Adapting Speech Recognition in Augmented Reality for Mobile Devices in Outdoor Environments*

Rui Pascoal[1], Ricardo Ribeiro[2], Fernando Batista[3], and Ana de Almeida[4]

1 **Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal**
2 **Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal; and INESC-ID Lisboa, Lisbon, Portugal**
3 **Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal; and INESC-ID Lisboa, Lisbon, Portugal**
4 **Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal; and CISUC Centro de Informática e Sistemas da Universidade de Coimbra, Coimbra, Portugal**

Abstract

This paper describes the process of integrating automatic speech recognition (ASR) into a mobile application and explores the benefits and challenges of integrating speech with augmented reality (AR) in outdoor environments. The augmented reality allows end-users to interact with the information displayed and perform tasks, while increasing the user's perception about the real world by adding virtual information to it. Speech is the most natural way of communication: it allows hands-free interaction and may allow end-users to quickly and easily access a range of features available. Speech recognition technology is often available in most of the current mobile devices, but it often uses Internet to receive the corresponding transcript from remote servers, e.g., Google speech recognition. However, in some outdoor environments, Internet is not always available or may be offered at poor quality. We integrated an off-line automatic speech recognition module into an AR application for outdoor usage that does not require Internet. Currently, speech interaction is used within the application to access five different features, namely: to take a photo, shoot a film, communicate, messaging related tasks, and to request information, either geographic, biometric, or climatic. The application makes available solutions to manage and interact with the mobile device, offering good usability. We have compared the online and off-line speech recognition systems in order to assess their adequacy to the tasks. Both systems were tested under different conditions, commonly found in outdoor environments, such as: Internet access quality, presence of noise, and distractions.

1998 ACM Subject Classification I.2.7: Natural Language Processing

Keywords and phrases Speech Recognition, Natural Language Processing, Sphinx for Mobile Devices, Augmented Reality, Outdoor Environments

Digital Object Identifier 10.4230/OASIcs.SLATE.2017.21

* This work was supported by national funds through Fundação para a Ciência e Tecnologia (FCT) with reference UID/CEC/50021/2013.

1 Introduction

New technologies related with Machine Learning, Natural Language Processing (NLP) and Augmented Reality (AR) are currently arising. One of the world's leading information technology research and advisory company – Gartner, Inc. – predicts that, after a long period of technological development and refinement, the implementation of advanced Machine Learning technologies and conversational systems solutions for AR applications are achieving its peak.

Augmented Reality (AR) is an example of what Fred Brooks[1] calls "Amplification Intelligence": the use of the computer as a tool to perform human tasks in an easier way. Specifically, AR can be used to perform tasks more intuitively, and efficiently interact with the information displayed in a screen. Natural Language Processing (NLP) providing means to help with this interaction. As stated by Chowdhury [5], NLP is a research and application area that explores how computers can be used to understand and manipulate speech and natural language text to do useful things. Thus, wise interactions with AR technology are needed [10]. A richer end-user experience involves speech interaction rather than simply reading or writing at a screen. Because they include cameras, a smartphone or smart glass can run AR applications, and that led to accelerating the development of innovative AR [15] applications. Our goal is to address the end-users interaction needs for a quick augmented reality technology, especially for outdoor activities, such as walking, cycling, etc.

There are benefits of using automatic speech recognition (ASR) with AR technology in outdoor environments [15], but such environments involve a number of additional challenges. Speech recognition in an augmented reality interface contributes to the efficiency, intelligent information and communication use, good perception, and "common sense", as mentioned by Barry [3] and Ronald Azuma [2] in outdoor environments [1].

In a previous research we have developed an AR prototype and performed empirical tests with 12 end-users, revealing that speech provided a more quick interaction than gestual, where users were able to perform operations about 1 to 2 seconds faster in average [14]. Our current goal is to integrate an off-line speech recognition system in the AR mobile application. We have adopted the open source CMU Sphinx[2] system [13],[12]. The Sphinx-4 speech recognition system has been jointly developed by Carnegie Mellon University, Sun Microsystems Laboratories, and Mitsubishi Electric Research Laboratories (MERL) [12]. It does not need Internet access and it should also provide a socially acceptable interface, natural to interact with [1], overcoming the AR technology challenges [2].

The work performed in the scope of this paper provides access to five interface features for outdoor usage. It allows requesting geographical, biometric, and climatic information without communication overload [14]. Our solution is intended to provide a good usability while managing and interacting with the information on the AR application. We have followed available receipts on how to integrate Sphinx in the Android operating system. Such operating system is the starting point for our future plans on mobile augmented reality devices, also based on the android operating system, including Recon Jet™ glasses[3] or Epson Moverio BT-200™[4]. CMU sphinx speech recognition can also be equally integrated in other mobile Operating Systems, such as Windows or iOS. This module does not require Internet access, and provides means to address the noise observed in outdoor environments.

---

[1] Frederick Phillips Brooks, software engineer and computer, known by the project OS/360 operating system developed by IBM for the System/360 mainframe. He wrote a book of The Mythical Man-Month.

[2] `http://cmusphinx.sourceforge.net`

[3] `https://reconinstruments.com/products/jet/`

[4] `https://tech.moverio.epson.com/en/bt-200/`

The following questions arise when speech recognition is used in augmented reality technology in mobile operating systems and in outdoor environments.

