Silent Speech Interaction for Ambient Assisted Living Scenarios

António Teixeira[1,2(✉)], Nuno Vitor[1], João Freitas[3,4], and Samuel Silva[2]

[1] Department of Electronics Telecommunication and Informatics, University of Aveiro, Aveiro, Portugal
ajst@ua.pt
[2] Institute of Electronics and Informatics Engineering of Aveiro (IEETA), Aveiro, Portugal
[3] Microsoft Language Development Center (MLDC), Lisbon, Portugal
[4] DefinedCrowd, R. da Prata 80, Lisbon, Portugal
http://wiki.ieeta.pt/wiki/index.php/User:Teixeira

Abstract. In many Ambient Assisted Living (AAL) contexts, the speech signal cannot be used or speech recognition performance is highly affected due to ambient noise from televisions or music players. Trying to address these difficulties resulted in the exploration of Silent Speech interfaces (SSI), making use of other means to obtain information regarding what the user is uttering, even when no acoustic speech signal is produced.

The automatic recognition of what has been said, based only on images of the face, is the purpose of Visual Speech Recognition (VSR) systems, a type of SSI. However, despite the potential of VSR for enabling the interaction of older adults with new AAL applications, and current advances in SSI technologies, no real VSR application can be found in the literature.

Based on recent work in SSI, for European Portuguese, a first working application of VSR targeting older adults is presented along with and results from an initial evaluation. The system performed well, enabling real-time control of a media player with an accuracy of 81.3% and performing classification in around 1.3 s. At this stage, the results vary from speaker to speaker and the system performs better if the words are correctly articulated. The effect of distance of the speaker to the video apparatus (a Kinect One) proved not to be an issue in terms of the system accuracy.

Keywords: Silent Speech Interfaces (SSI) · Visual Speech Recognition (VSR) · Ambient Assisted Living (AAL) · Elderly

1 Introduction

With the increasing percentage of elderly in the world population, Information and communications technology (ICT) providers are more and more challenged

J. Zhou and G. Salvendy (Eds.): ITAP 2017, Part I, LNCS 10297, pp. 369–387, 2017.
DOI: 10.1007/978-3-319-58530-7_29

to help them stay active at their homes. However, the adoption of Ambient Assisted Living (AAL) applications by elderly users strongly depends on its usability and quality of the interaction.

Speech, as we are all aware, is the easiest way for humans to communicate, can be used at a distance while keeping hands free. But in many AAL situations (e.g. noisy environments due to sound of televisions or music), the speech signal cannot be used or speech recognition performance is highly affected. Elderly speech also affects the performance of speech recognizers.

Silent Speech interfaces (SSI) [8], can be used to address these challenges, since it looks beyond the acoustic signal during spoken communication.

Informally, one can say that an SSI system extends the human speech production process by exploring biometric signals other than the voice. In fact, audible speech is just the end result of the complex process of speech production involving, for example, cerebral and motor activities, and a wide range of technologies support the acquisition of data pertaining these different parts of the process. For instance, surface electromyography can capture muscle activity and video can provide data regarding lip movement.

Nowadays, there are several studies in SSI considering every stage of the speech production process (see Sect. 3). Depending on the speech production signal or signals that the SSI system targets, SSI approaches can be either invasive or not. An advantage of vision-based approaches, which target the visible effects of speech (mainly: lips position, jaw position, tongue tip when lips are open), is that they typically require no attachment or insertion of devices. These systems, referred as Visual Speech Recognition (VSR) systems can use different types of cameras (e.g. RGB, depth cameras, etc.).

Despite the potential of VSR for older adults' interaction with new AAL applications and the advances in SSI technologies, to best of our knowledge, no real VSR application can be found in the literature.

Main objective is to develop and evaluate a VSR application, directed to older adults in a real AAL scenario (described in Sect. 2) by leveraging previous work of the authors in SSI [8], multimodal interaction [30] and AAL [7].

Paper is structured as follows: next section presents, briefly, the scenario chosen for the proof-of-concept; Sects. 3 and 4 present some background regarding the process of speech production and a survey of related work in SSI, focusing in VSR; Sect. 5 presents the developed prototype; evaluation results are the subject of Sect. 6; paper ends with conclusions and future work, in Sect. 7.

2 Scenario

For a first proof-of-concept of the potential of SSI for A we chose to develop an application to control a multimedia player in European Portuguese, a relevant scenario for AAL.

The most important areas in AAL are related to allowing the user with some kind of limitations to control entertainment systems, access to social networks,

or similar. Controlling such systems allows the user to access to access memories and information from friends and family.

Considering our goal, we chose to take advantage of one of the world's most used open multimedia player, VLC Media Player (VLC) [17]. VLC is seen by users as simple to use and supports a wide range of multimedia formats. Thus, given that it is open source we can adapt it to the targeted AAL scenario.

In VLC it is possible to load a set of videos and then select the video that the user wants to watch. Also, it also supports common media controls such the sound volume, the speed of the video and the stop and play functions. In our scenario these controls are used by detecting the silent speech command, i.e. the movements of the lips and the chin of the user. This allows the user to control the system in a noisy ambient, in case the user wants some privacy or in situations where the user has speech production limitations.

The considered vocabulary is in European Portuguese and the selected words were considered to be the most natural for the targeted scenario. Several iterations were done to reach this set of words. The Table 1 has, in the first column, the set of words chosen in Portuguese and in the second column their translation in English. To help with the recognition accuracy, we avoided phonetically similar words, particularly in commands with two words.

