Construction of Medical Reports Using Automatic Speech Recognition System

Thiago Ferreira de Toledo[1]; Huei Diana Lee[1, 2]; Wu Feng Chung[1, 2]; Cláudio Saddy Rodrigues Coy[2]; Newton Spolaôr[1]

[1]Laboratory of Bioinformatics, Western Paraná State University. Presidente Tancredo Neves Avenue, 6731, ZIP code 85867-900, Foz do Iguaçu, Brazil

[2]Service of Coloproctology, Faculty of Medical Sciences, University of Campinas. Tessália Vieira de Camargo Street, 126, ZIP code 13083-887, Campinas, Brazil

{thiagonautf, hueidianalee, wufengchung, claudiocoy, newtonspolaor}@gmail.com

Abstract The general purpose of Automatic Speech Recognition systems is to allow the interaction of humans with computers through speech, converting it to a processed signal and then into text. In general, these systems must be able to face adversity regarding variability of the environment and channel as well as age, style and speed of speech, speech with accents, various speakers, and spontaneous speech. Automatic Speech Recognition can be used to reduce the time needed to prepare medical reports, since manual transcription demands additional time to the health professional. Based on this scenario, this study aims to introduce a method for developing a web prototype system for the preparation of medical reports using an ASR for Brazilian Portuguese. In order to select, evaluate and synthesize speech recognition technologies, a systematic review is being conducted with the purpose of mapping the state-of-the-art and the main advances in the area. Results of the initial evaluation of the so far selected Automatic Speech Recognition systems available for Brazilian Portuguese are also presented.

1. Introduction

Speech is the main and most efficient means of communication among human beings. For this reason, there is a growing interest in allowing computers to understand human speech, from technological curiosity to the desire to automate tasks using machines with more natural interfaces [1].

Specifically, from the point of view of automating tasks in a more naturally manner, Automatic Speech Recognition (ASR) systems can be used to allow the interaction between human beings and computers through speech. In this context, before transforming speech into text, it is submitted to signal processing which uses algorithms to extract information from the signal to be used in specific applications [2].

In general, ASR systems are speaker-dependent, i. e., they are adapted to recognize the voice of an individual, aiming to attain higher precision. However, they have worse flexibility than speaker-independent systems. In turn, independent-speaker systems must be able to face adversity with regard to environmental variability and the channel, age, style and speech speed [3]. Other issues not yet overcome are related to distance, noisy environments, speaking with accents, various speakers and spontaneous speech [4].

Although ASR systems present some limitations, their use can reduce the time needed for drafting texts by typing them in text processors (transcript). An example of text to be transcribed is a medical report, which consists of a written opinion filled by an expert [5]. Manual typing of these documents demands additional time to the health professional, which could be devoted to patient care.

Based on this scenario, the Laboratory of Bioinformatics (LABI) at the Western Paraná State University (Unioeste/Foz do Iguaçu) in partnership with the Department of Coloproctology, Faculty of Medicine (FCM) at the State University of Campinas (UNICAMP), is developing a web system prototype to generate medical reports, with the support of ASR. This prototype would be useful, for example, when the user speaks into a microphone, to narrate a medical report and the spoken content would be converted into a text transcript.

In this context, this work aims to present a method to build this web system prototype to assist health professionals in the preparation of medical reports using ASR technology for the Brazilian Portuguese.

This work is organized as follows: the overall architecture of an ASR system is presented in Section 2. The explanation of the problem and related work are exposed in the Section 3. Section 4 describes material and methods considered to develop the web system prototype, including an initial speech recognition Application Programming Interface (API) evaluation as well as the presentation of the systematic review of related literature. Then, the results obtained with the so far selected ASR systems for the Brazilian Portuguese are reported and discussed (Section 5). Finally, the conclusions are presented in Section 6.

2. General Architecture of an Automatic Speech Recognition

An ASR system works, in general, as follows (Figure 1.1). First, the input speech signal passes through the Front End, which performs signal pre-processing. This task aims to extract information to generate spectral characteristics, which corresponds to information of the wave signal behavior. That is, the signal is represented by numbers to feed a digital system. The spectral characteristics are passed then to a phoneme probability estimator which estimates the smallest sound unit, and is located in the acoustic model. The acoustic and language models are used by the decoder to interpret the meaning of the speech and identify the corresponding words. Then, the decoder converts the words into the text format [1].

In Figure 1.1, the block diagram of the overall architecture of an ASR is presented [1, 4], showing the sequence of necessary processes, from the input of the speech signal to its decoding, and finally to be transformed into a transcription.

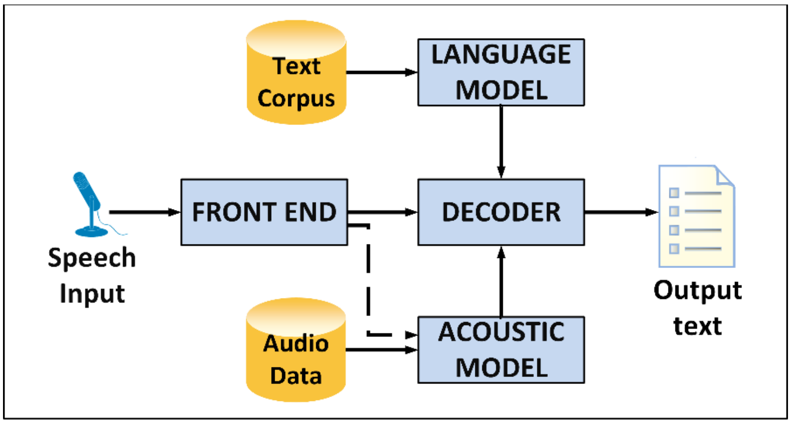

Fig. 1.1 block diagram of the typical architecture of an ASR (adapted from [1, 4]).

