working at 40 kHz), a crystal oscillator at 7.2 MHz and frequency dividers to obtain 40 kHz and 36 kHz, and all amplifiers and linear filters needed to process the echo signal (Freitas et al., 2012).

The Kinect sensor was placed at approximately 0.7m from the speaker. It was configured, using Kinect SDK 1.5, to capture a colour video stream with a resolution of 640x480 pixel, 24-bit RGB at 30 frames per second and a depth stream, with a resolution of 640x480 pixel, 11-bit at 30 frames per second.

The sEMG acquisition system consisted of 5 pairs of EMG surface electrodes connected to a device that communicates with a computer via Bluetooth. As depicted in Figure 5 the sensors were attached to the skin using a single use 2.50 cm diameter clear plastic self-adhesive surfaces and considering an approximate 2.00 cm spacing

between the electrodes center for bipolar configurations. Before placing the surface EMG sensors, the sensor location was previously cleaned with alcohol. While uttering the prompts no other movement, besides the one associated with speech production, was made. The five electrode pairs were placed in order to capture the myoelectric signal from the following muscles: the *levator angulis oris* (channel 2); *zygomaticus major* (channel 2); the *tongue* (channel 1 and 5), the *anterior belly of the digastric* (channel 1); the *platysma* (channel 4) and the last electrode pair was placed below the ear between the mastoid process and the mandible. The sEMG channels 1 and 4 used a monopolar configuration (i.e. placed one of the electrodes from the respective pair in a location with low or negligible muscle activity), being the reference electrodes placed on the mastoid portion of the temporal bone. The positioning of the EMG electrodes 1, 2, 4 and 5 was based on previous work (e.g. Schultz and Wand, 2010) and sEMG electrode 3 was placed according to recent findings by the authors about the detection of nasality in SSIs (Freitas et al., 2014).

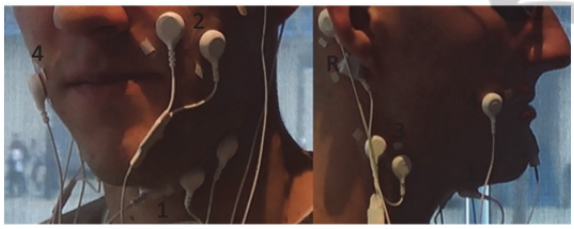

Figure 5: sEMG electrodes positioning and the respective channels (1 to 5) plus the reference electrode (R).

The UDS device was placed at approximately 40.0 cm from the speaker and was connected to an external sound board (Roland, UA-25 EX) which in turn is connected to the laptop through a USB connection. The two supported recording channels of the external sound board were connected to the I/O channel of the sEMG recording device and to the UDS device. The Doppler echo and the synchronization signals were sampled at 44.1 kHz and to facilitate signal processing, a frequency translation was applied to the carrier by modulating the echo signal by a sine wave and low passing the result, obtaining a similar frequency modulated signal centered at 4 kHz.

In order to register all input modalities via time alignment between all corresponding four input streams, we have used an I/O bit flag in the sEMG recording device, which has one input switch for debugging purposes and two output connections, as depicted in Figure 6. Synchronization occurs when the output of a synch signal, programmed to be automatically emitted by the sEMG device at the beginning of each prompt, is used to drive a led and to provide an additional channel in an external sound card. Registration between the video and depth streams is ensured by the Kinect SDK. Using the information from the led and the auxiliary audio channel with synch info, the signals were time aligned offline. To align the RGB video and the depth streams with the remaining modalities, we have used an image template matching technique that automatically detects the led position on each colour frame. For the UDS acquisition system, the activation of the output I/O flag of the sEMG recording device, generates a small voltage peak on the signal of the first channel. To enhance and detect that peak, a second degree derivative is applied to the signal followed by an amplitude threshold. To be able to detect this peak, we have previously configured the external sound board channel with maximum input sensitivity. The time-alignment of the EMG signals is ensured by the sEMG recording device, since the I/O flag is recorded in a synchronous way with the samples of each channel.

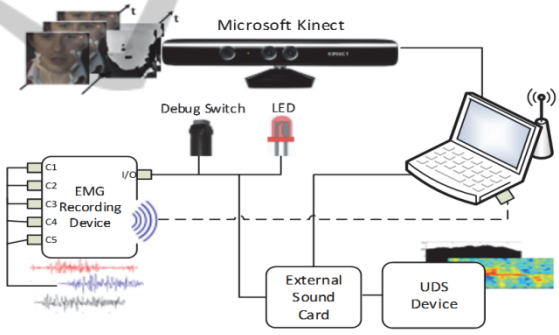

Figure 6: Diagram of the time alignment scheme showing the I/O channel connected to the three outputs – debug switch, external sound card and a directional led.

