Multimodal Corpora for Silent Speech Interaction

João Freitas[1,2], António Teixeira[2], Miguel Sales Dias[1,3]
[1]Microsoft Language Development Center, Lisboa, Portugal
[2]Dep. Electronics Telecommunications & Informatics/IEETA, University of Aveiro, Portugal
[3]ISCTE-Lisbon University Institute, Lisboa, Portugal
E-mail: t-joaof@microsoft.com, ajst@ua.pt, miguel.dias@microsoft.com

Abstract

A Silent Speech Interface (SSI) allows for speech communication to take place in the absence of an acoustic signal. This type of interface is an alternative to conventional Automatic Speech Recognition which is not adequate for users with some speech impairments or in the presence of environmental noise. The work presented here produces the conditions to explore and analyze complex combinations of input modalities applicable in SSI research. By exploring non-invasive and promising modalities, we have selected the following sensing technologies used in human-computer interaction: Video and Depth input, Ultrasonic Doppler sensing and Surface Electromyography. This paper describes a novel data collection methodology where these independent streams of information are synchronously acquired with the aim of supporting research and development of a multimodal SSI. The reported recordings were divided into two rounds: a first one where the acquired data was silently uttered and a second round where speakers pronounced the scripted prompts in an audible and normal tone. In the first round of recordings, a total of 53.94 minutes were captured where 30.25% was estimated to be silent speech. In the second round of recordings, a total of 30.45 minutes were obtained and 30.05% of the recordings were audible speech.

Keywords: Silent Speech, Multimodal HCI, Data Collection

1. Introduction

Silent Speech designates the process of speech communication in the absence of an audible and intelligible acoustic signal (Denby et al., 2009). By extracting information of the human speech production process, an SSI is able to interpret and process the acquired data. Several SSI based on different sensory types of data have been proposed in the literature (e.g. Electro-encephalographic sensors (Porbadnigk et al., 2009), Electromagnetic Articulography sensors (Fagan et al. 2008), etc.). Nonetheless, acquiring data from a single input modality limits the amount of useful information available for capture and further processing. Furthermore, in order to develop a multimodal SSI, it is necessary to collect data from multiple input modalities in a synchronous way due to the nonexistence of SSI multimodal data available for research. However, satisfying the requirements and gathering all the necessary equipment for collecting such corpora is a complex and cumbersome task (Hueber et al., 2007). Hence, the availability to the community of multimodal corpora would not only allow to increase the number of data resources accessible for further research, but would also pave the way for the development of a multimodal SSI, which could provide a more complete representation of the speech production model behavior during speech.

The work presented in this paper, creates the conditions to explore and analyze more complex combinations of input modalities for SSI research. By exploring non-invasive and state-of-the-art modalities such as Ultrasonic Doppler (Srinivasan et al., 2010), we have selected several sensing technologies based on: the possibility of being used in a natural manner without complex medical procedures from the ethical and clinical perspectives, low cost, tolerant to noisy environments and able to work with speech-handicapped users or elderly people, for whom speaking requires a substantial effort. Based on these requirements, we collected data from four SSI modalities with the following specifications: (1) video input, which captures the RGB color 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 (resulting in our this case, in a 3D point cloud in the sensor reference frame, represented by a grayscale image), providing useful information about the mouth opening and tongue position, in some cases; (3) surface EMG (sEMG) sensory data, which provides information about the myoelectric signal produced by the targeted facial muscles during speech movements; (4) Ultrasonic Doppler Sensing (UDS), 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).

In literature several studies that combined 2 input modalities, in addition to audio can be found (e.g. Denby and Stone, (2004) and Tran et al. (2008)). Nonetheless, to the best of our knowledge, this is the first silent speech corpora 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 and research on multimodal SSIs.

2. Data Collection Setup

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 sEMG recording device, which had an available output channel, as the source that generates the alignment pulse for all the remaining modalities. After the data collection system setup was ready, a database described in this paper, was collected for further analysis.

2.1 The individual data input modalities

The devices employed in this data collection, depicted in Figure 1, were: (1) a Microsoft Kinect for Windows that acquires visual and depth information; (2) an sEMG sensor acquisition system from Plux (2014), 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 70cm from the speaker. It was configured, using Kinect SDK 1.5, to capture a color 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 to code the *Z* dimension, at 30 frames per second. Kinect was also configured to use the Near Depth range (i.e. range between 40cm to 300cm) and to track a seated skeleton.