- **Question 1: Is the automatic recognition system for mobile AR devices efficient enough in outdoor environments, containing phenomena, such as noise and distractions?** In fact, some problems may arise, such as the user remembering what are the possible requests, or the noisy conditions that may prevent the system from correctly recognizing the information provided.
- **Question 2: How is the performance achieved with Sphinx speech recognition instead of using a web-based speech recognition system?** This is a mandatory question, because in outdoor environments we may have Internet connection constraints and the default speech recognition system may not be effective to respond to the execution of hands-free features.

Other issues and difficulties arise when speech signal is corrupted by many sources, e.g., the wind is bad for performance of recognition system, because wind speed does interference with the sound input of the microphone. In addition the system has to cope with non-grammaticality of spoken communication and ambiguity of language [17], e.g., there are several ways of saying "i want to take a picture". The current challenge is an augmented reality system being able to interpret several ways to request operations by end-user speech.

This paper is structured as follows. Section 2 presents the Related Work. Section 3 describes our system architecture. Section 4 discusses the results and the associated difficulties. Finally, Section 5 presents the major conclusions and point out future research directions.

2 Related Work

AR is a new technology, but should not be categorized as mere technology. Instead, AR is an advanced computer interface, as mentioned by Alan Craig [6], which development started more than forty years ago. Still, there is a strong requirement to be adopted to people, being required usability of technology (as mentioned by Sawyer [16]), in various areas of society.

In previous work[5], the first author used a classification taxonomy of the different kind of environments that may exist abroad (as mentioned by Pascoal and Guerreiro [14]), and in these various environments performed tests with end-users, e.g., in the following four environments: (i) silent; (ii) silent with distraction; (iii) noisy with distraction; and, (iv) very noisy with distraction. That suggested environment taxonomy involves the factors noise and distractions. It helps to cluster results of end-user tests with a speech recognition prototype. The noise can be traffic, industry, animals, or wind speed, and so on. The distractions can be movement of people, animals, information overload, or forgetting system keywords.

By analyzing the various systems of speech recognition developed in recent years, anyone can identify that the software and hardware architecture adopted between them differs widely. However one difference is, e.g., Google speech recognition[6] needs Internet access, but CMU Sphinx does not need Internet access. That's why the developer will focus primarily on

[5] Chapter 12: Information Overload in Augmented Reality – The Outdoor Sports Environments, from book: Information and communication overload in digital age (2017). `www.igi-global.com`, as mentioned by Pascoal and Guerreiro [14] in Information and Communication Overload in the Digital Age (pp. 271–301).

[6] The Google speech recognition or Google Cloud Speech API, which enables developers to convert audio to text by applying powerful neural network models in an easy to use API, available at `https://cloud.google.com/speech/`.

Sphinx. However, the authors will show some differences between Sphinx and Google Speech. At the same time, they will perform tests with both tools (Sphinx vs Google speech), because the ultimate goal is to extract the best and the most suitable of these two systems.

Finally, the mission is to contribute to the implementation of a mobile AR system, which has the aim of being used outside, overcoming the constraints and limitations of current mobile AR applications[7]. Moreover, requirements faced by the application developers were identified, e.g., to overcome the technological and environmental limitations, because they are clearly interrelated. In addition to the human limitations in understanding due to information overload [4], restrictions also often relate to the limited capabilities of mobile devices, and the fact that AR equipment should be usable in a wide range of environmental conditions, as mentioned by Ronald Azuma [1, 2].

2.1 Potential Distractions in Outdoor Environments

What happens most often is that, in the case of an outdoor end-user like a cyclist when overloaded with distractions moves more slowly, and cannot keep in proper lane of the road correctly, but to compensate the risk of collision, experienced athletes psychologically maintain, or safeguard, for maintaining a greater distance of other cyclists and other obstacles there is ahead. It is a self-protection to reduce accidents. The conclusion is that attention is higher when using a voice interface, e.g., to send messages compared to text messages with hands. However, although attention is higher, the conduction is still impaired, therefore, there is a cognitive interference previewing messages, as mentioned by Sawyer [16].

If end-users are distracted with information overload on a smartphone or smart glasses or distracted with environment they cannot be remember of the keywords to interact with an AR application. This is what happened when Pascoal and Guerreiro [14] executed quantitative approach tests at twelve end-users using an AR prototype. They saw a task execution degradation (fifth task – "ok agent") immediately after the another task (fourth task – "ok message"), i.e., six users have not complied with expectations (i.e., fifty per cent), but the previous task the execution was successful with everybody.

Now, on this work, the authors also evaluate with a quantitative approach, the performance of users executing only one task, but with two kind of automatic speech recognizers (Sphinx versus Google) and in distinct environments, some are noisy and with distractions and others are silent, but all running outside and with normal Internet access.

2.2 Human-Computer Interaction with Speech Recognition

The human-computer interaction must be a natural interface, meaning the interfaces of AR applications must be intuitive for users and easily controlled using the natural human movements. For hands-free, a good method could be with microphone interaction, by keywords, like "computer" or "photo". The quantitative methods shown in the Discussion section are faster and usable for end-users.