Table 1. Set of words chosen regarding the AAL context of using the VLC

Portuguese	English
Ver filme	Watch a movie
Parar	Stop
Continuar	Continue
Aumentar volume	Increase volume
Baixar volume	Decrease volume
Mais rápido	Faster
Mais lento	Slower
Próximo filme	Next movie

3 Background

Due to the complexity of the speech production and perception process, in this section are presented some topics needed to understand the SSI basics.

3.1 Speech Production

Speech production requires a complex series of events and is considered the most complex motor task performed by humans [29]. In a fluent conversation, we are able to produce two or three words per second.

The speech production process can be divided into several stages [8,9,18]. In order, these stages are, according to [8, p. 4]: (1) Conceptualization and Formulation, (2) Articulatory Control and (3) Articulation.

In the first stage, the brain converts communication intentions into messages, creates linguistic representation required for the expansion of this preverbal messages and produces a phonetic plan [5]. Articulatory control, the second stage, using information from first stage, generates the electrical impulses need to control the articulators. These commands must simultaneously control all the aspects of the articulation, including the lips, jaw, tongue and velum [24]. The changes in articulators, in the last stage, continuously change vocal tract characteristics (mainly shape and stiffness), producing the acoustic speech signal and other effects (e.g., alterations in the face).

In speech production, the articulatory muscles, like the tongue, have a vital role because they can shape the air stream to produce a recognizable speech. Mandibular movement also has an important role in this process. Despite the relevance of cavities, surfaces and organs such as the lungs in speech production, the articulators have a key role in the pronunciation of the different sounds of a language. Their position defines the articulatory and resonant characteristics of the vocal tract.

Articulators can be active or passive. The active articulators include the lips, tongue, lower jaw and velum, being the tongue the most important, as it participates in the production of (almost) all sounds. The passive include the teeth, alveolar ridge and hard palate. Figure 1a sagittal view of several articulators. The most visible effects of the speech production chain are the movement of the lips, tongue, lower jaw and the chin.

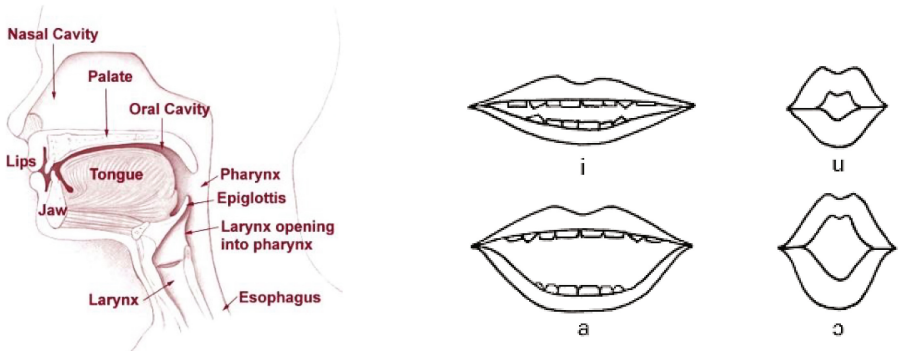

Fig. 1. Sagittal view of the vocal tract depicting its main regions and several articulators [8], at left, and example of visible effects of speech production, at right.

3.2 SSI Basics

A Silent Speech interface (SSI) is a system that interprets human signals other than the audible acoustic signal enabling speech communication [6]. A SSI system is commonly characterized by the acquisition of information from the human speech production process such as articulations, facial muscle movement or brain activity. It is possible to say that the SSI systems extends the human speech production process by exploring biometric signals other that voice, using sensors, cameras, etc. [8].

There are multiple works in SSI done in every stage of the speech production stage. For example, in the first stage (Conceptualization) works were done on the interpolation of the signals from implants in the speech-motor cortex [3] and from Electroencephalography (EEG) sensors [23]. Using information from Articulation stage, works were made regarding the movement of the lips [1,32] or the movements of speaker's face estimated through Ultrasonic Doppler sensing [10], for example.

The usual architecture of a SSI system comprises modules for: signals acquisition and processing, extraction of Features, and classification.

The acquisition of signals from any stage of the human speech production process can invasive or non-invasive and obtrusive or not. An invasive modality needs a medical attention to be used or requires the use of sensors. An obtrusive modality requires wearing some type of equipment, such as, for example sensors.

Choosing the best SSI is not an easy task because they have different advantages and disadvantages regarding price, usability, accuracy, and speaker's dependence.

3.3 Methods to Collect Visual Information for SSI Systems

The most used way to collect visual information from the speech production process is through cameras, non-invasive and non-obtrusive method. There are some different cameras on the market like: RGB cameras that collect information on the color space; and depth cameras, that collect depth information through the stereo vision approach, infrared or time of flight technology.

The introduction Microsoft Kinect for Windows made simple to have both simultaneously, as it provides both types of technologies at an affordable price.

RGB cameras are used to collect information pixel per pixel in a RGB color space (Red, Green and Blue). Today these type of cameras use CMOS or charge-coupled device (CCD) image sensor and operate in general in a Bayer filter arrangement, where green gives twice as many detectors as red and blue (red-green-blue-green (RGBG) color filter array (CFA)) in order to give better luminance resolution than chrominance resolution.

To get the information of depth of the various pixels in a image, depth cameras use one of two different methods [16]: stereo vision or Time of Flight (TOF). Stereo vision uses two (or more) images taken at the same time from separate cameras and the differences are analyzed to yield depth information [2].