5.4 Usability Evaluation

The last stage of the PhD will be focused on evaluating the interface usability. Here, usability evaluation tests using the proposed SSI prototype will be conducted, allowing to identify shortcomings, refine previously established user requirements and improve the existent prototype. In a first phase of the usability tests, the features provided by the interface will be tested in the form of a task that the user must accomplish. For each subject it will be analysed if the task was accomplished; how many tries were required; if the

application flow ran smoothly and how long the user took to adapt to the system. In these tests, usability should be evaluated in terms of efficiency, effectiveness and satisfaction. Concerning efficiency, the required time to execute a task using the system, the number of actions and the time spent with application instructions, should be considered. In terms of effectiveness, it should be measured if the task was completed with success, how frequent it recurs to application features and the quality of the output. Finally, one should assess if the user enjoyed and presented a positive attitude towards the system.

6 EXPECTED OUTCOME

So far, this PhD thesis have spawned the following contributions:

- A new taxonomy that associates each type of SSI to a stage of the human speech production model.
- A state-of-the-art overview of SSIs, which includes the latest related research.
- SSI technologies and techniques applied for the first time to EP (e.g. sEMG and UDS)
- Results that indicate the difficulty on distinguishing minimal pairs of words that only differ on nasal sounds when using surface EMG or Video.
- Analysis of velum movement detection using surface electrodes in the regions of the face and neck and the best electrode location to do so.
- A silent speech *corpus* that combines more than two input data types and the first to synchronously combine the corresponding modalities.

Upon completion we expect to obtain the following outcomes:

- Development of a multimodal SSI prototype based on Video, Depth, UDS, and sEMG where eventually the weakest points of one modality can be minored by other(s).
- A careful analysis of what modalities to fuse, when and how, in order to provide adequate response to users' goals and context, striving for additional robustness in situations, such as noisy environments, or where privacy issues and existing disabilities might hinder single modality interaction.
- Assess the usability of the proposed system in real-world scenarios.

REFERENCES

Beddor, P. S., 1993. The perception of nasal vowels. In *M. K. Huffman and R. A. Krakow, Nasals, Nasalization, and the Velum, Phonetics and Phonology*, Academic Press Inc., Vol. 5, pp. 171-196.

Bell-Berti, F., 1976. An Electromyographic Study of Velopharyngeal Function, *Speech Journal of Speech and Hearing Research*, Vol.19, pp. 225-240.

Brumberg, J. S., Nieto-Castanonf, A., Kennedye, P. R. and Guenther, F. H., 2010. Brain–computer interfaces for speech communication. *Speech Communication*, Vol. 52, Issue 4, pp. 367-379.

Denby, B. and Stone, M., 2004. Speech synthesis from real time ultrasound images of the tongue, *Internat. Conf. on Acoustics, Speech, and Signal Processing*, Montreal, Canada, Vol. 1, pp. I685–I688.

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J.M. and Brumberg, J.S, 2010. Silent speech interfaces. *Speech Communication*, Vol. 52, Issue 4, pp. 270-287.

Dias, M. S., Bastos, R., Fernandes, J., Tavares, J. and Santos, P., 2009. Using Hand Gesture and Speech in a Multimodal Augmented Reality Environment, *GW2007*, LNAI 5085, pp.175-180.

Fagan, M. J., Ell, S. R., Gilbert, J. M., Sarrazin, E. and Chapman, P.M, 2008. Development of a (silent) speech recognition system for patients following laryngectomy. *Med. Eng. Phys.*, Vol. 30, Issue 4, pp. 419-425.

Ferreira, F., Almeida, N., Casimiro, J., Rosa, A. F., Oliveira, A., and Teixeira, A. 2013. Multimodal and Adaptable Medication Assistant for the Elderly CISTI'2013 (8th Iberian Conference on Information Systems and Technologies).

Florescu, V-M., Crevier-Buchman, L., Denby, B., Hueber, T., Colazo-Simon, A., Pillot-Loiseau, C., Roussel, P., Gendrot, C. and Quattrochi, S., 2010. Silent vs Vocalized Articulation for a Portable Ultrasound-Based Silent Speech Interface. *Proceedings of Interspeech 2010*, Makuari, Japan.