The authors use a method with some kind of reverse word stemming, where all words with a common root are mapped from a single word (e.g., photo): all instances of photographing and beyond as "i would like a photo please" and so on, are mapped into "photo", because "photo" is a single infinitive concept. The authors used this method because by experience on

[7] Other constrains are memory, storage capacity, battery autonomy and bandwidth on embedded devices are also very limited. For these reasons, has concentrated on simple tasks with restrictive grammars, as mentioned for David Huggins et al. [9].

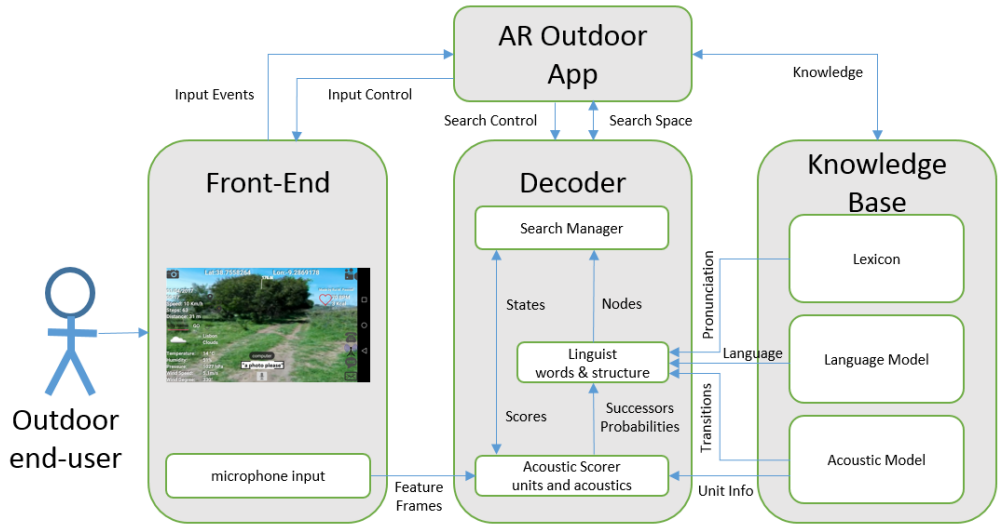

Figure 1 Sphinx System Architecture [12, 13].

field, most end-users do not remember, or do not want to remember, rigid phrases or words: it is more cognitively easy and appropriate to have several ways of saying the same, and this is where probabilistic learning for natural language processing comes to help.

These kinds of human-computer interactions have been discussed in the International Symposium on Mixed and Augmented Reality (ISMAR), the leading international academic conference in the field of Augmented Reality and Mixed Reality. To create the best human-computer interaction, in other words, mobile human-device AR abroad with AR applications, that is, to give the user the ability to walk around large environments, outdoor is essential good guidance tracking abroad, as mentioned by Ronald Azuma [1] [2] and by Alan Craig [6]. Tangible AR interaction naturally leads to combining real object input with gesture and voice interaction, which often leads to multimodal interfaces, as shown at "A Review of Ten Years of ISMAR". They also discussed about AR technology interaction with speech commands, i.e., the survey work giving an overview of recent research in the field and conducts deploy of AR interactions with voice commands, as mentioned by Feng Zhou et al, from University of Canterbury, New Zealand [18].

3 System Architecture

Figure 1 shows the architecture of Sphinx and Figure 2 shows how are structured the possible interactions a end-user can have with the AR application using Sphinx as the automatic speech recognition module.

In what concerns the architecture of Sphinx (Figure 1), the speech signal is parameterized at the Front-End module, which communicates the derived features to the Decoder block. This block has three components: the search manager, the linguist, and the acoustic scorer. These work in tandem to perform the decoding. Inside of the Front-End there are several communicating blocks, each with an input and an output, linked to the output of its predecessor. When a block is ready for more data, it reads data from the predecessor and interprets it to find out if the incoming information is speech data or a control signal. A control signal might indicate the beginning or end of speech – important for the Decoder

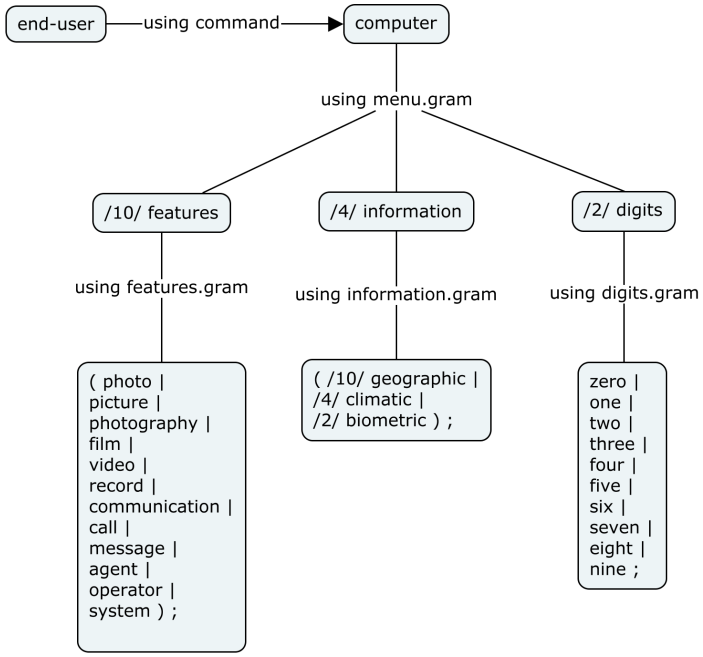

Figure 2 Search tree of possibilities for the best hypothesis.