Each element of the architecture of an ASR is described in more detail in what follows [1, 4]:

- **Signal Processing and Feature Extraction (Front End):** responsible for improving the speech signal by removing noises and distortions of the communication channel through which the information signals travel between the sender and the receiver. In addition, it converts the signal from time domain to frequency domain to allow to the analysis of the wave signal, and extract the sound characteristics of the signal to be compared with the acoustic models;
- **Acoustic Model (AM):** integrates knowledge about acoustic phonetics, i. e., physical properties of speech sounds, using resources generated by the Front End component to generate a score for each phoneme. This score is used to determine the most likely sequence of phonemes;
- **Language Model (LM):** estimates the probability of occurrence of a given sequence of words, regardless of the acoustic sequence, by assigning a score to that sequence. LM learns the correlation between words in a training corpus (typically text). Punctuation can be estimated more accurately if prior knowledge about the domain or task is known;

- **Decoder:** combines the scores generated by the AM and LM, based on the most evident characteristics of speech sounds and the hypothesis of the sequence of words with higher scores. As a result, we have the sequence of words coming from speech recognition, that is, the text transcribed from what was spoken by the user.

3. Problem Description and Related Work

Medical reports have as standard structure, preamble, questions to be answered, history of diseases, description, discussion, conclusion and response to questions. It should include a description of all signs and symptoms, results of tests performed, treatment adopted, evolution presented and expected from the patient [5].

In order to generate automated transcriptions for medical reports, it is possible to use ASR systems to perform the automatic transcription of texts into digital format from spoken dictation. This idea was already applied in a related work in the radiology area [2, 6], using ASR to generate radiology medical reports.

According to these papers, the use of ASR has helped to create radiological medical reports more efficiently. This is due to the fact that health professionals do not need to type the entire document. As a result, they focus only on document post-processing to complete the report.

However, these papers used commercial systems for English [6] and Finnish [2]. In the case of systems for Brazilian Portuguese, there are still some limitations, such as the scarcity of corpora resources for public use [7, 8, 9, 10] and the current restriction to recognize only isolated words [11, 12, 13]. Regarding the first limitation, it should be emphasized that, although the Portal Domínio Público[1] makes available access to various documents in text and respective audio format, the amount of available resources is not enough to train a state-of-the-art ASR. In fact, a typical language model is trained with more than 230 billion words [14], an amount larger than the one currently stored in Portal Domínio Público.

Speech recognition systems can also be used to add new functionality to existing applications. In this sense, systems such as those developed in LABI/Unioeste in partnership with FCM/UNICAMP can be extended. In what follows, potential applications of ASR to support some of these systems are described:

- **Automation of a process for mapping medical reports to a structured representation [15]:** a collaborative web system was developed for the ontology-based process for mapping medical reports to provide an automated way to transform unstructured textual medical reports into a structured representation such as computational databases. Currently, the system receives reports prepared by specialists by typing text into text processors. Thus, an alternative

[1] http://www.dominiopublico.gov.br/

way of feeding the system consists in the application of automatic transcription techniques by means of an ASR for the preparation of these reports;

- **Real-time remote monitoring and interaction for videocolonoscopy [16, 17]:** an original method, applied to the telemedicine area, was developed to perform the monitoring of colonoscopy exams. In addition, it enables the real-time interaction of medical experts during the exam. This interaction includes chat, audio and video. In this way, ASR can be used so that, as the expert speaks, the system would generate and make available the correspondent textual representation. This legend would serve as a support for the understanding of the speech by other participants, avoiding distortions and even difficulties in recognize patterns of difficult speech, such as accents. In addition, audio subtitling could serve as an adjunct to a textual report for each exam transmission.

4. Materials and Methods

In this section, we describe the materials and methods, including technologies, method and steps required for the development of the web system prototype. Tools used and the evaluated ASRs are also presented in this section, as well as the protocol of the systematic review in progress.

4.1 Steps for the Development of the Web System Prototype

The following technologies and tools are being used to develop the ASR Web System Prototype (WSP) for generating medical reports:

- Technologies: Java[2], JavaServer Faces (JSF) 2[3], PrimeFaces[4], Hibernate[5], JavaScript[6], XHTML (eXtensible Hypertext Markup Language)[7] and CSS (Cascading Style Sheets)[8];
- Tools: Eclipse Mars 2 release (4.5.2)[9], Apache Tomcat 8.0.36[10], database MySQL Communuty Server 5.7[11] and Maven[12].

[2] http://docs.oracle.com/javase/7/docs/technotes/guides/language/

[3] http://www.oracle.com/technetwork/java/javaee/javaserverfaces-139869.html

[4] http://www.primefaces.org/

[5] http://hibernate.org/orm/

[6] http://www.ecma-international.org/ecma-262/5.1/

[7] https://www.w3.org/TR/xhtml1/

[8] http://www.w3.org/Style/CSS/

[9] https://eclipse.org/mars/

The WPS will be accessed via Internet or Intranet, by the following devices: microcomputers, notebooks, tablets and smartphones. The general architecture of the WPS is illustrated in Figure 4.1.