Freitas, J. Teixeira, A. Dias M. S. and Bastos, C., 2011. Towards a Multimodal Silent Speech Interface for European Portuguese, *Speech Technologies*, Ivo Ipsic (Ed.), InTech.

Freitas, J. Teixeira, A., Vaz, F. and Dias, M.S., 2012. Automatic Speech Recognition based on Ultrasonic Doppler Sensing for European Portuguese, *Advances in Speech and Language Technologies for Iberian Languages*, vol. CCIS 328, Springer.

Freitas, J., Calado, A., Barros, M. J. and Dias, M. S., 2009. Spoken Language Interface for Mobile Devices. *Human Language Technology. Challenges of the Information Society Lecture Notes in Computer Science*, Vol. 5603, pp. 24-35.

Freitas, J., Teixeira, A., Silva, S., Oliveira, C. and Dias, M.S., 2014. Velum Movement Detection based on Surface Electromyography for Speech Interface. *Conference on Bio-inspired Systems and Signal Processing, Biosignals 2014*, Angers, France.

Fritzell, B., 1969. The velopharyngeal muscles in speech: an electromyographic and cineradiographic study. *Acta Otolaryngolica*. Suppl. 50.

Galatas, G., Potamianos, G., Makedon, F., 2012. Audio-visual speech recognition incorporating facial depth information captured by the Kinect. *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pp. 2714-2717.

Hardcastle, W. J., 1976. *Physiology of Speech Production - An Introduction for Speech Scientists*. Academic Press, London.

Herff, C. Janke, M. Wand, M. and Schultz, T., 2011. Impact of Different Feedback Mechanisms in EMG-based Speech Recognition. In *Proceedings of Interspeech 2011*. Florence, Italy.

Kalgaonkar, K., Raj B., Hu., R., 2007. Ultrasonic doppler for voice activity detection. *IEEE Signal Processing Letters*, vol.14, Issue 10, pp. 754–757.

Kalgaonkar, K., Raj., B., 2008. Ultrasonic doppler sensor for speaker recognition. *Internat. Conf. on Acoustics, Speech, and Signal Processing*.

Kuehn D.P., Folkins JW, Cutting CB., 1982. Relationships between muscle activity and velar position, *Cleft Palate Journal*, Vol. 19, Issue 1, pp. 25-35.

Levelt. W., 1989. *Speaking: from Intention to Articulation*. Cambridge, Mass.: MIT Press.

Lubker, J. F., 1968. An electromyographic-cinefluorographic investigation of velar function during normal speech production. *Cleft Palate Journal*, Vol. 5, Issue 1, pp. 17.

Martins, P. Carbone, I. Pinto, A. Silva, A. and Teixeira, A., 2008. European Portuguese MRI based speech production studies. *Speech Communication*. NL: Elsevier, Vol.50, No.11/12, ISSN 0167-6393, pp. 925–952.

McGill, S., Juker, D. and Kropf, P., 1996. Appropriately placed surface EMG electrodes reflect deep muscle activity (psoas, quadratus lumborum, abdominal wall) in the lumbar spine. In *Journal of Biomechanics*, Vol. 29 Issue, 11, pp. 1503-7.

Microsoft Kinect, *Online: http://www.xbox.com/en-US/kinect*, accessed on 9 December 2013.

Patil, S. A. and Hansen, J. H. L., 2010. The physiological microphone (PMIC): A competitive alternative for speaker assessment in stress detection and speaker verification. *Speech Communication*. Vol. 52, Issue 4, pp. 327-340.

Pêra, V. Moura, A. and Freitas, D. 2004. LPFAV2: a new multi-modal database for developing speech recognition systems for an assistive technology application. In *SPECOM-2004*, pp. 73-76.

Phang, C. W., Sutanto, J., Kankanhalli, A., Li, Y., Tan, B. C. Y, and Teo, H. H., 2006. Senior citizens' acceptance of information systems: A study in the context of e-government services. *IEEE Transactions On. Engineering Management*, Vol. 53, Issue 4, pp. 555–569, 2006.

Plux Wireless Biosignals, Portugal, *Online: http://www.plux.info/*, accessed on 9 December 2013.

Porbadnigk, A., Wester, M., Calliess, J. and Schultz, T., 2009. EEG-based speech recognition impact of temporal effects. *International Conference on Bio-inspired Systems and Signal Processing, Biosignals 2009*, Porto, Portugal, pp.376–381.