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 2, the sensors were attached to the skin using a single use 2.50cm diameter clear plastic self-adhesive surfaces and considering an approximate 2.00cm 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 *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 from channel 3 was placed according to recent findings by the authors about the detection of nasality in SSIs (Freitas et al., 2014), a distinct characteristic of European Portuguese (EP) (Strevens, 1954).

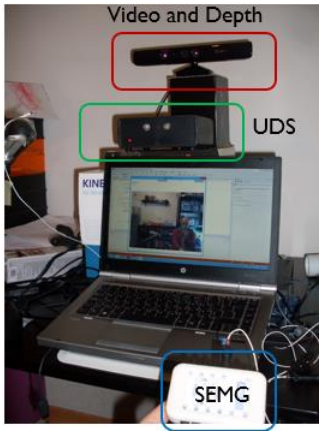

Figure 1: Acquisition devices and laptop with the data collection application running.

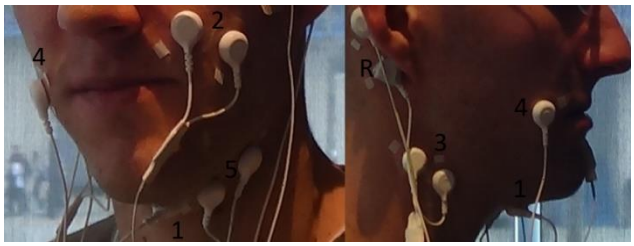

Figure 2: Surface EMG electrodes positioning and the respective channels (1 to 5) plus the reference electrode (R).

The Ultrasonic Doppler sensing device was placed at approximately 40cm from the speaker and was connected to an external sound board (Roland, UA-25 EX in the first setup and a TASCAM US-1641 in the second setup) which in turn was connected to the laptop through a USB connection. Two 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.

2.2 Registration of all input modalities

In order to register all the mentioned input modalities via time alignment between all corresponding 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 4. 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 color 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.

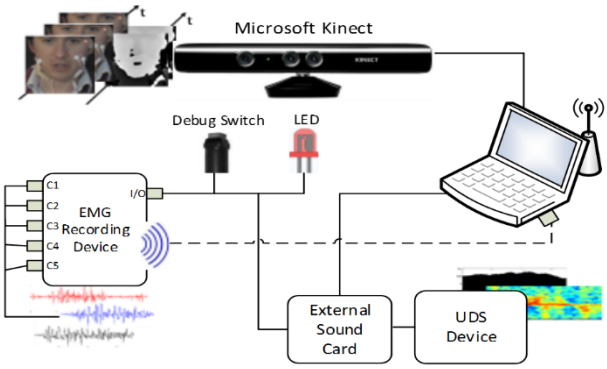

Figure 4: Diagram of the time alignment scheme showing the I/O channel connected to the three outputs – debug switch, external sound card and a directional led.

2.3 Acquisition Methodologies

The recordings took place in a quiet room with controlled illumination and an assistant responsible for monitoring the data acquisition and also for pushing a record/stop button in the recording tool interface in order to avoid unwanted muscle activity.

The data acquisition is divided into two distinct rounds hereon referred as first and second round of recordings. The main difference between them is the acquisition of an audible acoustic signal (second round) versus silently articulating the words (first round).

The first round of our database contains the recordings of 9 sessions of 8 native EP speakers (one speaker recorded two sessions) - 2 female and 6 male – with no history of hearing or speech disorders, with an age range from 25 to 35 years old and an average age of 30 years. Due to hardware limitations and the differences found between silently articulated speech and audible uttered speech related with the lack of acoustic feedback (Wand and Schultz, 2011), in this first round we have chosen to record only silent speech. Thus, no audible acoustic signal was produced by the speakers during the recordings and only one speaker had past experience with silent articulation.

In a second round, the previous sound card was replaced by a TASCAM US-1641, as depicted in Figure 3, and for comparison purposes and by taking advantage of the extra input channels provide by this device, we decided to collect a second round of recordings where the audio channel from the UDS device is also synchronously acquired. As such, we have collected in this round 3 speakers, one from the previous data collection and two elderly speakers without any history of speech disorders known so far and also native EP speakers. The first speaker was a male with 31 years old and the two elderly speakers, were two female with 65 and 71 years old, respectively. In this second stage of data collection, each speaker recorded two sessions without removing the EMG electrodes or changing the recording position.