Time of Flight cameras (Fig. 2) use modulated infrared light, not visible to humans. Then, a sensor captures the reflected light to extract distance information [15].

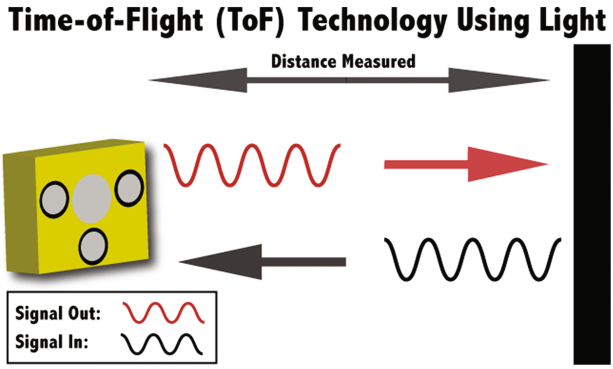

Fig. 2. Simplified illustration of the principle used in Time of Flight depth cameras. [31].

In Visual Recognition Systems, the camera is one of the key issues. The resolution of the camera is extremely important since it will define the detail of each image representing the data collected. Frame rate (fps) is another important specification regarding the amount of information that the camera could record in a second. This becomes a key factor in terms of speech recognition systems considering the movement of the lips.

Microsoft and Prime Sense released the Kinect (for Xbox 360) in 2010. With its 2 cameras and the capability to track 48 points from the human skeleton, this device brought a complete new approach in fields Human Computer Interaction, face tracking, and Audio-Visual Speech Recognition. Despite the many systems created using this versions of Kinect [14,22,35], this camera was far from being perfect due to its low resolution (640×480) and limitations of depth information extraction technique (structured light).

Kinect One (see Fig. 3) was release by Microsoft in 2013. This new version brought several improvements such as a better resolution (Full HD, 1920×1080 in RGB images); better depth images, thanks to the Time of Flight (TOF) technology; greater accuracy over its predecessor; capability to process 2 gigabits of data per second; capability to track up to 6 skeletons at once; and a wider

Fig. 3. Kinect one for windows

field of view. This new version soon became an important piece in visual speech recognition systems because of its relation in performance over price.

4 Related Work and State-of-the-Art

In this section is presented some relevant related work in SSI: starting from recent work in SSI in general; continuing with recent work in VSR, the SSI method adopted for the work described in this paper; and ending with information regarding SSI for Portuguese, the language adopted for the worked reported.

4.1 Representative Recent Developments in SSI

EEG is commonly used in SSI, being a representative example the work for Japanese, with EEG signals from 63 channels, by Matsumoto [19], showing that classification accuracies can be improved if an adaptive collection is made. An increase from 56–72% to 73–92% was reported, using SVM with Gaussian Kernel as classifiers.

In 2014 Freitas and coworkers created a multimodal SSI system, for European Portuguese language, combining sensing technologies such as Video and Depth input, Ultrasonic Doppler sensing and Surface Electromyography. These streams of information are synchronously acquired with the aim of supporting research and development of a multimodal SSI [12]. Due to the number and variety of streams, this system continues to be a good example of the state-of-the-art in multimodal SSI. The approach is non-invasive, however it is obtrusive, as EMG sensors were needed for the Surface Electromyography signals (see Fig. 4).

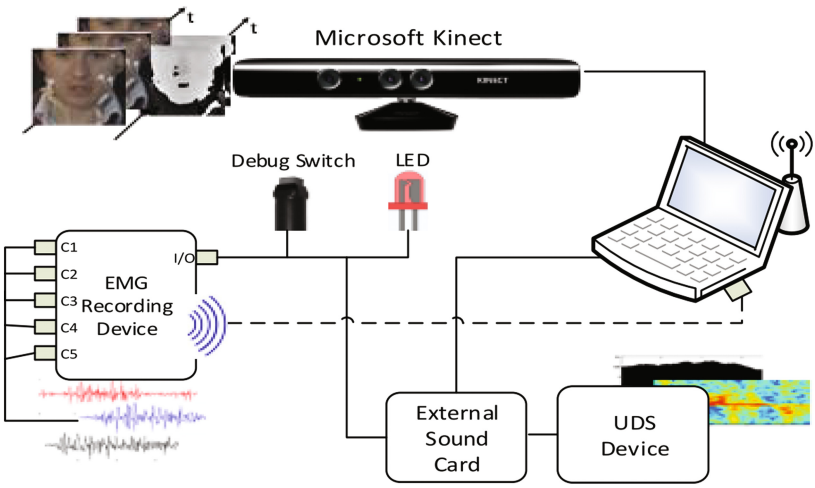

Fig. 4. Diagram of the alignment scheme of João Freitas and co-workers [12].

A vocabulary of 32 words in European Portuguese was used regarding an AAL context, divided in sets of digits, pairs of common words and AAL words.

For classification, Dynamic Time Warping and k-Nearest Neighbor classifiers were used. His results points towards performance advantages using a multimodal solution to implement an SSI, especially for Ultrasonic Doppler sensing and Surface Electromyography. However, a final conclusion can not be taken regarding which approach represents a higher gain.

The best results had nearly 94% accuracy (for AAL words with features from Video+Depth+UDS+EMG with DTW classification) and the worst were nearly 65% (for a Vocabulary Mix using features from Video+Depth with DTW+kNN classification).

