ISCA Archive
http://www.isca-speech.org/archive

Speech Production in Automatic Speech Recognition
Lyon, France
August 30, 2013

Multimodal Silent Speech Interface based on Video, Depth, Surface Electromyography and Ultrasonic Doppler: Data Collection and First Recognition Results

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, Universidade de Aveiro, Portugal
[3]ISCTE-Lisbon University Institute/ADETTI-IUL, Lisboa, Portugal

i-joaof@microsoft.com, ajst@ua.pt, miguel.dias@microsoft.com

Abstract

Silent Speech Interfaces use data from the speech production process, such as visual information of face movements. However, using a single modality limits the amount of available information. In this study we start to explore the use of multiple data input modalities in order to acquire a more complete representation of the speech production model. We have selected 4 non-invasive modalities – Visual data from Video and Depth, Surface Electromyography and Ultrasonic Doppler - and created a system that explores the synchronous combination of all 4, or of a subset of them, into a multimodal Silent Speech Interface (SSI). This paper describes the system design, data collection and first word recognition results. As the first acquired corpora are necessarily small for this SSI, we use for classification an example based recognition approach based on Dynamic Time Warping followed by a weighted k-Nearest Neighbor classifier. The first classification results using different vocabularies, with digits, a small set of commands related to Ambient Assisted Living and minimal nasal pairs, show that word recognition benefits can be obtained from a multimodal approach.

Index Terms: silent speech interfaces, multimodal, video and depth information, surface electromyography, ultrasonic Doppler sensing

1. Introduction

The use of Silent Speech Interfaces (SSI) is an alternative to conventional speech interfaces based on acoustic signals. Silent Speech designates the process of speech communication in the absence of an audible and intelligible acoustic signal. By extracting information of the human speech production process, an SSI is able to interpret, process or route the acquired data. The adoption of this type of interface brings advantage for people who have undergone a laryngectomy (i.e. total or partial removal of the larynx), or elder people for whom speaking requires a substantial effort. Since it is based on non-acoustically acquired speech cues, it also allows communication in noisy environments and is suitable for situations where privacy and confidentiality is required.

Several SSI based on different sensory types of data have been proposed in the literature: Implants in the speech-motor cortex [2], Electro-encephalographic (EEG) sensors [3]; Surface Electromyography (sEMG) sensors [4]; Electromagnetic Articulography (EMA) sensors [5]; Ultrasound (US) used concurrently with optical imaging of the tongue and lips [6]; Ultrasonic Doppler Sensing (UDS) [7]; Non-Audible Murmur (NAM) microphone [8]; glottal activity using electromagnetic [9], or vibration [10] sensors. Overviews can be found in [1] and [11]. However, the production of human speech is composed by several stages that go from intention to articulation effects [12], thus, acquiring data from a single stage limits the amount of useful information available for capture and further processing. Therefore, taking a higher level perspective, if multiple modalities could be used, a more complete representation of the speech production model could be obtained, benefiting speech recognition performance [13]. As eventually the weakest points of one modality can be minored by other(s), their combination should be investigated and compared with the performance observed by each one separately. In literature several multimodal SSIs can be found and almost all the approaches mentioned earlier also consider an audible acoustic signal stream in their experiments, with exception for some work based on brain computer interfaces [2][3]. Since this paper only addresses silent speech, we will not focus on the combination of audio with other input modalities.