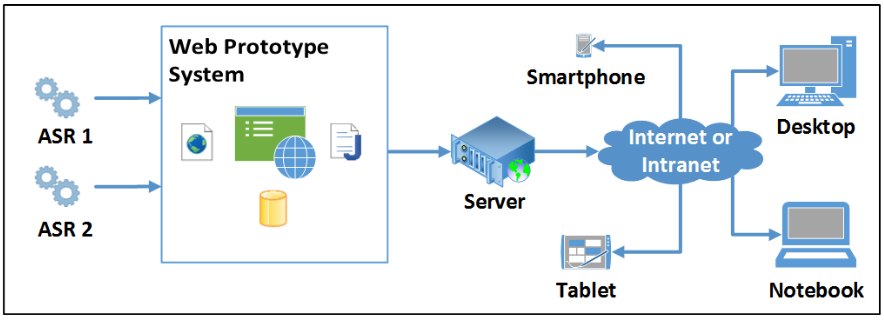

Fig. 4.1 WPS architecture diagram.

The following equipments are being used for the development of the WPS, the execution of the experiments and their hosting:

- Notebook Lenovo G40-80 with Windows 10 Home of 64 bits, processor Intel Core i5-5200U (2.2 GHz, 3 MB de cache), graphics card AMD Radeon R5 M230 2 GB Dedicated, 1 TB hard disk storage and 4 GB DDR3L memory of 1600 Megahertz (MHz);
- Headset microphone Multilaser Ph031 model with a sensitivity of -58 dB, frequency response of 30 Hz-16 Hz.

The prototype building will be based on the Delivery in Stage development method, proposed in the Software Engineering field. This model contemplates the delivery of a new functionality, previously planned and defined, at the end of a cycle [18]. In this approach, requirements analysis is performed to identify the needs of the system, architectural design consists of mapping the system to represent its different parts and, finally, the detailed design specifies the structure and internal behavior of each part of the system, describing them in more detail. Afterwards, based on the requirements defined, the prototype is developed, tested and delivered. This method was chosen due to its ability to deliver useful features before completing the prototype as a whole. Figure 4.2 illustrates the diagram with the steps for the development of a system based on the model Delivery in Stage approach.

[10] http://tomcat.apache.org/

[11] http://dev.mysql.com/downloads/mysql/

[12] https://maven.apache.org/

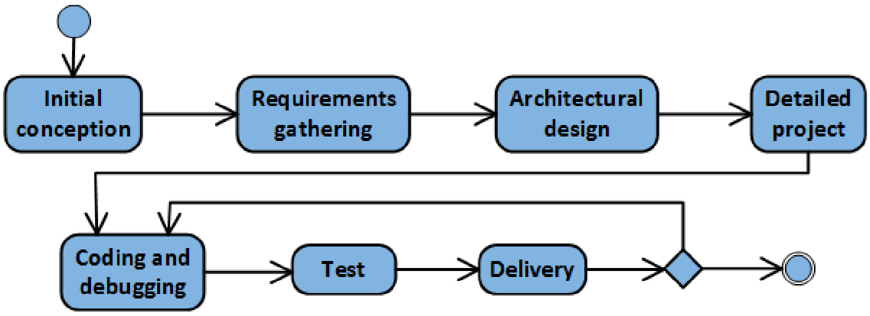

Fig. 4.2 Model diagram Delivery in Stage (adapted from [18]).

The WPS architecture will be based on the Model View Control[13] (MVC), which consists of the detachment of responsibilities regarding the recording of information in the database (Model). Vision is responsible for displaying the data contained in the database and other elements for the user. Intermediation between the Vision and the Model is performed by the Controller. For example, the Controller can manage the update of the state of the Model or send a command to the View to change the display presentation.

An important point to be observed during the development of the prototype is the minimization of the user's cognitive overload, which is related to the excess of information. This problem occurs when the brain's ability to process information is exceeded [19]. In this way, the prototype screens will be constructed with as few visual elements as possible.

4.2 Assessment of Automatic Speech Recognition Systems

The performance evaluation process of an ASR is an important task for verifying the level of reliability of the results generated by the system. One way to evaluate an ASR is to use the Word Error Rate (WER). This measure is derived from the Levenshtein distance, which is a metric for measuring the difference between two sequences of the strings. In WER, the minimum number (or weighted sum) of insertions, deletions and substitutions needed to transform the string into another are verified [20].

In this work, the evaluation of the ASRs is performed collecting audio recordings of volunteers and submitting then to each of the ASR system APIs. The text generated by the system is then compared with the reference text.

A 617 words reference text was used as a base[14] , from which the majority of the content is related to the medical area. Altogether, ten volunteers (five men and

[13] http://heim.ifi.uio.no/~trygver/themes/mvc/mvc-index.html

[14] http://drauziovarella.com.br/cancer/colonoscopia/

five women) with ages ranging from 19 to 43 collaborated to read this text at normal speed. At the time of reading the text, the Voice Recorder software version 10.1611.3051.0 present in Windows 10, was used for audio recordings.

To calculate the WER, the Sclite software, a module of the NIST Scoring Toolkit SCTK[15], was employed. The Sclite compares the text output generated by the ASR with the reference text. Scores and differences between upper and lower case letters are not considered.

The ASR considered in this work for evaluation in the Brazilian Portuguese were: Audimus[16], Bing Speech API[17] (developed by Microsoft), Coruja[18], IBM Speech to Text[19], two versions of Voxsigma Speech to Text[20] (a stable version and a beta version, which at the time of this evaluation was carried out, was under development) and Web Speech API[21] (developed by Google).

The statistical analysis of the results were made using the Action Stat software[22].

4.3 Systematic Review Protocol

The Systematic Review (SR) is a means to identify, evaluate and interpret all the relevant research available for a particular research question [21]. As part of the method for the development of this work, a SR is being performed to synthesize the relevant evidences in the context of ASR, as well as to find the state-of-the-art tools that can be coupled to the WPS.

Figure 4.3 illustrates the SR protocol, organizing it into three steps: planning, executing and describing the results.