Fig. 5. Tongue magnetometer and Outer Ear Interface [28].

Also in 2014, from Georgia Institute of Technology, USA, a wearable system (obtrusive and intrusive) was created [28] to capture tongue and jaw movements during silent speech in English (Fig. 5).

To achieve that, a two system part was created: one part with a Tongue Magnet Interface, which utilizes the 3-axis magnetometer aboard Google Glass to measure the movement of a small magnet glued to the user's tongue, and the second part a Outer Ear Interface which measures the deformation in the ear canal caused by jaw movements using proximity sensors embedded in a set of earmolds. The classification was done using hidden Markov model-based techniques to select one of the 11 phrases.

During pronunciation of 11 distinct phrases, the average user dependent recognition accuracy was 90.5% using both parts of the system. Using just the part of the Outer Ear Interface (non-intrusive but still obtrusive) the system performs with an accuracy of 85.5%.

4.2 Silent Speech Based on Visual Information

One of the first studies in Visual Speech Recognition (VSR) was in 1994. This study was based on a word recognition system with a lip modeling approach for the recognition task [25]. This system had a 85% accuracy using the height and width of the lips, but only 2 words were tested.

In 2007, Werda created an Automatic Lip Feature Extraction prototype, named as ALiFE, that could automatically localize lip feature points in a speaker's face and carry out a spatial-temporal tracking of these points [33]. The points of interest in Werda work were the top center of the upper lip, bottom center of the lower lip and corners (Fig. 6). By using these points it was possible to extract features like the width (distance between the corners points), the height (distance between the top and bottom points) and also the area consisting of the inside of the mouth. They used multiple speakers in their tests (females and males). French was the used language in the tests and the accuracy obtained was 72.7%.

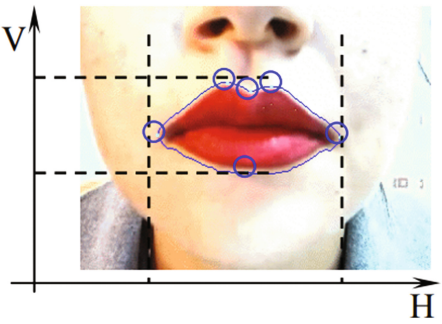

Fig. 6. Points of interest detection by the projection of final contour on horizontal and vertical axis (H and V) [33]

More recently, using the Kinect RGB camera and depth information, without the information of sound (VSR), to obtain 18 points of the lips (Fig. 7) and extracting the angles between all these points, Yargic and Dogan [35] created a system for a Turkish vocabulary of 15 words (color names), obtaining an accuracy rate of 78.2%. They used KNN classifiers.

Another recent VSR system was proposed by Frisky and colleagues [13] applying a video content analysis technique. Using spatiotemporal features descriptors, features were extracted from video containing visual lip information. A preprocessing step is employed to remove noise and enhance the contrast of images of every frame. This system achieved an accuracy between 25.9% and 89.02%.

One of the main features that is extracted in SSI based in visual information are the lips and their position/movement over time. Studies are being developed to find them as accurate and quick as possible [4].

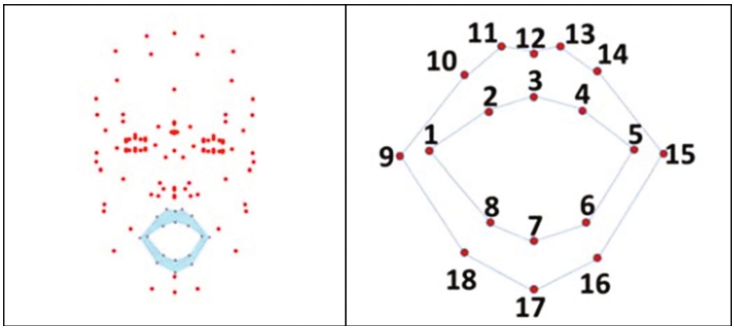

Fig. 7. Features used by Yargi and Dogan: 18 lip feature points and their assigned ID values [35].

4.3 Silent Speech for Portuguese

Regarding the Silent Speech Interfaces (SSI) for European Portuguese (EP), in 2010, Freitas started working, during his PhD, on a solution that addressed the issues raised in adapting existing work on SSI to a new language. Initial work focused on Visual Speech Recognition (VSR) and Acoustic Doppler Sensors (ADS) for speech recognition, evaluating this methodologies in order to cope with EP language characteristics. Dynamic Time Warping (DTW) was used, achieving an Word Error Rate (WER) of 8.6% [9].

In 2013, in a new work on SSI for EP, Freitas et al. [11] 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 SSI. For classification, Dynamic Time Warping (DTW), followed by a weighted k-Nearest Neighbor (kNN) classifier, was used. Results showed 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 Ultrasonic Doppler, or Video and Depth. Results also indicate a slight better performance when using a decision fusion approach with DTW followed by a kNN classifier [11].

One of the most recent works in a Silent Speech for Portuguese (and also Visual Silent Recognition) is [1]. In his dissertation Abreu used Kinect One to extract geometric and articulatory features from the lips. For the lips' segmentation, Abreu considered two color spaces: RGB and YCbCr. From the RGB frames he used the green channel in order to extract the external points of the lips and from the YCbCr color space the Cr channel was used to obtain the internal points of the lips. After the features extracted, Abreu made some normalizations such as length normalizations to the feature vectors to be sent to the classifiers and some distance normalization.