– or might indicate data dropped or some other problem. If the incoming data is speech, it is processed and the output is buffered, waiting for the successor block to request it. Additional blocks can also be introduced between any two blocks, to permit noise cancellation or compensation on the signal. Inside of the Decoder we have the Search Manager, the Linguist, and the Acoustic Scorer. The Search Manager has as primary function to construct and search a tree of possibilities for the best hypothesis. The construction of the search tree is done based on information obtained from the Linguist. In addition, the Search Manager communicates with the Acoustic Scorer to obtain acoustic scores for incoming data. The Linguist translates linguistic constraints provided to the system into an internal data structure called the grammar, which is usable by the Search Manager. The Acoustic Scorer has the task to compute the state output probability or density values for the various states, for any given input vector. It also provides these scores on demand to the search module. In order to compute these scores, the Scorer must communicate with the Front-End module to obtain the features for which the scores must be computed [12].

The authors observed and fit the AR application during interaction tests with five words and some equivalent sentences to the root word (e.g., “photo” = “a photo, please” or “photo” = “please, i would like a photo, thank you”), the difficulties to running the following solicited tasks: photo, film, communication, message, agent, biometric, climatic, and geographic.

Next, an abstraction through a frame conceptual map for speech recognition with these root words, i.e., keywords used by end-users when interact with the AR features.

The tree Figure 2, based on information obtained from the linguist, consists in all active paths in the search. The linguist translates linguistic constraints provided to the system into an internal data structure called the grammar. The numbers inserted in menu.gram mean weights of importance, e.g., the word “features” has a weight of 10, “digits” has a weight of 2 (worst execution priority), and so on. Figure 2 shows another detail, it’s the parenthesis in grammars of features and information. Only one word can be requested, and

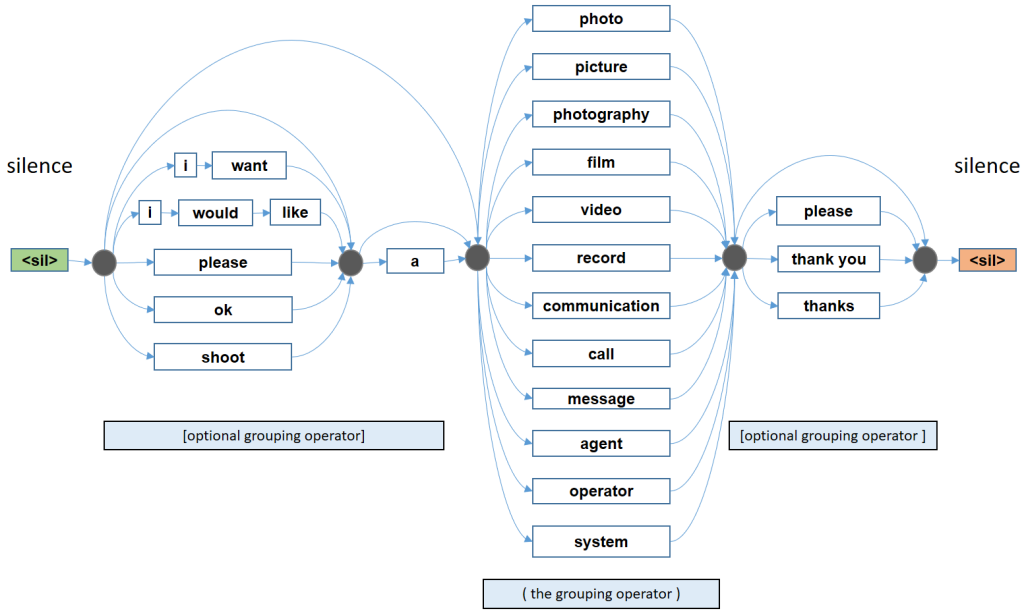

Figure 3 Grammar graph to execute features (made by authors).

in information.gram has another detail it is the associated weight, i.e., "geographic" word is more likely to be chosen than "biometric" word because has only a weight of 2.

However, as previously mentioned, the authors will use a method with some kind of reverse word stemming, where all words with a common root are mapped from a single word (e.g., "photo"). And all instances of photographing as "a photo, please" and so on, are mapped to "photo", because "photo" is a single infinitive concept. Moreover, the end-user can also use other sentences if he/she likes or if he/she remembers. See Figure 3.

However, AR can be applied in every sense, not only visually. The usual researchers of AR fields, focused on mixing images and graphics real and virtual. However, *AR can be extended to include sound. Users can use headsets equipped with microphones*, as mentioned by Ronald Azuma [1, 2].

Next, we will see all instances of the word "photo" in a text. This is a process to check if string (the word spoken by the user) matches with defined grammar, as mentioned by L. Karttunen [11] and Walker et al. [17].