[15] http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sctk.htm

[16] http://www.voiceinteraction.com.br/?page_id=423

[17] https://www.projectoxford.ai/speech

[18] http://www.laps.ufpa.br/falabrasil/downloads.php

[19] https://speech-to-text-demo.mybluemix.net/

[20] http://www.vocapia.com/voxsigma-speech-to-text.html

[21] https://developers.google.com/web/updates/2013/01/Voice-Driven-Web-Apps-Introduction-to-the-Web-Speech-API

[22] http://www.portalaction.com.br/action-stat-pro

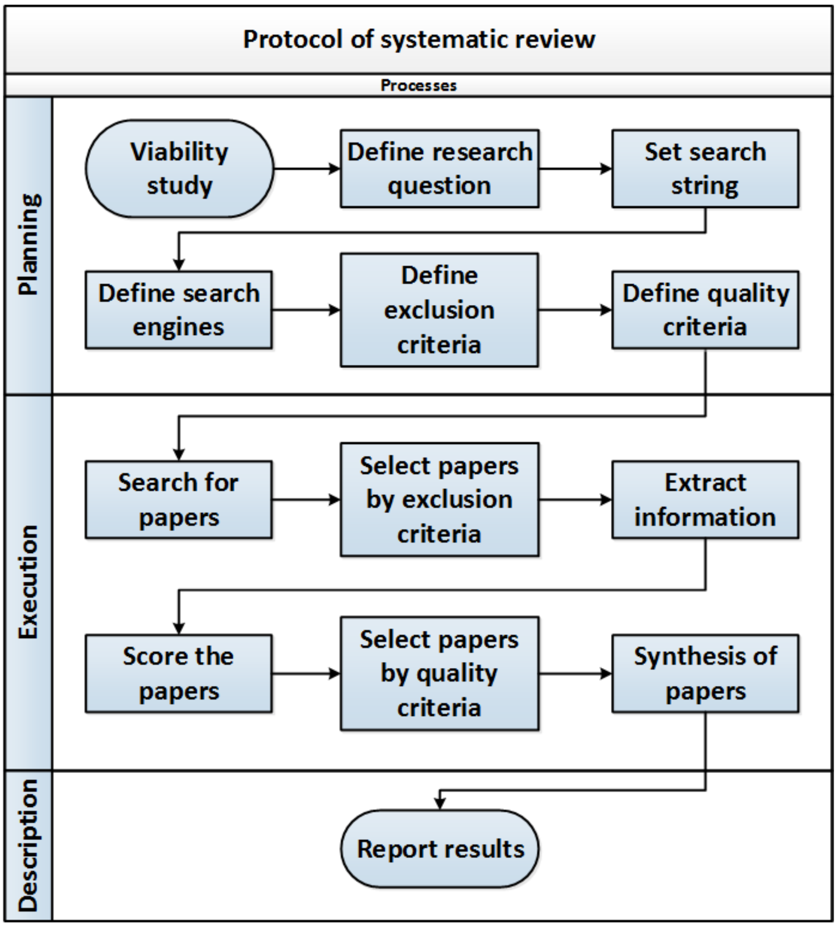

Fig. 4.3 SR protocol flowchart.

In the planning stage the following are defined: research question, search strategy, selection and quality criteria. In the next stage, execution, search is carried out and a first selection is conducted, for example, removing papers that accomplish one or more selection (exclusion) criteria. Then, information is extracted and papers are scored. Afterwards, a second selection phase includes works that meet the quality criteria and, finally, the synthesis is carried out. In the description step, the results are reported.

The Systematic Review conducted in this work aims to answer the following research question: which computer systems or automatic speech recognition software have been used in recent work?

Regarding the search strategy, the following search string was defined: *(("automatic speech recognition") AND ("open source" OR application OR system OR software OR program OR tool))*.

This string was used to search the databases of Scopus, CAPES, IEEE Xplorer, Wiley, Web of Science, CiteSeerX, ACM and PubMed. In order to obtain the most recent work in the area, this SR focused on papers published after 2010.

The Systematic Review was divided into four phases:

1. Selection of work by exclusion criteria, the following criteria are applied:

- Duplicated title, i.e., publication with the same title used by other publications found by SR;
- Work that deals exclusively with language, acoustic or lexical models, corpus and extraction of characteristics;
- Work that deals with speech recognition, visual speech recognition, Text-To-Speech;
- Patents;
- Work that proposes methods or techniques for hardware upgrade for the speech recognition;
- Work that do not meet the research question;
- Work whose institutional access has not been released or publications that are not assigned to a full text accessible from the institutional access;
- Work written in languages different from Portuguese, Spanish or English;
- Publications considering an ASR recognition language different from Italic language (Portuguese, Spanish, French, Italian, Catalan, Romanian and Latin) and Germanic language (Icelandic, Norwegian, Swedish, Danish, Scottish, English, Frisian, Dutch, Low German and German).

2. Extraction of information from selected papers, according to the following fields: country in which the research was developed and ASR language, system technology (ASR system name, classifier, feature extraction and corpus), accuracy rate achieved and continuous recognition (if the ASR used in the work is for recognition of keywords, isolated words or digits, the field is marked with “no”);
3. Evaluation of the quality level of the work, based on the following criteria: (1) the recognition language of the ASR (2) presents discussion of results? (3) presents accuracy rate? (4) is continuous speech recognition performed? (5) allows to be trained with other corpora? (6) proposes technique to improve ASR technology? The scoring of each publication is applied as described in what follows:

- Three points: the publication uses an ASR for the Portuguese language;
- Two points: the publication uses an ASR of Spanish or English language;
- One point for each satisfied requirement: the publication uses other languages, papers that presented discussion on the results, presents accuracy rate, performed continuous speech recognition, allows to be trained with other corpora and proposes technique to improve ASR technology.