The selected vocabulary consisted of 25 European Portuguese words, which were divided into 2 sets: one with a widely used set of words used in speech recognition literature, digits from zero to nine and the other taken from a Ambient Assisted Living context.

The classification was done using SVM classifiers and the best accuracy of his system (ViKi - Visual Speech Recognition for Kinect) was 68% based on geometric features and 34% of recognition accuracy based on articulatory features. An hybrid solution using both geometric and articulatory was also tested achieving an accuracy of 49%.

5 Proof-of-Concept Prototype

Our proof-of-concept was developed with the Kinect One camera from Microsoft using VSR of a small set of commands (e.g. "See Movie"), uttered by the user, positioned at some distance of the front of the camera. The recognized commands are passed to VLC player in order to control it.

5.1 Requirements

One of the most important requirement is that the system has to permit some real daily life experiences, for example controlling a television at a certain distance (e.g. from the couch). However, given a typical living room scenario with a television turned on, it is probable to exist some audio noise. In this case, a SSI based on visual speech recognition allows to recognize speech without using acoustic information.

Another requirement is that the proof-of-concept prototype must detect the user's face and start and stop recording data automatically (we excluded push-to-talk solutions). This way a more natural solution is achieved with clear advantages for people with motor limitations.

5.2 Architecture

The system follows the architecture of traditional VSR systems [1,14,27] and takes the advantages of the Kinect One Camera to extract the features from the lips and chin of the user. A diagram with the main actions and modules is presented in Fig. 8.

System architecture follows the classic approach in pattern recognition, integrating feature extraction and classifiers, and is divided into 4 main blocks (Activity Detection, Feature Extraction, Classification and Data Base Creator). There are 2 modes: training and testing/real use. In training mode features extracted are stored in a database, and used to train the classifiers. The test path is the one where the system is used to control the VLC. It cannot be used without a previously creating the database and training the classifiers.

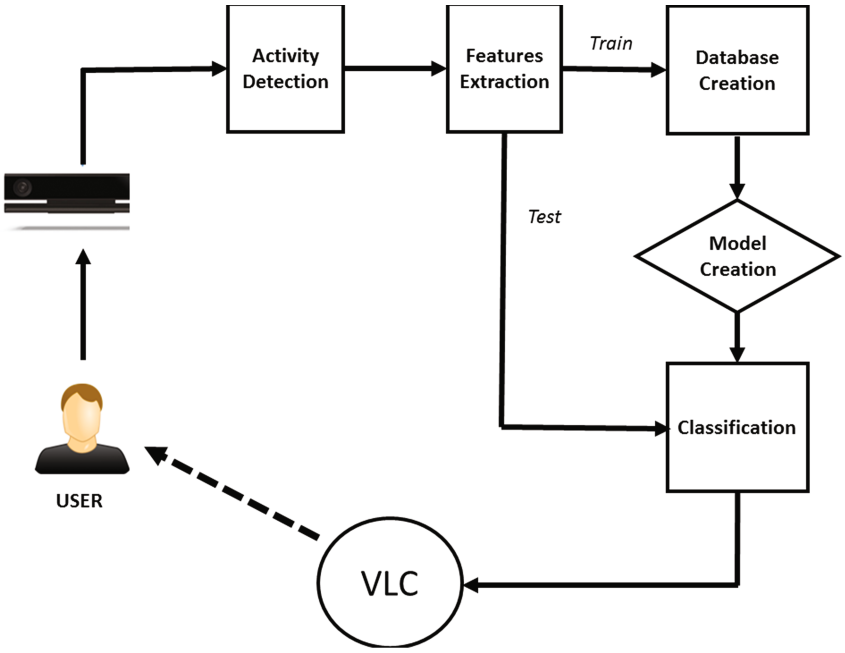

Fig. 8. A diagram illustrating the main modules of the prototype and how they are used. There are 2 modes: training and testing/real use. In training mode features extracted are stored in a database, used to train the classifiers.

5.3 Activity Detection

The first step is the Activity Detection. In this step the system searches for the face of the speaker, when found (with Microsoft Kinect SDK), a rectangular box is drawn surrounding the face of the user from every frame that arrives from the Kinect Camera. The SDK provides additional information like if the speaker is happy, wearing glasses, if the right or left eye is closed etc., as shown in Fig. 9.

For the acquisition to start right from the start of the word, the following process was adopted: first the user has to have the face and lips stable for around one second. Then, the system informs the speaker that it is ready to record a word by displaying a text message and making the window background green in order to be easier for the speaker to see this state change. In the state of Ready to Record, the system starts recording as soon as the speaker opens his mouth (information obtained with Kinect SDK [20]). This is used as an indication that the speaker may utter a command. With the same approach, the system stops recording when the speaker has the face and lips stable for at least one second.

5.4 Features

To build our system we selected the position of the lips and chin as the features for our classifier.

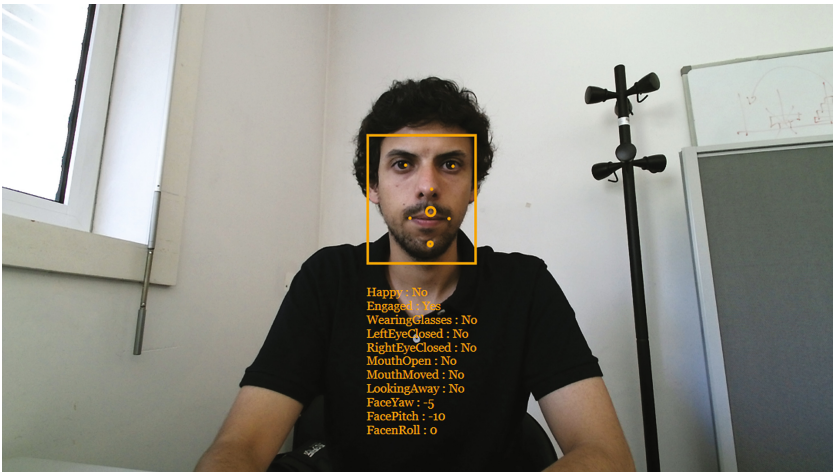

Fig. 9. Face detection by Kinect and other speaker information shown.