The work presented here aims at creating the conditions to explore more complex combinations of Human-Computer Interaction (HCI) input modalities for SSI – exploring more non-invasive and recent modalities such as UDS - and to make an analysis of first evaluations. As such, we have selected multiple HCI technologies based on: the possibility of being used in a natural manner without complex medical procedures, low cost, tolerance to noisy environments, ability to work with speech-handicapped users and cost. Given these requirements, a novel type of SSI based on the following specifications was defined as our target: (1) Facial information acquired from Visual and Depth sensors; (2) sEMG of the articulator muscles; and (3) Capture of facial movements during speech using UDS. Since this is the first reported SSI that uses more than 2 data types and consequently, the first to combine the corresponding modalities, no corpora exists for the selected language – European Portuguese (EP) - or any other that the authors know of, causing the necessity of creating a database with this type of multimodal data. The amount of data collected, although being sufficient for a first proof-of-concept SSI, it is yet not sufficient to allow for a generalization of the observed data and apply classification methods such as Hidden Markov Models (HMMs). For that reason, we decided to use an example based classification method that uses Dynamic Time Warping (DTW) scores followed by a k-Nearest Neighbor (k-NN) classifier [14], which has achieved successful results in the literature for speech recognition tasks [15].

The remaining sections of this study are structured as follows: Section 2 presents a description of related work on previous multimodal approaches used in SSIs and on the most recent and relevant developments concerning the selected modalities. Section 3 describes the methodology and the system used to acquire multiples streams of data, giving particular attention to the synchronization solution used to register/align all the input signals. Section 4 presents the first word

recognition results, using example based techniques in a multimodal SSI. Finally, section 5 presents the conclusions of this research.

2. Related work

In 2004, Denby and Stone [16], 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. [6], 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. [17], 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 colored 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. in 2008 [21]. 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. [22], 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 [23]; the importance of acoustic feedback [24]; EMG-based phone classification [25]; and session-independent training methods [26]. For what UDS is concerned, it has been applied to several areas (e.g. voice activity detection [27], speaker identification [28], and synthesis [29]) including speech recognition with promising results [7][30].

3. Methodology / System design

Before conducting this research it was necessary to plan and gather all the necessary equipment, which in the case of UDS, led us to the development of custom built equipment [30]. The next step was to create the necessary conditions to register all signals via proper synchronization. For this purpose, we selected the sEMG recording device as the central unit that generates the alignment pulse for all the remaining modalities as described in section 3.3. After the system setup was ready, a proof-of-concept database was collected for further analysis.

3.1. The individual modalities

The work described here addresses 4 modalities: (1) Video which captures the image pixels of the speakers' mouth region and its surroundings, including chin and cheeks. (2) Depth which captures depth information of the same areas, providing useful information about the mouth opening and tongue position in some cases. (3) Surface EMG sensory data, which provides us information about the myoelectric signal of 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 proportional to the movements of the speaker's face. Based on the analysis of the Doppler signal, patterns of movements of the facial muscles, lips, tongue, jaw, etc., can be extracted [7].

The devices employed in this data collection were: (1) a Microsoft Kinect [19] that acquires visual and depth information; (2) an sEMG acquisition system from Plux [20] that captures the myoelectric signal from the facial muscles; (3) a custom built dedicated circuit board (hereon referred 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 [30]. All devices were connected to the same laptop.

The Kinect sensor was placed at approximately 0.7m from the speaker and it was configured to capture a color video stream with a resolution of 640x480 pixels, 24-bit RGB at 30 frames per second and a depth stream with a resolution of 640x480 pixels, 11-bit at 30 frames per second. Kinect was also configured to use the Near Depth range (i.e. range between 0.4m to 3m) 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. The sensors were attached to the skin using a single use 2.5cm diameter clear plastic self-adhesive surfaces and considering an approximate 2cm spacing between the electrodes center for bipolar configurations. The recordings took place in a single session, meaning that the sensors were never removed during the recordings of each speaker. 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 EMG channels 1 and 4 used a unipolar configuration (i.e. used one of the electrodes from the respective pair as a reference electrode located in a place 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. [4]) and EMG electrode 3 was placed according to recent findings about the detection of nasality in SSIs [31], a distinct characteristic of EP [32].

The UDS device was placed at approximately 40cm 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 of a frequency $f_a = 36kHz$ and

low passing the result, obtaining a similar frequency modulated signal centered at $f_1 = f_0 - f_a$, i.e., $f_1 = 4kHz$.