Figure 3: TASCAM US-1641 device used in the second round of recordings.

Before each recording session, the participants received a 30 minute briefing that included instructions, speaker preparation and voluntarily signing of a consent form which accurately described the experiment and its duration and what kind of data was going to be collected.

Each recording session took between 40 to 60 minutes, generating an average 3.81GB of data per speaker, that includes: session metadata, such as devices configuration; RBG and depth information of a 128x128 pixel square centered at the mouth center and the coordinates of 100 facial points, in the sensor reference frame, for each Kinect image; sEMG data from the 5 available channels; two channel wave per prompt containing the UDS and the synchronization signals; and a compressed video of the whole session. In the second round of recordings, we recorded a three channel wave containing the audio from the UDS device microphone, the Ultrasonic and the synchronization signal.

2.4 Corpora

For this data collection we have selected a vocabulary of 32 EP words, which can be divided into 3 distinct sets. The first set, used in previous literature work for other languages (e.g. (Srinivasan et al., 2010) and for EP in prior work of the authors (Freitas et al., 2012), consists of 10 digits from zero to nine. The second set contains 4 minimal pairs of common words in EP that only differ on nasality of one of the phones (minimal pairs regarding this characteristic, e.g. Cato/Canto [katu]/[kɐ̃tu] or Peta/Penta [petɐ]/[pɐ̃tɐ] – see (Freitas et al., 2011) for more details), and is directly related with previous investigation by the authors on the detection of nasality with SSIs. Table 1 shows the last (third) set, with 14 common words in EP, taken from context free grammars of an Ambient Assisted Living (AAL) application that supports speech input and chosen based on past experiences of the authors (Teixeira et al., 2012). A total of 99 scripted prompts per session were presented to the speaker (three additional silence prompts were also included in the beginning, middle and

end of the session), in a random order with each prompt being pronounced individually, in order to allow isolated word recognition. All prompts were repeated 3 times per recording session.

Ambient Assisted Living Word Set				
Videos (*Videos*)	Ligar (*Call/Dial*)	Contatos (*Contacts*)	Mensagens *(Messages)*	Voltar *(Back)*
Pesquisar *(Search)*	Anterior (*Previous*)	Fotografias (*Photographs*)	Família (Family)	Ajuda *(Help)*
Seguinte *(Next)*	Lembretes (*Reminders*)	Calendário *(Calendar)*	E-Mail *(E-Mail)*	-

Table 1: Set of words of the EP vocabulary, extracted from AAL contexts.

3. Characterization of the Acquired Database

In this section we present some statistics of the acquired data. In the first round of recordings no audio was collected thus an automatic algorithm was used to estimate speech statistics. For the second round of recordings, audible utterances were recorded and the audio was used as auxiliary information for manually annotating the data.

3.1 First Round of Recordings

The data collected in the first round of recordings has a total elapsed duration of 56.11 minutes, with an average duration of 5.99 minutes per session and 3.74 seconds per utterance, not considering silence utterances. By applying a Voice Activity Detection (VAD) technique based on UDS alone, we estimate that 30.25% is silent speech (i.e. continuous facial movements) and that 69.75% is the silence before and after each utterance. The VAD algorithm uses the energy of the UDS pre-processed spectrum information around the carrier and a mean reference value extracted from the silence prompts of each speaker to distinguish silent articulation. Each session presents an average speech duration of 1.81 minutes and 4.18 minutes of non-speech. The female speakers had an average speech duration of 42.79% per session, while this figure for male speakers was only 23.29%. Table 2 details the audio duration of the collected data by word set.

Word Set	Total Recorded Duration (minutes)	Silent Speech	Non-Speech
Digits	15.28	26.78%	73.22%
Nasal Pairs	13.02	28.90%	71.10%
AAL	25.63	33.00%	67.00%
All word sets	53.94	30.25%	69.75%

Table 2: Audio duration, speech time and non-speech time distribution by word set (excluding silence utterances) for the first round of recordings.

3.2 Second Round of Recordings

In the second round of recordings since synchronously acquired audio was available the estimation of the speech and non-speech characteristics was performed based on the manual annotation of the speech signal by the first author. As described in Table 3, in this second round, a total elapsed duration of 30.45 minutes, with an average duration of 5.07 minutes per session and 3.17 seconds per utterance.