In more detail, we extract the position of the lips given by the distance between the upper and lower lip (height) and the distance between the left and right corners (width), the protrusion of the lips (upper lip and bottom lip) and the chin position (x and y coordinates). The position of the lips was chosen because it has proven to give good results in previous works [1]. The chin position was added because of the role of the lower jaw in the human speech production process. To obtain these 6 features we used Kinect SDK, namely the `HighDetailFacePoints` in `Kinect20.face.lib` [21] (Fig. 10).

In order to deal with the different distances between the speaker and the recording device a z-score normalization is applied. To facilitate the Classification stage, we assume a fixed length of 2 s for feature vectors (resulting in feature vectors of 60 dimensions at a 30 fps recording rate).

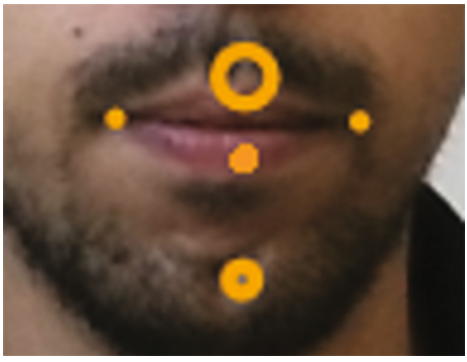

Fig. 10. Points tracked in mouth and chin for feature extraction.

5.5 Classifiers

In terms of classifiers, the Support Vector Machine (SVM), Random Forest, Sequential Minimal Optimization (SMO), AdaBoost and Naive Bayes algorithms available in Weka [34] were evaluated, offline, with databases recorded using the train path of the developed system. We used a linear Kernel for the SVM classifier. This initial list of classifiers resulted from the authors' previous experience in classification tasks in SSI and speech segmentation, such as [26].

As the speed of the algorithm is mandatory for real usage, three classifiers were chosen based in performance and speed in those evaluations: Random Forest, SMO and Naive Bayes Algorithms.

A Winner Take All approach was adopted to combine the decision of these 3 classifiers.

6 Evaluation

Besides evaluating the influence of each classifier and how the distance of the user to the Kinect affected the results, evaluation not included in this paper, the prototype was also tested live with three users. This first evaluation consisted in classifying a word in real time for VLC controlling purposes and was aimed at getting some more insight regarding the system performance to inform future improvements. Speaker dependency of the system was also tested, training the system with a database recorded for one speaker and testing with another. Information regarding participants, databases recorded for training and obtained results are presented in the next subsections.

6.1 Participants

Three persons participated in the evaluation of the system: (a) one of the authors, an Engineering post-graduate student, 23 years old, male; (b) a 22 year old male, also a student of the same course; and (c) a 29 years old female with an MSc in Gerontology and a PhD in Science and Health Technologies, natural from Madeira island, Portugal, and speaking with the regional accent.

6.2 Databases for Training

To train the classifiers to be used in the live evaluation, five different databases were created: 3 databases for Speaker 1 (each recorded at a different distance – 0.6 m, 1 m and 2 m away from the Kinect Camera); 1 database for Speaker 2; and 1 database for Speaker 3. Speakers 2 and 3 recorded at 1 m from the Kinect camera. The databases were recorded at a research lab in low noise conditions. Speaker 1 recorded all the databases without producing audible speech (silent speech) and Speakers 2 and 3 recorded the databases pronouncing the words.

6.3 Results for the Live Evaluation

Live evaluation consisted in classifying a word in real time for VLC controlling purposes. The first tests were performed in matching conditions of test and train regarding speaker and distance (i.e., same speaker and distance in test and database used for training). After, the effect of distance was assessed followed by some speaker dependency tests, assessing if the developed system could perform well when trained with data from other speakers.

Matching Conditions

The results obtained, in terms of hits and misses of the commands are presented in Table 2. Different distances were tested for Speaker 1 since 3 databases were recorded for him.

Table 2. Performance of the system in live evaluation with 3 speakers in matched conditions (test and train using data recorded for the same speaker and distance).

Test speaker	Distance (m)	Hit	Miss	Hit (%)
Speaker 1	0.6	52	28	65.0
	1	44	36	55.0
	2	56	24	70.0
Speaker 2	1	46	34	57.5
Speaker 3	1	25	55	31.3

The best result was achieved for Speaker 1 at a distance of 2 m away from the Kinect, with 70% of correctly detected commands (hits). Speaker 3 had the worst results, possibly influenced by her accent.

Effect of User's Distance to the Kinect

To test the distance dependency, Speaker 1 tested at 2 different distances with classifiers trained with databases recorded at other distances. The following combinations were used: the speaker at 0.6 m from the Kinect and the classifiers trained with the data at 1 m and 2 m; the speaker at 1 m from the Kinect and train data recorded at 0.6 m and 2 m. The results can be seen in Table 3.

