Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 196 (2022) 235–244

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2021

A web-based Voice Interaction framework proposal for enhancing Information Systems user experience

Tiago F. Pereira[a,*], Arthur Matta[a], Carlos M. Mayea[a], Frederico Pereira[a], Nelson Monroy[b], João Jorge[b], Tiago Rosa[b], Carlos E. Salgado[a], Ana Lima[a], Ricardo J. Machado[c], Luís Magalhães[c], Telmo Adão[a], Miguel Ángel Guevara López[a], Dibet Garcia Gonzalez[a]

[a]*Center for Computer Graphics, Campus de Azurém, Edifício 14, Guimarães, 4800-058, Braga, Portugal*
[b]*PHC, Av. Prof. Dr. Cavaco Silva 7A, Porto Salvo, Lisboa, 2740-120, Portugal*
[c]*Minho University, Campus de Azurém, Edifício 14, Guimarães, 4800-058, Braga, Portugal*

Abstract

Nowadays, numerous organisations of different dimensions and business sectors operate in highly challenging and dynamic environments, wherein the supporting information systems (IS) are becoming increasingly complex. In this context, assistive tools capable of tackling such complexity have the potential to aid users improving their performance and effectiveness, as well as to streamline businesses' processes and promote entrepreneurial-level competitivity. Following this line of research, a web-based speech-to-term recognition approach is presented as a solution to endow IS with advanced capabilities for providing an easier (more natural) and straightforward interaction with baseline functionalities, by combining relevant techniques such as Voice Activity Detection (VAD), Automatic Speech Recognition (ASR), Natural Language Processing (NLP), and an Ontological Database (OD), mapping the IS' functionalities and characteristics, is proposed. The developed interoperable system allows the conversion of speech to text - deriving into IS instructions - that is, in turn, submitted to on an ontological database wherein a term-based query is performed to elicit a set of available commands to be executed in the web context. These commands, fully mapped in the ontological database, are divided into three categories: a) navigation by menus/links, b) buttons interaction (e.g., submit forms) and c) completion of form fields. The proposed framework was experimentally tested in close to real conditions, resorting to an Enterprise Resource Planning (ERP) tool supplied by ERP Company.

Peer-review under responsibility of the scientific committee of the CENTERIS –International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2021

* Corresponding author. Tel.: +351 914 267 120.
E-mail address: fernando.pereira@ccg.pt

1877-0509
Peer-review under responsibility of the scientific committee of the CENTERIS –International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2021
10.1016/j.procs.2021.12.010

Keywords: Automatic Speech Recognition; Deep Learning; Voice Activity Detection; Ontologies; Graph Database; Ontological Component; Semantic Interoperability.

1. Introduction

For the vast majority of todays' computer workstations, the main interfaces for human information input are the keyboard and mouse peripherals. The introduction of additional information input interfaces (such as audio) can prospectively increase the productivity of users interacting with an information system (IS). Technologies such as voice activity detection (VAD), speech-to-text (S2T) and an Ontological Component (OC), mapping the structure and cataloguing data of interest, have the potential to unlock advanced user-system interaction through human-intrinsic communication means - also known as Natural User Interfaces (NUI) - in particular, voice.

Voice Activity Detection is a term used for signal processing methods that allow the detection of speech contained in an audio signal. In speech processing, the discrimination between speech and non-speech is, still to this day, a relevant issue affecting web applications for real-time speech recognition. Speech processing algorithms are often significantly demanding of computing resources and, as speech is naturally discontinuous; the inclusion of VAD methods in these algorithms is a design strategy to alleviate unnecessary processing [1]. In the proposed framework, the VAD implementation integrates a Front-End module, where the audio signal acquisition is performed from a microphone device. The signal is analysed for speech detection, segmented, and finally, speech only signals streamed to a Back-End processor. The implemented VAD must be capable of detecting speech for typical background noise, a feature that is commonly referred to as noise-robust.

Nowadays, speech recognition (SR) implemented into voice assistants represents a supplementary input source for devices like mobile phones, tablets and virtual assistants, allowing user interaction with these devices. Benefiting from this widespread demand, SR technology has achieved a level of maturity justifying its implementation in web systems as an additional input source of information.

Relative to the Ontological Component, Ontologies can be considered a data model that represents a set of concepts within a domain and the relationships between them. Fundamentally, we can characterise it as a technique for organising information, especially with regard to the formal representation of knowledge. These are generally created by specialists and, as their structure is based on the description of concepts and the semantic relationships between them, they make it possible to generate a formal and explicit specification of a shared conceptualisation. In this context, we perform an analysis and demonstration of the adopted approach. To this end, we define an ontological architecture where the entire ERP is mapped, specifically, all of its menus, features and expressions involved. This ontological systematisation allows to catalogue the terminologies inherent to an ERP as well as the attribution and definition of synonyms for expressions and functionalities associated.

The rapid progress in efficiency that has been seen in the three afore-mentioned technologies and their combination, create the opportunity for an increasingly natural and faster user interaction with potentially complex IS, therefore permitting an improvement in productivity.

The remaining of this document is composed of a state-of-the-art (section two) wherein terms and concepts used in the acquisition and processing of audio, speech-to-text conversion models and, also, on the Ontology and interoperability component are detailed. Next, in section three, materials and methods used to implement the proposed solution are discussed. In the fourth section, an analysis of the obtained results is carried out and lastly, a conclusions and future work section is presented.