Word Set	Total Recorded Duration (minutes)	Speech	Non-Speech
Digits	8.78	28.13%	71.87%
Nasal Pairs	7.48	71.87%	73.13%
AAL	14.20	32.91%	67.09%
All Word Sets	30.45	30.05%	69.95%

Table 3: Audio duration, speech time and non-speech time distribution by word set (excluding silence utterances) for the second round of recordings.

In Table 4 the session statistics for the first and second round are presented. Based on these values, a larger duration of the sessions were only silent speech was considered, can be noticed. This suggests a slower articulation when no acoustic feedback, however it might also be related or influenced by the lack of experience verified in most speakers when articulating the words without any acoustic feedback.

Data Collection Stage	Average Duration per session (minutes)	Average Speech per session (minutes)	Average Non-Speech per session (minutes)
1st round	5.99	1.81	4.18
2nd round	5.07	1.52	3.55

Table 4: Audio duration, speech time and non-speech time distribution by word set (excluding silence utterances) for the second round of recordings.

If instead of estimating the characteristics of the first round based on the automatic algorithm we use the speech/non-speech distribution estimated in the second round. Then, by applying it to the average duration per session of the first round, we get a 1.80 minutes of speech

and 4.19 minutes for non-speech data, a similar result to what was obtained using the UDS algorithm.

4. Conclusion

This paper describes a multimodal data collection with 5 streams of data: Video, Depth, Surface EMG, Ultrasonic Doppler Sensing and audio. By using the surface EMG recording device we were able to synchronously combine these silent speech modalities and acquire information from multiple stages of the human speech production model. The data collection is divided into two rounds of recordings: in a first round only silent speech (i.e. no acoustic signal was produced by the speaker) was recorded; in a second set of recordings, audible speech was captured in addition to the remaining modalities. We have also used an algorithm based on UDS energy for estimating total speech time in the absence of the acoustic signal and some statistics of how the data was distributed.

5. Future Work

The collected data opens several doors in terms of future research. This data will potentially allow for the development of a multimodal SSI based on these modalities, where the strongest points of one modality can eventually help to minimize the weakest point of other(s). It will also allow looking at other types of information, beyond the acoustic signal, for interesting research issues, such as elderly speech characteristics and nasal sounds production and recognition.

6. Acknowledgements

This work was partially funded by Marie Curie Actions Golem (ref.251415, FP7-PEOPLE-2009-IAPP) and IRIS (ref. 610986, FP7-PEOPLE-2013-IAPP), by FEDER through the Program COMPETE under the scope of QREN 5329 FalaGlobal and by National Funds (FCT-Foundation for Science and Technology) in the context of IEETA Research Unit funding FCOMP-01-0124-FEDER-022682 (FCT-PEst-C/EEI/UI0127/2011). The authors would also like to thank the experiment participants.

7. References

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J.M., Brumberg, J.S. (2009). Silent speech interfaces. *Speech Communication*, 52(4), pp. 270--287.

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

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.*, 30(4), pp. 419–425.

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

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

Freitas, J., Teixeira, A., Silva, S., Oliveira, C., Dias, M.S. (2014). Velum Movement Detection based on Surface Electromyography for Speech Interface", *Proceedings of Biosignals 2014*, Angers, France.

Hueber, T., Chollet, G., Denby, B., Stone, M. and Zouari, L. (2007). Ouisper: Corpus Based Synthesis Driven by Articulatory Data. *International Congress of Phonetic Sciences*, Saarbrücken, pp. 2193--2196.

Plux Wireless Biosignals, Portugal (2014). *Online: http://www.plux.info/*, accessed on 17 March 2014.

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

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

Srinivasan, S., Raj, B. and Ezzat, T. (2010). Ultrasonic sensing for robust speech recognition. *Internat. Conf. on Acoustics, Speech, and Signal Processing*, pp. 5102—5105.

Strevens, P. (1954). Some observations on the phonetics and pronunciation of modern Portuguese, *Rev. Laboratório Fonética Experimental*, Coimbra II, pp. 5--29.

Tran, V.A., Bailly, G., Loevenbruck, H. and Jutten, C. (2008). Improvement to a NAM captured whisper-to-speech system. *Proceedings of Interspeech 2008*, pp.1465-1468.

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

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