The results show evidence that the distance is not an issue (the hits are similar to the ones obtained in Table 2) and show that distance normalization is capable of handling the different user-Kinect distances of a typical AAL scenario.

In Table 3 the best live performance of this work was obtained (81.3%) with the Speaker at 1 m from the Kinect and the training database recorded at 2 m.

Table 3. Effect of mismatch in distance between live test conditions and the databases used to train the system classifiers, for Speaker 1.

Speaker distance (m)	Classifiers trained with (m)	Hit	Miss	Hit (%)
0.6	1	42	38	52.5
	2	44	36	55.0
1	0.6	49	31	61.3
	2	65	15	**81.3**

Speaker Dependency

To finish the evaluation, the speaker dependency of the system was tested. The objective was to understand if the system can be used by an user that has no training data. In other words, if the system can perform with Speaker X, in the Test part, against Speaker Y's data uses for training.

Three tests were made: Speaker 1, at 1 m from the Kinect, with classifiers trained with the databases of Speaker 2 and Speaker 3; Speaker 2, at 1 m from the Kinect, but using classifiers trained with Speaker 1's database, also at 1 m. The results are presented in Table 4.

Table 4. Results regarding speaker dependency tests. Tests by Speaker 1 and 2 with classifiers trained with databases of other speakers.

Tested by	Train database	Hit	Miss	Hit (%)
Speaker1	Speaker 2	14	66	17.5
	Speaker 3	12	68	15.0
Speaker 2	Speaker 1	17	63	21.3

The results shown that the system's accuracy decreases dramatically in comparison to the results obtained for test and train with the same speaker. Analyzing the results presented in Table 4, we can conclude that the system is clearly speaker dependent.

7 Conclusion

In AAL scenarios, means are often needed to control media applications in noisy environments, such as a living room. Thus, this paper describes a first working SSI prototype for Portuguese, potentially relevant for older adults, which allows the control of a media player application at multiple distances using (silent) speech, with promising results.

The developed prototype is divided into the following parts: activity detection (automatic recording based on the movement of the lips), feature extraction, train of classifiers, classification and integration with VLC player. The Microsoft Kinect for Windows was used to capture visual information of the face.

Three different adults, with different ages, genres and accents, tested the system. Using different databases with recordings from each of them, we evaluated different distances between the Kinect and the user, as well as speaker dependency of our solution. The system revealed good performance in real time control of VLC, with an accuracy of 81.3% and 1.3 s taken to perform a classification.

The results show some variation among the users that participated in the study. Some pronounce the words slowly with hyper articulation, others pronounce them fast with small movements of the lips. The system performs better if the words are correctly articulated during all the repetitions and if the words are correctly recorded during the 2 s available to extract the features from the lips and chin. The effect of distance of the speaker was also tested, proving not to be an issue in terms of the system's accuracy.

7.1 Future Work

In terms of future work there are several open possibilities, starting by the control of other relevant applications for AAL scenarios, such as Skype, Youtube, Facebook or Spotify.

Despite the good performance, improvements can be made to the process of command detection, to start recording, and use of a fixed recording time, contributing to an even more natural usage.

The developed system is speaker dependent. Even though it is already useful in many scenarios, and the recording of data for a new speaker is quite simple, evolution to a speaker independent system should be considered.

The evaluation reported, even though it serves the purpose of informing further development of the system, is quite limited. Extended evaluation is needed, and should be implemented to enable a more thorough evaluation of the next prototypes, first with non-elderly and, as soon as system is robust enough, with elderly.

The system created is a Visual Speech Recognition system, non-invasive and non-obtrusive. However, it would be interesting to create and evaluate a multimodal system combining the features used in the created prototype with features from other phases of the humans speech production.

Acknowledgements. Research partially funded by IEETA Research Unit funding (UID/CEC/00127/2013) and Marie Curie Actions IRIS (ref. 610986, FP7-PEOPLE-2013-IAPP). Samuel Silva acknowledges funding from FCT grant SFRH/BPD/108151/2015. The authors also thank the participants in the evaluation.