The grammar created by authors has simple words, like "photo", "film", or "biometrics", because the ability of human kind for long words is smaller than for short words. In general, the memory capacity for verbal contexts – digits, letters, words, and so on – strongly depends on the time it takes to speak aloud content and lexical function of the content, i.e., if the contents are known words or not. Several other factors also affect the measure of a person's memory and so it is difficult to establish the capacity of short-term memory by several chunks. That is why, in 2001, Nelson Cowan proposed that the activity of memory has a capacity of about four chunks in young adults (and lower in older children and adults).

Therefore, the authors specified grammars with key words and key sentences to execute associated methods (e.g., features grammar and information grammar). The recognition system must accept at least ten sentences or more, consisting of several words, which allow access to the five features (photo, film, communication, message, agent). To run this five features and programmatically speaking, authors suggest an if-else condition with a

Listing 1 Defined JSGF grammar for features.

```
grammar features;

public <features> = <startCommand> <mainCommand> <endCommand> ;
<startCommand> = [i want|i would like|please|ok|shoot] [a];

<mainCommand> = (photo|picture|photography|film|video|record|
                 communication|call|message|
                 agent|operator|system) ;

<endCommand> = [please|thanks|thank you];
```

conditional "OR", e.g., to execute photo method. This is hard coded, but this could be avoided by simplify code, when using JSGF defined grammar[8]. See Listing 1 for the defined JSGF grammar for features. This grammar will simplify a wide range of equivalent sentences, which are, several possibilities of saying the same, in various ways, like "i would like a picture please", or "please a call thanks". See also Figure 3.

Another situation that the authors encountered when developing the augmented reality application and when unit tests were in execution was the running of undesired features without being ordered to run. This can be a serious problem during search, as mentioned by Paul Lamere et al. [13]. It is a pruning problem encountered by the search module in decoding parallel feature streams. The pruning is based on combined scores, paths with different contributions from the multiple feature streams get compared for pruning [13]. To break or reduce these pruning problems, the developed features on AR application are being combined in a weighted manner, with weights that can be more easily controlled, e.g., the "film" feature has low weight than the "photo" feature, and the "stop" word to exit of some features and return for "main activity" should have a little less. This is a specific algorithm for application learning. The result generated will provide by the search module is in the form of a tree, which can be queried for the best recognition hypotheses, or a set of hypotheses.

3.1 Information and Features Management Given to Outdoor End-User

Based on earlier subtopics and on the point of view of an outdoor end-user, the main question of this subsection, is how to get all relevant information with minimal effort and how to minimize dependency of communication with the Internet network without information overload [14]. Also, to get relevant features easily, in other words, is the Sphinx Android application efficient enough in outdoor environments, e.g., silent, silent with distractions, noise with distractions and very noise with distractions?

Next are described in detail the three groups "suggested" of informative data that could be submitted to the outdoor end-users:

1. **Climatic data:** involves temperature, atmospheric pressure, altitude, and relative humidity. This data serves, not only to inform, but also to intelligently calculate together with the user's health status data.
2. **Biometric data:** involves the heart rate and calorie expenditure. These are important data to calculate and provide vital advices. Without this sensor, it will not be possible to deliver alerts, which could make the difference.

[8] The rules of Java™ Speech API Grammar Format – can be found at `http://www.w3.org/TR/jsgf/`.

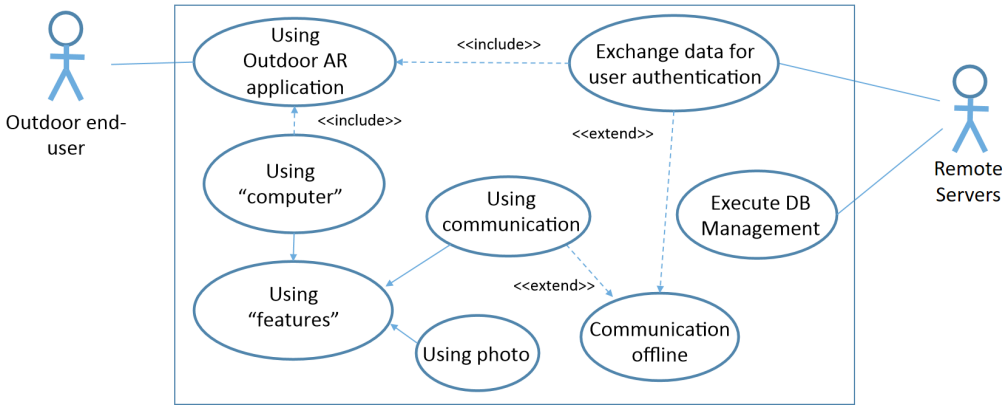

Figure 4 UML use case diagram for outdoor end-user to take a photo with hands-free.

3. **Geographic data:** involves the global position system (GPS), compass, and stopwatch, also, involves speed, measured steps, and distances, initial and final positions. Geolocation serves not only to inform, but also to index the server database, events, and points of interest. Can be a good tracker of an user in motion and what their cadence or rhythm is.