4. Selection of work based on quality criteria. In this phase, we kept publications that consider unpublished ASR. And some publications that use the same ASR – if an ASR was used in two or more publication, only the corresponding papers that obtain the highest scores are selected.

5. Results and Discussion

This section aims to discuss the preliminary results achieved so far. In addition to the methodology used for the construction of WPS, as mentioned. in this work, we also performed a preliminary experiment in order to evaluate the selected ASR systems. Results of the systematic review are also presented and discussed in this section.

5.1 Automatic Speech Recognition Systems Evaluated in the Preliminary Experiment

In order to perform a previous analysis of the available ASRs, seven systems for speech recognition for the Brazilian Portuguese were evaluated. The mean error measured in WER (the lower the error, the better the accuracy is) can be verified in Figure 5.1. Table 5.1 shows the WER mean error values and the respective standard deviation (bold lines of the table) are detailed for each volunteer in the evaluated ASRs.

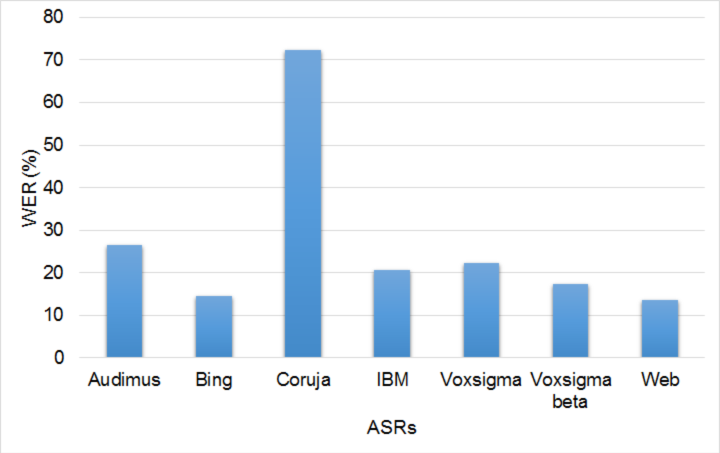

Fig. 5.1 ASRs evaluated.

Table 5.1: average error WER (%) with the results of the volunteers (SD: standard deviation) for the assessed ASR.

Volunteer	Audimus	Bing	Coruja	IBM	Voxsigma	Voxsigma beta	Web
Male 1	27.06	11.18	64.67	21.39	17.83	15.88	11.34
Male 2	27.39	16.37	83.79	26.58	23.01	19.12	15.40
Male 3	30.31	14.75	66.45	21.07	24.80	18.48	14.10
Male 4	26.09	12.80	66.29	17.99	21.56	17.50	7.94
Male 5	15.56	6.81	63.53	15.40	14.42	13.78	8.10
Female 1	28.20	20.91	82.17	19.93	22.37	18.64	17.99
Female 2	33.55	15.23	80.71	22.37	28.36	18.96	15.88
Female 3	31.12	12.48	85.25	19.94	27.07	18.48	14.59
Female 4	27.39	19.45	76.82	27.45	29.66	22,69	15.23
Female 5	19.45	15.56	53.16	14.42	13.29	10.21	14.75
Average	**26.61**	**14.55**	**72.28**	**20.65**	**22.24**	**17.37**	**13.53**
SD	**5.37**	**4.05**	**10.85**	**4.21**	**5.62**	**3.40**	**3.34**

In order to statistically analyze the accuracy level of the ASRs, the Anderson-Darling normality test for small samples is applied [24]. It was verified that the data set does not follow a normal distribution.

As the samples were extracted from the same population, present ordinal, continuous and unpaired data with no gaussian distribution, the Friedman nonparametric test was performed using the Simes-Hochberg multiple comparison method [25, 26], as a post-test, at a significance level of 5%. It was not found a significant difference between the samples (P-value = 0,0000000058). Table 5.2 shows the multiple comparison between the evaluated ASRs. In this table, the lines in bold represent the cases with statiscally significant difference whereas the italic line indicates that it was not found significant difference between the Bing Speech API and Web Speech API.

The WER of ASRs according to volunteer's gender is shown in Figure 5.2. The IBM ASR tool demonstrated lower variation between genders. One can note that all ASRs presented lower average error for male volunteers. However, the ASR Web still achieved the best result.

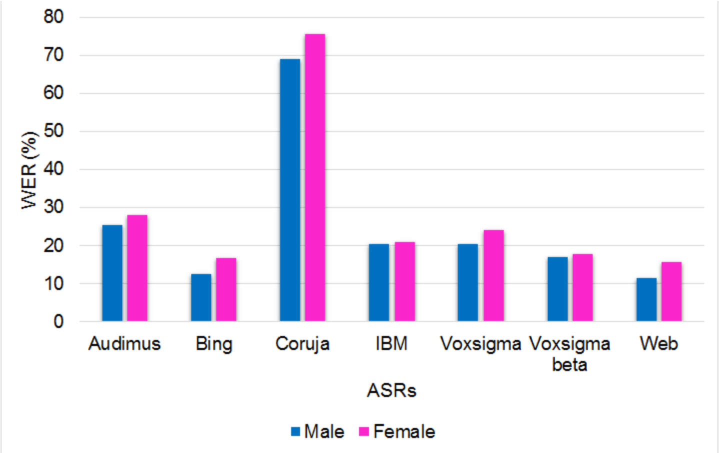

Fig. 5.2 Percentual of the WER achieved by the evaluated ASR, averaged across the results obtained from all volunteers in a specific gender.