3.2. Multimodal Acquisition

The recordings took place in a quiet room with controlled illumination with 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. Our database contains the recordings of 8 native EP speakers - 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. No audible acoustic signal was produced by the speakers during the recordings and only one speaker had past experience with silent articulation. Before each session the participants received a 30m briefing that included instructions and speaker preparation. Each recording session took between 40 to 60 minutes generating an average 3.2GB of data per speaker that includes: session metadata; RBG and depth information of a 128x128 pixel square centered at the mouth center and the coordinates of 100 facial points for each Kinect frame; EMG data from the 5 available channels; two channel wave per prompt containing the Doppler and the synchronization signal; and a compressed video of the whole session. The prompts were presented to the speaker in a random order.

3.3. Registration of all input modalities

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 1. Synchronization occurs when the output of a synch signal, 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 UDS recording. Registration between the video and depth streams is ensured by Kinect SDK.

Using the information from the led and auxiliary channel with synch info, the signals were aligned offline. To align RGB video and the depth streams, we have used an image template matching technique that automatically detects the led position on each color frame. By applying a threshold on the average luminance of that image area, we can determine if the led is ON or OFF. 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.

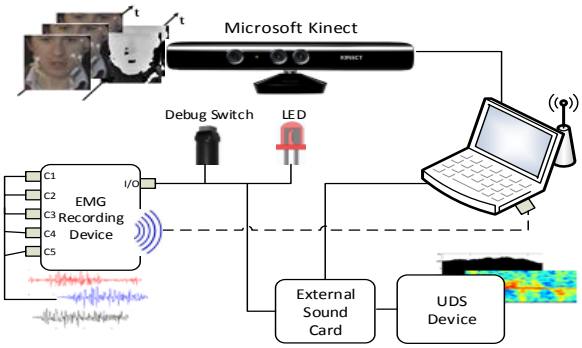

Figure 1: *Diagram of the alignment scheme showing the I/O channel connected to the three outputs – debug switch, external sound card and a directional led.*

3.4. Corpora

For this first experiment we have selected a vocabulary of 32 EP words, which can be divided into 3 distinct sets. The first set, used in previous work for other languages (e.g. [7]) and for EP in previous work [30], 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, e.g. Cato/Canto [katu]/[kɐ̃tu] or Peta/Penta [petɐ]/[pẽtɐ] – see [33] for more details), and is directly related with previous investigation on the detection of nasality with SSIs. Table 1 shows the last 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 in [34]. A total of 99 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, and occurring 3 times per recording session.

Table 1. *Set of words previously used in an AAL context.*

AAL Words				
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)*	

3.5. Feature Extraction

The first step for extracting visual and depth facial information is to determine a region-of-interest (ROI). For this process we use the real-time Active-Appearance Models (AAMs) [35], already provided by the Kinect SDK, to track 100 points in the speakers' face during the recordings. This allows us to have a 64x64 pixel ROI always centered at the speaker's mouth. After estimating the ROI, an appearance-based methodology based on previous work taken from the literature [22], is applied to both depth and color images. The challenge in extracting features based on appearance resides in collecting required information from the vast amounts of data present in image sequences. Each frame contains a large number of pixels and is obviously too large to model as a feature vector. In order to reduce dimensionality and to allow better feature classification, we start by applying a Discrete Cosine Transform (DCT) to extract the 45 coefficients with higher energy values. Then, to match with frame rates from other input modalities, we

interpolate the obtained frames from 30Hz to 100Hz. After this process, a two stage Linear Discriminant Analysis, LDA is applied in order to reduce dimensionality of the resulting feature vector. In the first stage, we apply LDA to each frame, reducing it to 15 features per frame. In the second stage, LDA is applied to the stacking of 12 adjacent frames, selecting the 10 mapped features with the highest eigenvalues. Finally, the first and second derivative are appended to the feature vector.