To show end-users' interactions with an AR system, is presented the following particular Use Case Diagram (Figure 4). It is an abstraction of outdoor end-users when they take a photo with hands-free. Previously, users may be authenticated on the server[9].

4 Discussion

Spoken language processing is a diverse subject that relies on knowledge of many levels, including acoustics, phonology, phonetics, linguistics, semantics, pragmatics, and discourse. Difficulties arise when speech signal corrupted by many sources (environmental noises). There are observed issues and difficulties, arise when speech signal corrupted by many sources, e.g., the wind is bad for performance of recognition system, because wind speed does interference with the sound input of the microphone used by every users. Also another aspect, we must take into account a human cognitive limitation, e.g., when users are receiving information and several tasks to perform, can lead to information and communication overload [14] [4]. In addition the system has to cope with non-grammaticality of spoken communication and ambiguity of language. Also, this is a huge field, because of the diverse nature of spoken language processing requires knowledge in computer science, electrical engineering, mathematics, syntax, and psychology, as mentioned by Willie Walker et al. [17].

The main task that was requested to the end-users was the execution of the photo functionality, according to the grammatical probability of the request to run a photo method, e.g., "please i would like a photo thank you".

So, with a quantitative and qualitative treatment of research hypotheses are implemented the user's information, as next, an AR prototype that simulates an outdoor environment to collect data, through observation of interaction tests with fourteen end-users. Next will show reaction's results with noise and distractions for better usability. See next Figure 5.

[9] This authentication is a future development of the authors. This is a requirement to track a particular user, a help to him, but needs Internet connection.

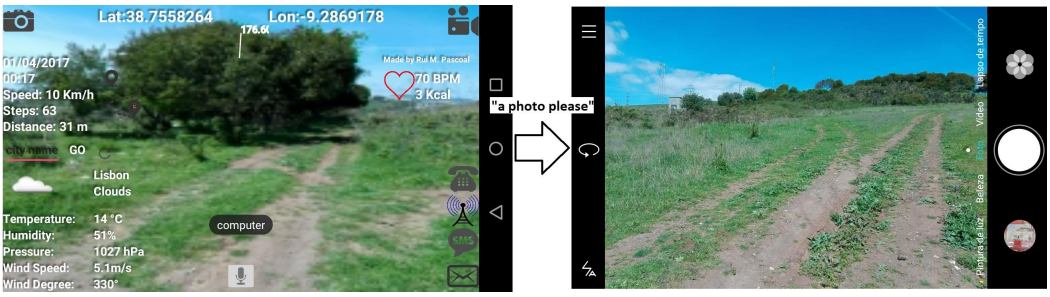

Figure 5 Execution of photo feature in AR prototype with recognition system.

Tests up through a practical implementation output, the provision of the information given to end-users, wise information[10], in which case the use is preferably an Android smartphone, because it is portable and mobile [6].

This application had some difficulties when implementing Sphinx on android, e.g., there were difficulties in executing functionalities through speech recognition, to execute the associated methods (`startCameraFoto()`), it had to be through the following condition:

Previously the grammars had to be built, for the recognition of the words to be used in the interaction with the android application, as well as the static variables, with the key words of access to the three grammars:

- "features"
- "digits"
- "information"

The recommendations for researchers and future researchers with the influencing factor of outdoor environments are as follows. These recommendations are based on the difficulties experienced by the authors, as well as on the results of empirical field tests with some end-users.

The Sphinx Android prototype served to analyze real interactions in outdoor environments. This empirical research was conducted to obtain a quantitative approach. Afterwards, a questionnaire was applied to have a qualitative evaluation of end-users. Finally, reviews were collected by structured interviews.

The end-users were clustered in four groups, e.g., users 1, 2, 10 and 13 are in the silent environment group, as shown above on Table 1. Next, we will see time differences with interactions and differences with Sphinx speech recognition vs Google speech recognition. This is suitable, to perceive the correlation between variables, and take conclusions.

Google speech recognition system performs very well, but needs Internet connection and in some cases like with users 5, 7 and 8, had a delay. Also Sphinx had a difficult processing orders in environments with very noise and distractions, and there were some users who spoke too fast and too far from the microphone (user 3 and 7). Four users felt overloaded with informations and two of them are female, and did not remember what words to say. They repeated the test, later.

There are more things to take into account, e.g., the outdoor environments can be very wild and often have many restrictions, one of the restrictions are the difficulty of accessing the

[10] To adjust and filter the geographical, biometric and climatic information to provide users, is being developed research in parallel correlations between these three variables, the authors resorted to methods of statistical learning [8, 10].

Table 1 Table with quantitative/qualitative results.