The speech recognition technology developed by Google consists of a modified Long Short-Term Memory (LSTM) based on the Recurrent Neural Network (RNN). The LSTM architecture contains special units called memory blocks in the recurring hidden layer. Memory blocks contain self-connecting memory cells that recall the temporal state of the network, as well as multiplicative units to control the flow of information [22].

The LSTM-RNN model presented better performances in the task of speech recognition with a great vocabulary for the English language [22]. An architecture based on RNN and LSTM is also used by Microsoft in conversation systems for speech understanding. This approach was tested at Cortana, a Microsoft's virtual assistant for user interaction with the operating system [23].

The Web Speech API is a Javascript specification, published by Speech API Communit Group[23], and was developed by Google. Its use is restricted for the Google Chrome browser desktop version and in smartphones and tablets with the Android operating system. To allow the WPS application in other browsers and other devices that do not have the Android operating system, Microsoft's ASR Bing Speech API will also be coupled to the prototype.

With these results, it is possible to observe that the proprietary technology systems presented performance superior to Coruja (single open source system). One possible explanation for this superiority consists in the use, by the proprietary systems, of a large set of training corpus, both in quantity and in speech variations, as well as the use of superior speech recognition technologies.

[23] https://www.w3.org/community/speech-api

Table 5.2: Multiple comparison between ASRs.

Factors	P-value	P-value adjusted
Audimus - Coruja	**0.0001**	**0.0022**
Audimus - Web	0.2142	0.7562
Audimus - IBM	0.0977	0.7562
Audimus - Bing	0.1784	0.7562
Audimus - Voxsigma	**0.0013**	**0.0213**
Audimus - Voxsigma beta	**<0.0001**	**0.0004**
Coruja - Web	**<0.0001**	**<0.0001**
Coruja - IBM	0.0297	0.2973
Coruja - Bing	0.0130	0.1428
Coruja - Voxsigma	0.5346	0.7562
Coruja - Voxsigma beta	0.6788	0.7562
Web - IBM	0.0038	0.0525
Web – Bing	*0.0097*	*0.1159*
Web – Voxsigma	**<0.0001**	**0.0002**
Web - Voxsigma beta	**<0.0001**	**<0.0001**
IBM – Bing	0.7562	0.7562
IBM – Voxsigma	0.1205	0.7562
IBM - Voxsigma beta	0.0097	0.1159
Bing- Voxsigma	0.0624	0.5619
Bing- Voxsigma beta	0.0038	0.0525
Voxsigma - Voxsigma beta	0.3006	0.7562

These ASRs also exhibited the best mean WER values. Therefore, based on this analysis, the Web Speech API and Bing Speech API ASRs were selected to be coupled to the WPS.

5.2 Systematic Review

The systematic review started with 6.113 papers titles that were returned during the first search on the selected databases. By applying the exclusion criterion regarding duplicate titles, 4.421 papers were selected. After applying the other exclusion criteria, 404 papers were selected. In the second phase, information was extracted from these 404 papers.

In the third phase the quality was evaluated, assigning scores to the papers. Then, in the fourth phase, 44 publications that presented unpublished ASR and 76 publications with maximum scores for each defined ASR were selected, yielding a total of 120 papers. Figure 5.3 presents the flowchart with the four phases specific to this work: selection by exclusion criteria, extraction of information, quality evaluation and selection by quality criteria.

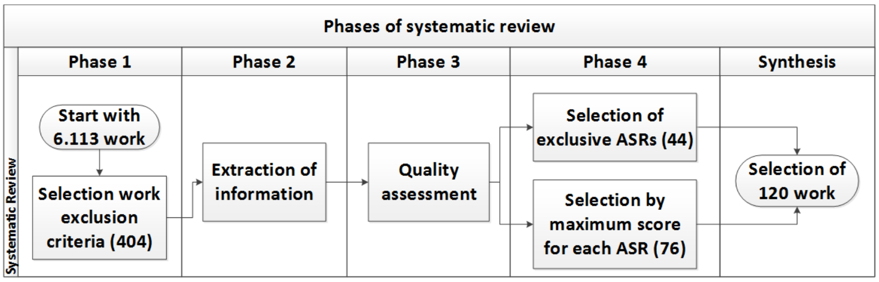
Phases of systematic review
Phase 1
Phase 2
Phase 3
Phase 4
Synthesis
Systematic Review
Start with 6.113 work
Selection work exclusion criteria (404)
Extraction of information
Quality assessment
Selection of exclusive ASRs (44)
Selection by maximum score for each ASR (76)
Selection of 120 work

Fig. 5.3 Flowchart of the SR phases.

It should be emphasized that one of the criteria for selecting papers maintains only articles written in English, Spanish and Portuguese with ASRs for the Italic and Germanic languages, and considers that the selection method prioritizes works whose language of the ASR is Portuguese or English. This influences the results summarized in Figure 5.4 as papers written in other languages or ASRs in another language were eventually not selected.

Analyzing Figure 5.4, It can be seen that the predominant language support of ASR is the English (83,01%). The 14 multilingual ASRs also feature training for the English language. The second predominant recognition language in ASRs is Portuguese (from Brazil and Portugal), corresponding to 5,66%, followed by the Dutch and French languages (1,89%) and the Romanian and German languages (0,94%).