For UDS feature extraction we followed a similar approach to [7] and started by pre-processing the UDS signal. The acquired signal was first zero-averaged, then a 3 sample moving average filter is applied to suppress the 4 kHz carrier and later a difference operator is applied. After the pre-processing stage, we split the signal into 50ms frames with a 10ms frame shift, and apply a Discrete Fourier transform (DFT) with a second-order Goertzel algorithm over the preprocessed signal for the interval around the carrier - 3500 Hz to 4750 Hz. Finally, a DCT is applied to the DFT results to de-correlate and compress the signal, extracting the first 30 coefficients, which contain most of the signal energy.

For sEMG feature extraction we have used a similar approach to the one described by [13] and [31], based on temporal features, since it has been shown in previous studies that time-domain features presents better accuracy results. The extracted features are frame-based and for any given sEMG signal frames of 30ms and a frame shift of 10ms is considered. After feature extraction, we applied LDA to reduce the feature vector to 30 coefficients per frame. The vocabulary of each word set was used as the LDA categorical variable.

4. Isolated Word Recognition Results

As a first approach, we decided to analyze if better word recognition performance could be obtained by fusing multiple classification approaches. Another question addressed in this paper is if fusion of different modalities should occur at an early stage after feature extraction (i.e. feature fusion), or should be applied upon classification results of each individual modality (i.e. decision fusion). Hence, we started by obtaining baseline results for each modality for each word set, as depicted in Table 2 and Table 3. Considering the reduced number of samples per speaker in the corpus we decided to use an example based technique that uses DTW alone or DTW followed by a weighted k-NN for a first classification approach. To split the data into train and test, we adopted a 10-fold cross-validation strategy [30]. The LDA described in section 3.5 is performed for each of the 10-fold partition.

In the DTW technique, we calculate the distance between the test word samples and each sample of the training set choosing the class with the minimum distance. When multiple distances are used for each training sample (as when we have multiple modalities and consequently multiple feature vectors), the class with the minimum sum of all distances for each sample is selected. In case of using the DTW followed by a weighted k-NN classifier, the DTW distances are used as an input to build the feature space. The *k* number of neighbors are dynamically determined based on the size *n* of the training set, having $k = \sqrt{n}$ [15]. For prediction, we used an Euclidean distance metric and an inverse distance weighting function, i.e. each point of the model has a weight equal to the inverse of its distance.

The results presented in Table 2 show that the best result is found for the combination of Video, Depth and UDS input modalities, with an average error rate of 69.6%, using DTW followed by k-NN and a Decision Fusion technique. However, the combination of Video and Depth presented very similar results with an average error rate of 69.7%, particularly if we consider the statistical dispersion of the results shown by the confidence interval. Regarding individual modalities, Video presented the best results with an average error rate of 70.1%, followed by Depth with 71.1%, both using the DTW+k-NN classifier. When comparing classification techniques, DTW followed by k-NN seems to outperform classification based on DTW only by a small margin of 0.6% in average.

Regarding which fusion approach is better, results vary according to classification technique and modalities. We find that, using feature fusion with DTW only, has better results in average, but when applying k-NN, decision fusion has better performance, except for the case where we combine all modalities.

When analyzing the results of individual word sets, as depicted in Table 3 the combination of multiple modalities seems to present a better performance, except for the case of the AAL word set, which has a larger vocabulary, and where Video presents the best performance using both classification pipelines with a 71.4% average error rate using DTW and 70.8% using DTW+k-NN. However, combinations between Video and Depth using a decision fusion technique present a very close performance result with 71.4% error rate but a larger confidence interval. In the remaining cases, the combination of Video and Depth present the best results with error rates of 72.1% for Digits, 66.2% for Nasal Pairs and 65.7% for the vocabulary mix, all achieved using the DTW+k-NN pipeline.

Discarding the combination of all modalities (with clear problems), a one-way ANOVA was performed, separately, to the average values of error and to the error rates for each set, having as factor the existence or not of several modalities. Significantly different and lower results were found for multimodal combinations error rates. As an example, for the average of the 3 word sets, was obtained $F(1,14) = 6.249$, $p=0.0255$.