User	Sex	Age	Environment	Quantitative Approach (in seconds)				Qualitative Approach	
				CMU Sphinx		Google ASR			
				word	sentence	word	sentence	Personal evaluation	Information overload
U1	Male	21	silent	0.5	0.5	0.5	0.5	good	no
U2	Female	38		0.5	0.5	0.5	0.5	good	yes
U10	Male	39		0.5	0.5	1.5	1.5	good	no
U13	Male	22		0.5	1.0	0.5	1.0	good	no
U3	Male	14	silent & distractions	1.0	1.5	0.5	0.5	good	yes
U4	Male	48		0.5	1.0	0.5	1.0	good	no
U9	Male	42		0.5	1.0	1.0	1.0	good	yes
U5	Male	22	noise & distractions	1.0	1.0	1.5	1.5	good	no
U6	Male	36		0.5	1.0	0.5	1.0	no answer	no
U11	Female	38		1.0	1.0	1.0	1.5	good	no
U14	Female	37		1.0	1.5	1.0	1.0	bad	yes
U7	Female	37	very noise & distractions	1.0	1.5	1.0	1.0	good	no
U8	Female	19		1.0	1.5	1.0	1.0	good	no
U12	Male	15		1.0	1.0	1.0	1.0	good	no

Internet. In addition to this negative and critical factor, unfortunately the Google recognizer is also very ugly and covers most of the field of view of users (display on prototype). Figure 7 (left image) shows what happens when Google is running.

Figure 7 (right image) shows the end result of an AR prototype with an ideal recognition system. The authors focus on the field of augmented reality and the processing of natural language to create the best way for users to interact with these new technologies.

Also on the Google side and for this chapter discussion can be emphasize the interesting work of Javier Gonzalez-Dominguez et al., at [7], The researchers developed an end-to-end work with multi-language architecture, which was deployed at Google, that allows users to select arbitrary combinations of spoken languages (eight languages simultaneously). They leverage recent advances in language identification and a novel method of real-time language selection to achieve similar recognition accuracy and nearly-identical latency characteristics as a monolingual system.

In the case of the sphinx recognizer the grammar language is only developed in the English language, and in addition it is necessary to follow a sequence of keywords, e.g., “computer” + “features” to the final word or sentence used to associate with the respective method, e.g., for feature communication, “please, i would like to call you, thank you”.

5 Conclusions

This work identified issues and questions about the interaction with an AR application using two speech recognition modules (Sphinx and Google), experimenting different solutions in different conditions, such as with or without noise and distractions. Mobility with the aim of social acceptance, as well as usability, are the most important real benefits for real end-users.

The authors also presented some advantages of the Sphinx system over the Google system. The flexibility in the usage of various kinds of acoustic and language representations, as mentioned by Paul Lamere [12] and the independence of Internet access make Sphinx more advantageous and useful than Google’s speech recognizer. The Discussion section showed results of interactions with fourteen end-users and the advantages of the Sphinx system.

In the empirical tests, the end-users were clustered by environment conditions. Concerning the quantitative perspective, the time taken to respond to speech requests was affected by

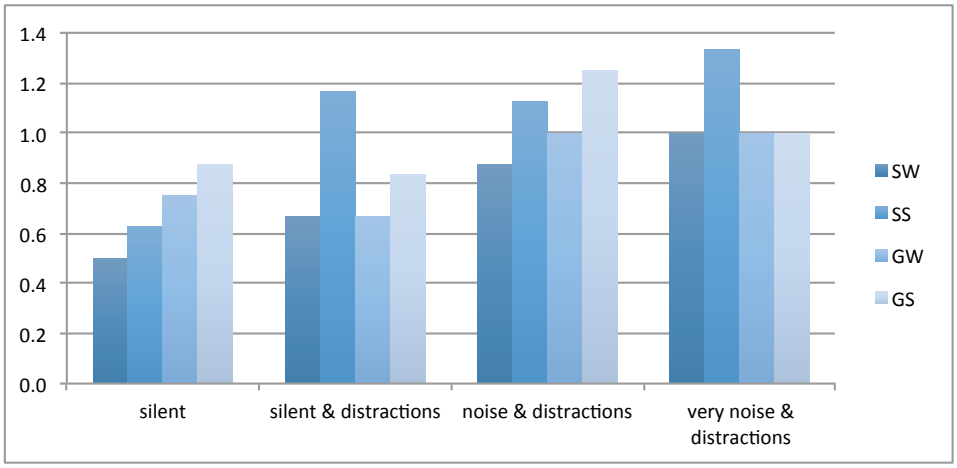

Figure 6 Time taken (seconds) using the two ASR systems under different conditions: Sphinx using a single word (SW), Sphinx using a sentence (SS), Google using a single word (GW), Google using a sentence (GS).

Figure 7 Google Speech recognition (left) and AR mockup with NLP – recognition system for outdoor environments (right).

recognition difficulties by both systems. Sphinx has shown difficulties in four cases, resulting in 1.5 seconds to generate the correct output for requests in the noisy with distractions context. Nevertheless, this system achieved good results in the quiet environment. Google speech recognition system performs well, but needs an Internet connection and in some cases was slower than Sphinx. Feedback from the users suggests that sometimes they spoke too fast or too far from the microphone.

Concerning the qualitative perspective of the evaluation, almost all users considered the interface good. The only exception was provided by a female user in the noisy environment with distractions context that considered that there was an information overload. Additionally, four other users in different contexts also felt that there was an information overload.

The authors propose the simultaneous adoption, deploy, and use of these two automatic speech recognition systems (Sphinx and Google) in the AR application. That is, when using features that do not require an Internet connection, the Sphinx recognizer can be used (e.g., to take a photo, film, and agent AR operational functions), and when access to the Internet is available, Google recognition may be used (e.g., making phone calls and sending messages, because these features require mandatory access to telecommunications infrastructures). In fact, as Google's recognizer was relatively faster in noisier environments than Sphinx recognizer, it can be considered more adequate for communications-based tasks, such as dialing or composing messages.