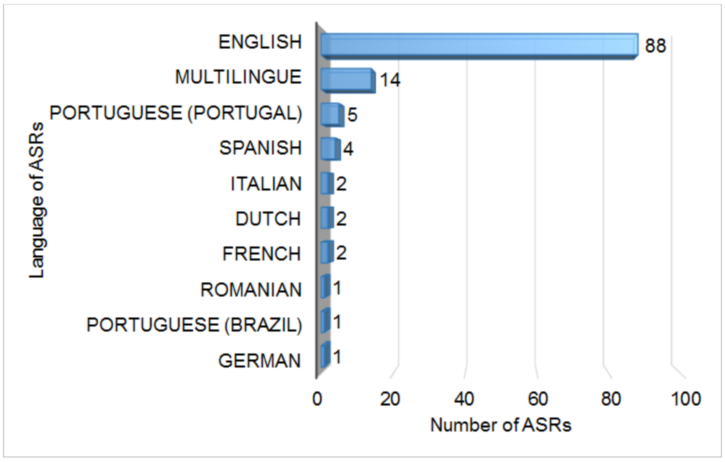

Fig. 5.4 Language of training for ASRs.

The ASRs selected in SR are mentioned in Figure 5.5. Note that systems with occurrences in two or more papers are shown in a specific column for the corresponding ASR. On the other hand, the ASRs that were used in only one work fit into a unique column: ASR that appeared in a single work. Altogether, 116 systems are considered in the figure, because works that presented more than one ASR (4) were not represented.

HTK[24] and Kaldi[25] are open-source ASR tools and are indicated respectively in 18 and 17 of the selected papers, respectivaly. The commercial systems BBN Byblos[26] and Audimus[27], appears both in 8 papers. Other ASRs used were: CMU Sphinx and the PocketSphinx[28], MATLAB[29], SRI DynaSpeak[30], Julius[31], SHOUT[32], Janus[33], IBM Atila [27], Dragon[34] and AMI[35].

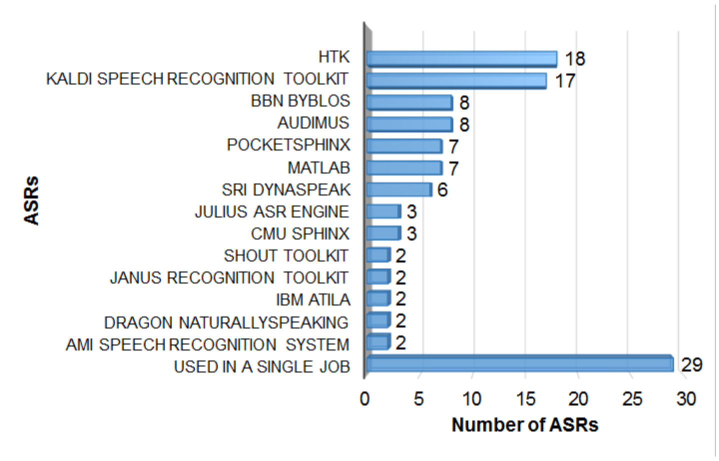

Fig. 5.5 ASRs selected in the systematic review.

In Figure 5.6, the amount of published works per year is illustrated. Although the 2015 year is associated with the lowest quantity of works, it should be empha-

[24] http://htk.eng.cam.ac.uk/

[25] http://kaldi-asr.org/

[26] http://www.raytheon.com/ourcompany/bbn/

[27] http://www.voiceinteraction.tv/

[28] http://cmusphinx.sourceforge.net/

[29] https://www.mathworks.com/products/matlab/

[30] http://www.speechatsri.com/index.shtml

[31] http://julius.osdn.jp/en_index.php

[32] http://shout-toolkit.sourceforge.net/index.html

[33] http://isl.anthropomatik.kit.edu/cmu-kit/english/1406.php

[34] http://www.nuance.com/dragon/index.htm

[35] http://speech.fit.vutbr.cz/projects/ami-project

sized that the SR execution was concluded in September of the same year. Thus, papers not indexed or published by that time were not considered.

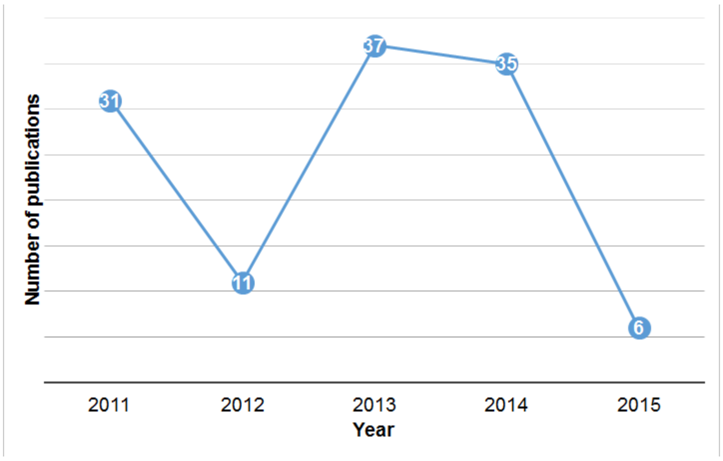

Fig. 5.6 Number of publications per year.

6. Conclusion

This work described a method that is being applied to develop a prototype web system. This system can be used to assist health professionals in generating medical reports through an ASR. It was also reported the systematic review protocol considered to contextualize relevant literature evidences and to align the present work with the state-of-the-art of Automatic Speech Recognition technologies.

Some of the main findings were achieved by the preliminary experiment, which highlighted the Bing Speech API and Web Speech API. This evaluation consisted of submitting audio files from volunteers to each of the ASRs. The result was obtained by comparing the generated text with the reference text.

After including these APIs in the WPS under development, we will conduct a broader experiment to evaluate the system´s performance in terms of speech recognition accuracy. A qualitative evaluation will also be performed in order to assess the usability and quality of the interface and functions of the web prototype system.

References

[1] Padmanabhan J and Premkumar MJJ. Machine Learning in Automatic Speech Recognition: A Survey. IETE Technical
Review. 2015; 32(4):240–251.