Table 2. *Average isolated word recognition error rate for individual and different combination of multiple modalities and for two classification techniques. It includes the results of all word sets.*

	DTW	*DTW+KNN*
Video	71.3%	70.1%
Depth	71.8%	71.1%
EMG	89.2%	89.4%
UDS	81.4%	81.7%
Video+Depth (FF)	71.4%	69.7%
Video+Depth (DF)	72.3%	69.7%
Video+Depth +UDS (FF)	**71.2%**	70.4%
Video+Depth +UDS (DF)	73.0%	**69.6%**
Video+Depth+UDS+EMG (FF)	86.7%	88.3%
Video+Depth+UDS+EMG (DF)	88.4%	90.2%

Table 3. *Average isolated word recognition error rate with 95%confidence interval of the 10-fold, for individual and different combination of multiple modalities, using 4 different vocabularies (including a random mixed selection of 8 words based on the other word set) and 2 different fusion techniques – feature fusion (FF) and decision fusion (DF).*

	Digits		**AAL words**		**Nasal pairs**		**Vocabulary Mix**	
	DTW	*DTW+k-NN*	*DTW*	*DTW+k-NN*	*DTW*	*DTW+k-NN*	*DTW*	*DTW+k-NN*
Video	74.2% ± 5.4%	72.5% ± 3.7%	**71.4% ± 4.2%**	**70.8% ± 4.7%**	72.5% ± 7.6%	67.9% ± 7.3%	67.2% ± 5.9%	69.3% ± 4.5%
Depth	74.2% ± 4.3%	74.1% ± 4.6%	72.6% ± 5.0%	71.7% ± 6.2%	70.9% ± 7.6%	67.8% ± 6.8%	69.3% ± 4.0%	70.9% ± 4.6%
EMG	87.9% ± 4.7%	85.8% ± 3.9%	92.3% ± 3.2%	91.1% ± 2.7%	89.6% ± 4.5%	92.2% ± 3.9%	87.0% ± 4.1%	88.6% ± 4.6%
UDS	81.7% ± 4.6%	83.4% ± 3.5%	83.0% ± 3.9%	82.5% ± 2.1%	82.4% ± 5.4%	78.1% ± 6.0%	78.6% ± 8.8%	82.8% ± 6.0%
Video+Depth (FF)	75.0% ± 5.3%	**72.1% ± 4.0%**	73.1% ± 6.4%	73.4% ± 5.6%	70.4% ± 7.0%	67.3% ± 9.3%	67.2% ± 6.4%	**65.7% ± 5.8%**
Video+Depth (DF)	**73.3% ± 4.7%**	72.9% ± 4.2%	71.4% ± 5.2%	73.7% ± 5.2%	75.1% ± 7.4%	**66.2% ± 9.0%**	69.3% ± 6.1%	65.8% ± 7.1%
Video+Depth +UDS (FF)	74.6% ± 5.4%	72.9% ± 3.3%	73.2% ± 5.6%	72.6% ± 5.2%	**69.9% ± 7.2%**	69.4% ± 7.8%	**67.2% ± 5.8%**	66.7% ± 5.5%
Video+Depth +UDS (DF)	73.3% ± 5.2%	72.5% ± 4.2%	72.3% ± 4.8%	72.6% ± 4.7%	74.6% ± 7.9%	66.8% ± 9.5%	71.9% ± 5.6%	66.3% ± 6.6%
Video+Depth+ UDS+EMG (FF)	85.4% ± 4.2%	91.7% ± 1.6%	92.0% ± 2.6%	89.6% ± 2.3%	86.1% ± 6.9%	86.4% ± 3.3%	83.3% ± 5.4%	85.4% ± 4.6%
Video+Depth+ UDS+EMG (DF)	86.2% ± 4.9%	89.2% ± 3.7%	93.8% ± 2.8%	93.2% ± 2.5%	89.5% ± 3.3%	90.1% ± 3.1%	83.9% ± 6.7%	88.1% ± 4.2%

5. Discussion

This study explores the use of a novel multimodal approach in order to capture a more complete representation of the speech production model, also addressing the problem of how to combine multiple modalities and how to model information with limited amounts of training data.