In societal terms, one important contribution is the adaptation of this kind of technologies, an AR environment with Speech-based interaction, for use in outdoor environments, which can be of great importance, for instance, for tourists (e.g., touristic itineraries with cultural information with landmarks, relevant cultural sites, and historical places).

6 Future Research Directions

The future of interactions with technology will be constantly progressing. New technologies such as NLP and AR in the everyday life of people will be more and more present. The authors and researchers found that the tendency imposed passes through the portability, mobility and simplicity. The use of information systems is transversal of every area of society, and will be increasingly present in the use of this advanced interface. The research done by the authors concerning the paradigms of interaction, noted that it is better to have a fast and practical interaction, to get the best possible benefit of these advanced technologies.

To other grammar possibilities as particular information needed to end-users like if an user requests specific information, e.g., "computer", "informations", "climatics", "tell me, can I play athletics?". Then computer will replay "no", if outlook = sunny and humidity = high. Or, if outlook = normal and windy = false, computer will replay "yes". These are particular cases of given attributes and arbitrary attributes (classification and association rules, respectively [8]). The authors are committed to, and middle the AR development of a capable system to intelligently answer these specific questions solicited by end-users. Also, the authors would like to develop a user login layout before entering in AR and NLP application. This is a requirement to track a particular end-user, but it needs Internet connection as well as to help the GPS tracking precision.

References

1 Ronald T. Azuma. The challenge of making augmented reality work outdoors. In Yuichi Ohta and Hideyuki Tamura, editors, *Mixed Reality: Merging Real and Virtual Worlds*, pages 379–390. Springer-Verlag, 1999.

2 Ronald T. Azuma. The most important challenge facing augmented reality. *Presence*, 25(3):234–238, 2016.

3 Ann Marie Barry. *Visual intelligence: Perception, Image, and Manipulation in Visual Communication*. SUNY Press, 1997.

4 David Bawden and Lyn Robinson. The dark side of information: Overload, anxiety and other paradoxes and pathologies. *Journal of Information Science*, 35(2):180–191, 2009.

5 Dimitris Christodoulakis, editor. *Second International Conference on Natural Language Processing*. Springer, 2000.

6 Alan B. Craig. *Understanding Augmented Reality: Concepts and Applications*. Morgan Kaufmann, 2013.

7 Javier Gonzalez-Dominguez, David Eustis, Ignacio Lopez-Moreno, Andrew W. Senior, Françoise Beaufays, and Pedro J. Moreno. A real-time end-to-end multilingual speech recognition architecture. *Journal of Selected Topics in Signal Processing*, 9(4):749–759, 2015.

8 Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The elements of statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

9 David Huggins-Daines, Mohit Kumar, Arthur Chan, Alan W. Black, Mosur Ravishankar, and Alexander I. Rudnicky. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 185–188, 2006.

10 Edward C. Kaiser, Alex Olwal, David McGee, Hrvoje Benko, Andrea Corradini, Xiaoguang Li, Philip R. Cohen, and Steven Feiner. Mutual disambiguation of 3D multimodal interaction in augmented and virtual reality. In *5th International Conference on Multimodal Interfaces (ICMI)*, pages 12–19, 2003.

11 L. Karttunen, Jean-Pierre Chanod, Gregory Grefenstette, and Anne Schille. Regular expressions for language engineering. *Natural Language Engineering*, 2(4):305–328, 1996.

12 Paul Lamere, Philip Kwok, Evandro Gouvea, Bhiksha Raj, Rita Singh, William Walker, Manfred Warmuth, and Peter Wolf. The CMU Sphinx-4 speech recognition system. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2–5, 2003.

13 Paul Lamere, Philip Kwok, William Walker, Evandro B. Gouvêa, Rita Singh, Bhiksha Raj, and Peter Wolf. Design of the CMU Sphinx-4 decoder. In *8th European Conference on Speech Communication and Technology (EUROSPEECH)*, 2003.

14 Rui Miguel Pascoal and Sérgio Luís Guerreiro. Information overload in augmented reality: The outdoor sports environments. In *Information and Communication Overload in the Digital Age*, pages 271–301. IGI Global, 2017.

15 Heather F. Ross and Tina Harrison. Augmented reality apparel: An appraisal of consumer knowledge, attitude and behavioral intentions. In *49th Hawaii International Conference on System Sciences (HICSS)*, pages 3919–3927, 2016.

16 Ben D. Sawyer, Victor S. Finomore, Andrés A. Calvo, and Peter A. Hancock. Google Glass: A driver distraction cause or cure? *Human Factors*, 56(7):1307–1321, 2014.

17 Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, and Joe Woelfel. Sphinx-4: A flexible open source framework for speech recognition, 2004. Sun Microsystems, Inc.

18 Feng Zhou, Henry Been-Lirn Duh, and Mark Billinghurst. Trends in augmented reality tracking, interaction and display: A review of ten years of ISMAR. In *7th International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 193–202, 2008.