Quatieri, T. F., D. Messing, K. Brady, W. B. Campbell, J. P. Campbell, M. Brandstein, C. J. Weinstein, J. D. Tardelli and P. D. Gatewood, 2006. Exploiting non-acoustic sensors for speech enhancement. *IEEE Trans. Audio Speech Lang. Process*, Vol. 14, Issue 2, pp. 533–544.

Rossato, S. Teixeira, A. and Ferreira, L., 2006. Les Nasales du Portugais et du Français: une étude comparative sur les données EMMA. *In XXVI Journées d'Études de la Parole*. Dinard, France.

Sá, F. Afonso, P. Ferreira, R. and Pera, V., 2003. Reconhecimento Automático de Fala Contínua em Português Europeu Recorrendo a Streams Audio-Visuais. In *The Proceedings of COOPMEDIA'2003 - Workshop de Sistemas de Informação Multimédia, Cooperativos e Distribuídos*, Porto, Portugal.

Schultz, T. and Wand, M., 2010. Modeling coarticulation in large vocabulary EMG-based speech recognition. *Speech Communication*, Vol. 52, Issue 4, pp. 341-353.

Seikel, J. A., King, D. W., Drumright, D. G., 2010. *Anatomy and Physiology for Speech, Language, and Hearing*, 4rd Ed., Delmar Learning.

Srinivasan, S., Raj, B. and Ezzat, T., 2010. Ultrasonic sensing for robust speech recognition. *Internat. Conf. on Acoustics, Speech, and Signal Processing 2010*.

Teixeira, A. and Vaz, F., 2000. Síntese Articulatória dos Sons Nasais do Português. *Anais do V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada* (PROPOR), ICMC-USP, Atibaia, São Paulo, Brasil, 2000, pp. 183-193.

Teixeira, A. and Vaz, F., 2001. European Portuguese Nasal Vowels: An EMMA Study. *7th European Conference on Speech Communication and Technology, EuroSpeech – Scandinavia*, pp. 1843-1846.

Teixeira, A., Braga, D., Coelho, L., Fonseca, J., Alvarelhão, J., Martín, I., Queirós, A., Rocha, N., Calado, A. and Dias, M. S., 2009. Speech as the Basic Interface for Assistive Technology. *DSAI 2009 - Proceedings of the 2th International Conference on Software Development for Enhancing Accessibility and Fighting Info-Exclusion*, Porto Salvo, Portugal.

Teixeira, A., Martins, P., Oliveira, C., Ferreira, C., Silva, A., Shosted, R., 2012. "Real-time MRI for Portuguese: database, methods and applications", *Proceedings of PROPOR 2012*, LNCS vol. 7243. pp. 306-317.

Toda, T., Nakamura, K., Nagai, T., Kaino, T., Nakajima, Y., and Shikano, K., Technologies for Processing Body-Conducted Speech Detected with Non-Audible Murmur Microphone. *Proceedings of Interspeech 2009*, Brighton, UK.

Toth, A. R., Kalgaonkar, K., Raj, B., Ezzat, T., 2010. Synthesizing speech from Doppler signals, *Internat. Conference on Acoustics Speech and Signal Processing*, pp.4638-4641.

Tran, V.-A Bailly, G. Loevenbruck, H. and Toda, T., 2009. Multimodal HMM-based NAM to-speech conversion. In *Proceedings of Interspeech 2009*, Brighton, UK.

Wand, M. and Schultz, T., 2011. Analysis of Phone Confusion in EMG-based Speech Recognition, *Internat. Conf. on Acoustics, Speech and Signal Processing 2011*, Prague, Czech Republic.

Wand, M. and Schultz, T., 2011. Investigations on Speaking Mode Discrepancies in EMG-based Speech Recognition, *Proceedings of Interspeech 2011*, Florence, Italy.

Wand, M. and Schultz, T., 2011. Session-Independent EMG-based Speech Recognition, *International Conference on Bio-inspired Systems and Signal Processing, Biosignals 2011*, Rome, Italy.

Yau, W. C., Arjunan, S. P. and Kumar, D. K., 2008. Classification of voiceless speech using facial muscle activity and vision based techniques, *TENCON 2008-2008 IEEE Region 10 Conference*, 2008.

Zhu, B., 2008. Multimodal speech recognition with ultrasonic sensors. *Master's thesis*. Massachusetts Institute of Technology, Cambridge, Massachusetts.