Video, Depth, EMG and UDS capture different types of data at different stages of the speech production model [11]. Our aim with this initial study was to analyze if a multimodal approach could help minimizing the weakest points of one or several modalities, towards a solution where a more complete representation of the speech production model is achieved. The chosen modalities capture information about several articulators, e.g. lips, chin, tongue, etc. Nonetheless, some modalities may obtain a more accurate representation of a particular articulator when compared to others. For example, Video is not able to get an accurate representation of the tongue. However, EMG channels 1 and 5 can, in theory, obtain information about tongue movements during speech.

Overall results point towards performance advantages in using a multimodal solution to implement an SSI, particularly for the UDS and EMG cases. However, it is not possible to drive a final conclusion on which approach presents the higher gain. In this study two multimodal combinations stand out: Video and Depth; and Video, Depth and UDS. The ANOVA analysis has shown that the combination of these approaches introduces a significant improvement in most cases when compared with the individual modalities or other modalities combination. However, if we take into account the confidence intervals depicted in Table 3, none of the multimodal combinations seem to clearly outperform Video or Depth modalities alone.

Recognition problems were also detected when using EMG, which based on literature seem to be caused by the small data set [23]. As such, the combination of other modalities with EMG did not improve the results. Even though it is acceptable for this experience, other literature studies have shown that error rates tend to improve substantially when considering audible speech articulation, simultaneously with individual modalities like EMG [23] [33] or UDS [30], as opposed to silent speech articulation, as in the research described in this paper.

We have also compared two example based classification techniques: DTW and DTW+k-NN. Results indicate an increase of performance when adding the k-NN classifier to the recognition pipeline, as this extra step based on clustering allows to discard, in some cases, DTW distance outliers.

Regarding the fusion of techniques results are not conclusive on which technique is better, since it varies according to classification technique and modalities. Nonetheless, results show that decision fusion works better when used with the DTW followed by a k-NN classifier. This can be explained by the richer feature space used as input for the k-NN classifier, consequently generating better results.

6. Conclusions

The work presented in this research, presents the first steps of an SSI for isolated word recognition, based on the registration of four non-invasive input modalities, all of relatively low cost.

This paper describes the design and registration solution built for such system, the data collection methodology and also an analysis of the isolated word recognition performance of this SSI. Our results show that a significant difference in recognition rates can be found between unimodal and multimodal approaches in favor or the latter, and that benefits can be obtained by aligning several modalities, especially when registering Video, Depth and UDS, or Video and Depth. Results also indicate a slight better performance when using a decision fusion approach with DTW followed by a k-NN classifier.

As future work we intend to develop a classification model based on the human speech production model, where the weakest points of each modality can be overcome, e.g. combine visual lip information with the tongue myoelectric signal obtained with the sEMG. We also intend to expand the collected database and explore more appropriate features for the multimodal scenario.

7. Acknowledgements

This work was partially funded by Marie Curie Golem (ref.251415, FP7-PEOPLE-2009-IAPP), by FEDER through the Program COMPETE under the scope of QREN 5329 FalaGlobal (PTDC/EEA-PLP/098298/2008) 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.

8. References

[1] Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J.M., Brumberg, J.S, "Silent speech interfaces". *Speech Communication*, Vol. 52, Issue 4, pp. 270-287, April 2009.

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

[3] Porbadnigk, A., Wester, M., Calliess, J. and Schultz, T., "EEG-based speech recognition impact of temporal effects", *Biosignals 2009*, Porto, Portugal, pp.376–381, January 2009.

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

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

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

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

[8] 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, September 2009.

[9] 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. "Exploiting non-acoustic sensors for speech enhancement", *IEEE Trans. Audio Speech Lang. Process*, Vol. 14, Issue 2, pp. 533–544, 2006.