[2] Koivikko MP, Kauppinem T, and Ahovuo J. Improvement of Report Workflow and Productivity Using Speech Recognition - A Follow-up Study. Journal of Digital Imaging. 2008; 21(4):378–382.

[3] Vimala C and Radha V. Isolated Speech Recognition System for Tamil Language Using Statistical Pattern Matching and Machine Learning Techniques. Journal of Engineering Science and Technology. 2015; 10(5):617–632.

[4] Yu D and Deng L. Automatic Speech Recognition: A Deep Learning Approach. 1st ed. London: Springer; 2015.

[5] Conselho Regional de Medicina do Paraná. Parecer nº 1936/2008 CRM-PR - Diferença entre Atestado e Laudo Médico. Portal Médico [Internet]. 2008 [Access in 2016 nov 14]. Available in: http://www.portalmedico.org.br/pareceres/CRMPR/pareceres/2008/1936 2008.htm.

[6] Prevedello LM, Ledbetter S, Farkas C, and Khorasani R. Implementation of Speech Recognition in a Community Based Radiology Practice: Effect on Report Turnaround Times. Journal of the American College of Radiology. 2014; 11(4):402–406.

[7] Silva E, Baptista L, Fernandes H, and Klautau A. Desenvolvimento de um Sistema de Reconhecimento Automático de Voz Contínua com Grande Vocabulário para o Português Brasileiro, Proceedings of the XXV Congresso da Sociedade Brasileira de Computação; 2005 jul 22-29, São Leopoldo. p. 2258-2267, 2005.

[8] Pessoa LAS, Violaro F, and Barbosa PA. Modelos da Língua Baseados em Classes de Palavras para Sistema de Reconhecimento de Fala Contínua. Revista da Sociedade Brasileira de Telecomunicações. 1999; 14(2):757–784.

[9] Fagundes R and Sanches I. Uma Nova Abordagem Fonético-Fonológica em Sistemas de Reconhecimento de Fala Espontânea. 2003; 18(3):225–239.

[10] Batista Santos P. Avanços em Reconhecimento de Fala para Português Brasileiro e Aplicações: Ditado no LibreOffice e Unidade de Resposta Audível com Asterisk [Master's dissertation]. Belém: Federal University of Pará; 2013.

[11] Santos S and Alcaim A. Um Sistema de Reconhecimento de Voz Contínua Dependente da Tarefa em Língua Portuguesa. Journal of Communication and Information Systems. 2002; 17(2):135–147.

[12] Gomez EMT. Reconhecimento de Fala para Navegação em Aplicativos Moveis para Português Brasileiro [Master's dissertation]. Sao Paulo: University of São Paulo; 2011.

[13] Oliveira ALC, Silva ES, Macedo HT, and Matos LN. Sistema de Atendimento com Interação de Fala para o Português do Brasil, EATIS 2012: Proceedings of the 6th Euro American Conference on Telematics and Information Systems; 2012 may 23-25, New York; 2012.

[14] Gold B, Morgan N, and Ellis D. Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics. 1st ed. New Jersey: Amy Neustein; 2010.

[15] Oliva JT. Automatização do Processo de Mapeamento de Laudos Médicos para uma Representação Estruturada [Master's dissertation]. Foz do Iguacu: Western Paraná State University; 2014.

[16] Machado RB; Lee HD; Ayrizono MLS; Leal RF; Coy CSR; Fagundes JJ. Prototype of a Computer System for Managing Data and Video Colonoscopy Exams. Journal of Coloproctology. 2012; 32(1):50–59.

[17] Takaki WSR. Proposta de um Sistema Embarcado para Transmissão de Vídeos em Tempo Real com Aplicação em Telemedicina [Master's dissertation]. Foz do Iguaçu: Western Paraná State University; 2015.

[18] Waslawick RS. Engenharia de Software: Conceitos e Práticas. 1st ed. Rio de Janeiro: Elsevier; 2013.

[19] Zahabi M, Kaber DB, and Swangnetr M. Usability and Safety in Electronic Medical Records Interface Design: a Review of Recent Literature and Guideline Formulation. Human Factors and Ergonomics Society. 2015; 57(5):805–384.

[20] Davis KH., Biddulph R, and Balashek S. Automatic Recognition of Spoken Digits. The Journal of the Acoustical Society of America. 1952; 24:637–642.

[21] Kitchenham BA. Guidelines for Performing Systematic Literature Reviews in Software Engineering. Universidade de Keele, Keele; 2007.

[22] Sak H, Senior A, and Beaufays F. Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition [Internet]. 2014, [Access in 2016 nov 13]. Available in: https://arxiv.org/abs/1402.1128.

[23] Hakkani-Tür D, Tur G, Celikyilmaz A, Chen YNV, Gao J, Deng L, and Wang YY. Multi-Domain Joint Semantic Frame Parsing Using Bi-directional RNN-LSTM, INTERSPEECH 2016: Proceedings of the 17th Annual Meeting of the International Speech Communication Association; 2016 sep 8-12, Sao Francisco, 2016.

[24] Hou A, Parker LC, and Wilman DJ Harris WE. Statistical Tools for Classifying Galaxy Group Dynamics. The American Astronomical Society. 2009; 702(2):1199–1210.

[25] Simes RJ. An Improved Bonferroni Procedure for Multiple Tests of Significance. Biometrika. 1986; 73(3):751–754.

[26] Hochberg Y. A Sharper Bonferroni Procedure for Multiple Significance Testing. Biometrika. 1988; 75(4):800–802.

[27] Soltau H, Saon G, and Kingsbury B J. The IBM Attila Speech Recognition Toolkit. 2010 IEEE Spoken Language Technology. 2010:97–102.