References

1. Abreu, H.: Visual speech recognition for European Portuguese. Master thesis, Universidade do Minho (2014)
2. Bradski, G., Kaehler, A.: Learning OpenCV: computer vision with the OpenCV library. O'Reilly Media, Inc. (2008)
3. Brumberg, J.S., Nieto-Castanon, A., Kennedy, P.R., Guenther, F.H.: Brain-computer interfaces for speech communication. Speech Commun. **52**(4), 367–379 (2010). http://dx.doi.org/10.1016/j.specom.2010.01.001
4. Dalka, P., Bratoszewski, P., Czyzewski, A.: Visual lip contour detection for the purpose of speech recognition. In: Proceedings of the International Signals and Electronic Systems (ICSES) Conference, pp. 1–4, September 2014
5. De Smedt, K.: 11 computational models of incremental grammatical encoding. In: Computational Psycholinguistics: AI and Connectionist Models of Human Language Processing, pp. 279–307 (1996)
6. Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J.M., Brumberg, J.S.: Silent speech interfaces. Speech Commun. **52**(4), 270–287 (2010)
7. Freitas, J., Candeias, S., Dias, M.S., Lleida, E., Ortega, A., Teixeira, A., Orvalho, V.: The IRIS project: a liaison between industry and academia towards natural multimodal communication. In: Proceedings of the IberSPeech. Las Palmas, Spain (2014)
8. Freitas, J., Teixeira, A., Sales Dias, M., Silva, S.: An Introduction to Silent Speech Interfaces. Springer, Heidelberg (2016)
9. Freitas, J., Teixeira, A., Bastos, C., Dias, M.: Towards a multimodal silent speech interface for European Portuguese. In: Speech Technologies, pp. 125–149. InTech (2011)
10. Freitas, J., Teixeira, A., Dias, M.S.: Towards a silent speech interface for portuguese. In: Proceedings o the Biosignals, pp. 91–100 (2012)
11. Freitas, J., Teixeira, A., Dias, M.S.: Multimodal silent speech interface based on video, depth, surface electromyography and ultrasonic doppler: data collection and first recognition results. In: International Workshop on Speech Production in Automatic Speech Recognition (2013)
12. Freitas, J., Teixeira, A.J., Dias, M.S.: Multimodal corpora for silent speech interaction. In: LREC, pp. 4507–4511 (2014)
13. Frisky, A.Z.K., Wang, C.Y., Santoso, A., Wang, J.C.: Lip-based visual speech recognition system. In: Proceedings of the International Security Technology (ICCST) Carnahan Conference, pp. 315–319, September 2015
14. Galatas, G., Potamianos, G., Makedon, F.: Audio-visual speech recognition incorporating facial depth information captured by the kinect. In: 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO), pp. 2714–2717. IEEE (2012)
15. Gokturk, S.B., Yalcin, H., Bamji, C.: A time-of-flight depth sensor-system description, issues and solutions. In: Conference on Computer Vision and Pattern Recognition Workshopp, CVPRW 2004, pp. 35–35. IEEE (2004)
16. Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D.: Rgb-d mapping: using kinect-style depth cameras for dense 3D modeling of indoor environments. Int. J. Robot. Res. **31**(5), 647–663 (2012)
17. Lanaria, V.: VLC, the world's most popular media player, turns 15 years old: here's why you should download it now (2016)
18. Levelt, W.J.: Models of word production. Trends Cogn. Sci. **3**(6), 223–232 (1999)

19. Matsumoto, M.: Silent speech decoder using adaptive collection. In: Proceedings of the Companion Publication of the 19th International Conference on Intelligent User Interfaces, IUI Companion 2014, ACM, New York, pp. 73–76 (2014). http://doi.acm.org/10.1145/2559184.2559190
20. Microsoft: Face tracking (2016). https://msdn.microsoft.com/pt-pt/library/dn782034.aspx
21. Microsoft: high detail face points (2016). https://msdn.microsoft.com/en-us/library/microsoft.kinect.face.highdetailfacepoints
22. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Efficient model-based 3D tracking of hand articulations using kinect. In: BmVC, vol. 1, p. 3 (2011)
23. Porbadnigk, A., Wester, M., p Calliess, J., Schultz, T.: Eeg-based speech recognition impact of temporal effects. In: 2nd International Conference on Bio-inspired Systems and Signal Processing (Biosignals 2009) (2009)
24. Rabiner, L., Juang, B.H.: Fundamentals of speech recognition. Prentice hall, Englewood Cliffs (1993)
25. Rao, R.A., Mersereau, R.M.: Lip modeling for visual speech recognition. In: Proceedings of the Conference on Signals, Systems and Computers Record of the Twenty-Eighth Asilomar Conference vol. 1, pp. 587–590, 1 October 1994
26. Rodriguez, Y.L., Teixeira, A.: On the detection and classification of frames from European Portuguese oral and nasal vowels. In: Proceedings of the FALA 2010 (2010)
27. Saenko, K., Darrell, T., Glass, J.R.: Articulatory features for robust visual speech recognition. In: Proceedings of the 6th International Conference on Multimodal Interfaces, ICMI 2004, ACM, New York, pp. 152–158 (2004). http://doi.acm.org/10.1145/1027933.1027960
28. Sahni, H., Bedri, A., Reyes, G., Thukral, P., Guo, Z., Starner, T., Ghovanloo, M.: The tongue and ear interface: a wearable system for silent speech recognition. In: Proceedings of the 2014 ACM International Symposium on Wearable Computers, ISWC 2014, ACM, New York, pp. 47–54 (2014). http://doi.acm.org/10.1145/2634317.2634322
29. Seikel, J.A., King, D.W., Drumright, D.G.: Anatomy and physiology for speech, language, and hearing. Delmar Learning, 4th edn. (2009)
30. Teixeira, A., Almeida, N., Pereira, C., Silva, M., Vieira, D., Silva, S.: Applications of the multimodal interaction architecture in ambient assisted living. In: Dahl, D. (ed.) Multimodal Interaction with W3C Standards: Towards Natural User Interfaces to Everything, pp. 271–291. Springer, New York (2016)
31. TeraRanger: Time-of-flight principle (2016). http://www.teraranger.com/technology/time-of-flight-principle/
32. Wand, M., Koutn, J., et al.: Lipreading with long short-term memory. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6115–6119. IEEE (2016)
33. Werda, S., Mahdi, W., Hamadou, A.B.: Lip localization and viseme classification for visual speech recognition. arXiv preprint arXiv:1301.4558 (2007)
34. Witten, I.H., Frank, E., Hall, M.A.: Data Mining - Practical Machine Learning Tools and Techniques, 3rd edn. Morgan Kaufmann, San Francisco (2011)
35. Yargic, A., Dogan, M.: A lip reading application on MS Kinect camera. In: 2013 IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA), pp. 1–5. IEEE (2013)