[10] Patil, S. A. and Hansen, J. H. L., "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, April 2010.

[11] Freitas, J. Teixeira, A. Dias M. S. and Bastos, C., "Towards a Multimodal Silent Speech Interface for European Portuguese", *Speech Technologies*, Ivo Ipsic (Ed.), ISBN: 978-953-307-996-7, InTech, 2011.

[12] Levelt, W., "Speaking: from Intention to Articulation", Cambridge, Mass.: MIT Press, 1989.

[13] Jou, S.C.S., "Automatic Speech Recognition on Vibrocervigraphic and Electromyographic Signals". *Ph.D. dissertation*. NASA, 2008.

[14] Sainath, T.N., Ramabhadran, B., Nahamoo, D., Kanevsky, D., Van Compernolle, D., Demuynck, K., Gemmeke, J.F., Bellegarda, J.R. and Sundaram, S., "Exemplar-Based Processing for Speech Recognition: An Overview", *Signal Processing Magazine, IEEE* , vol.29, Issue 6, pp.98,113, Nov. 2012.

[15] De Wachter, M., "Example based continuous speech recognition," *Ph.D. dissertation*, K. U. Leuven, ESAT, Belgium, 2007.

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

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

[18] Tran, V.A., "Silent speech communication: whisper-to-speech conversion", *Ph.D. dissertation*, Institut National Polytechnique de Grenoble, 2010.

[19] Microsoft Kinect, *Online: http://www.xbox.com/en-US/kinect*, accessed on 11 Mar. 2013.

[20] Plux Wireless Biosignals, Portugal, *Online: http://www.plux.info/*, accessed on 24 Feb. 2013.

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

[22] Galatas, G., Potamianos, G., Makedon, F., "Audio-visual speech recognition incorporating facial depth information captured by the Kinect," *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pp.2714,2717, 27-31 Aug. 2012.

[23] Wand, M. Schultz, T., "Investigations on Speaking Mode Discrepancies in EMG-based Speech Recognition", *Interspeech 2011*, Florence, Italy, 2011.

[24] Herff, C. Janke, M. Wand, M. Schultz, T., "Impact of Different Feedback Mechanisms in EMG-based Speech Recognition", *Interspeech 2011*. Florence, Italy, 2011.

[25] Wand, M. Schultz, T., "Analysis of Phone Confusion in EMG-based Speech Recognition", *IEEE International Conference on Acoustics, Speech and Signal Processing 2011*, Prague, Czech Republic, 2011.

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

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

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

[29] Toth, A.R., Kalgaonkar, K., Raj, B., Ezzat, T., "Synthesizing speech from Doppler signals", I*EEE International Conference on Acoustics Speech and Signal Processing*, pp.4638-4641, 2010.

[30] 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.

[31] Freitas, J., Teixeira, A., Silva, S., Oliveira, C., Dias, M.S., "Nasality Detection for Speech Interface based on Surface Electromyography", *submitted for Computer Speech and Language*.

[32] Strevens, P., "Some observations on the phonetics and pronunciation of modern Portuguese", *Rev. Laboratório Fonética Experimental*, Coimbra II, pp. 5–29, 1954.

[33] Freitas, J., Teixeira, A. and Dias, M. S., "Towards a Silent Speech Interface for Portuguese: Surface Electromyography and the nasality challenge", *Int. Conf. on Bio-inspired Systems and Signal Processing*, Vilamoura, Algarve, Portugal, 2012.

[34] Teixeira, V., Pires, C., Pinto, F., Freitas, J., Dias, M.S., Rodrigues, E.M., "Towards elderly social integration using a multimodal human-computer interface", Proceedings of the 2nd International Living Usability Lab Workshop on AAL Latest Solutions, Trends and Applications, AAL 2012, Feb. 2012.

[35] Cootes, T.F., Gareth J.E., and Christopher J.T., "Active appearance models." Pattern Analysis and Machine Intelligence, IEEE Transactions on Vol. 23, Issue 6, pp. 681-685, 2001.