2. On Voice Activity Detection, Automatic Speech Recognition and Ontological Approaches.

Nowadays, technologies like SR and the hardware that supports it have reached a level of maturity that reliably constitute an approach to consider on the improvement of user interfaces' web-based implementations. The most

important operations a user performs are: (i) navigation to a web site, even by links that are not accessible in the current page, (ii) input form fulfilling (e.g., text, number, selection list) and (iii) actions execution (e.g., submit or cancel forms).

2.1. Voice Activity Detection

VAD methods vary in their processing techniques, but its purpose is generally identical. The VAD method should be able to discriminate speech data from non-speech data in a given audio signal. In their great majority, the speech segments are grouped in data chunks sorted chronologically for further processing, thus removing useless information (noise) from the data. Overall, VAD methods may be classified in two categories: based on energy thresholding, or in machine learning methods. Energy thresholding methods are sustained on the fact that speech adds energy to a signal, thus enabling discrimination between higher and lower energy (no-speech) sections of the signal. These implementations are generally simple, and thus widely adopted in systems with restricted resources [2]. Machine learning based VAD methods are of a seemingly simple design, based on one or more speech features selection, a learning technique, and training over large enough data suit-able for intended use cases. These methods often have great accuracy, but proposed implementations are modern and some techniques have yet to be refined, as they are frequently complex and resource intensive [3].

2.2. Automatic Speech Recognition

Automatic Speech Recognition (ASR) technology allows an electronic device to identify vocalized words and has been a research field since the 1950s [4]. ASR can be seen as a mathematical model that converts speech sounds into text, and it should not be confused with Voice Recognition, which aims to identify the person speaking instead of the speech [5]. Among the approaches for speech recognition, there are Hidden Markov Models (HMM) and Deep Neural Networks (DNN), which have been widely explored [6,7,8,11,12,13,14]. Essentially, most speech recognition systems have three main components [6]:

- Feature extraction: the audio is converted into a group of feature vectors that feed the following stage of the process; this part is sensitive to noise, pronunciation, speaker age or gender, etc.
- Acoustic model: is a context-independent model, trained to recognize phonemes from the vector of characteristics, used to build the words.
- Language model: it manages the grammar rules and defines the most probable order of words in the sentence, commonly represented as n-gram models that contain statistics of word sequences.

Many current speech recognition systems are mainly used as virtual assistants on mobile phones e.g. Siri, Alexa or Cortana. These technologies are becoming increasingly refined, arousing the interest of many scientific and professional communities in turning them into specialized ASR, for example, in the context of IS operability. Many frameworks, such as Microsoft Bing Speech Application Program Interface (API), Google Speech API or IBM Watson Speech-to-Text offer immediate development solutions to integrate in custom projects, although, under a commercial licensing. On the other hand, there are also open-source toolkits that aim developers to create speech-to-text systems in different programming languages and platforms like Kaldi [7], Wav2letter++ [8], Mozilla DeepSpeech [12], CMU Sphinx [13], Vosk [14], among others. In particular, this work focuses on adapting opensource speech-to-text engines, more specifically Mozilla DeepSpeech, CMU Sphinx and Vosk. We also use Google Speech API, but only for comparison purposes, due to its excellent results when evaluated in [9], [10] and [11].

2.3. Ontological Approaches

The existence of interoperability in companies is essential to efficiently face the modern-day challenges of fierce competition, find new business opportunities and provide a better service to productive needs [15]. Document exchange, being semantically consistent, is among the ways to achieve interoperability. However, the heterogeneity of structures within an organisation presents challenges for establishing interoperability [16]. The structure of semantic interoperability based on ontology has proven to be an effective solution for business interoperability [17].

The main issues ontologies seek to address are related to a lack of interoperability that many present-day companies still deal with [18]. Organisations comprise all information flows between people, processes and machines, including written and spoken communication, constituting an organised system for the collection, organisation, storage and communication of information [19].

Due to all this information flows, it is essential to have a standardised procedure to harmonise the way we treat a domain within an organisation, and this is where some obstacles to the process start to emerge [20]. In order for all of this to work properly, it is necessary to have a uniform and coherent business ecosystem in terms of data and its applications, but in some cases, we are confronted with unstructured data, with applications that operate in isolation and, additionally, with excessive use of paper or spreadsheets documents [21].

So, if we want a company with an interoperable system, we initially have to provide it with the necessary means and processes, carry out a survey of the existing computer applications and the terminologies used so that we can proceed to their standardisation through a catalog of terminologies. It is precisely here that ontologies act as a solution to that problem [22]. In this way, at the level of business/organisational processes, interoperability components, namely ontologies, have been gaining more and more emphasis and greater preponderance because they can be used to explain both the semantics of activities in general as well as the semantics of an organisation and the specific activity, for example, the resources used or the context of the application described by a domain ontology [23].

3. Proposal of a framework for IS-aware speech recognition

According to the different solutions envisioned for the diverse parts of the proposed approach, a framework connecting the three main parts has been discussed among the team and validated with the customers involved. Initially a trigger would be set at the client ERP, detecting a valid voice command, which would then be interpreted and processed for transformation into a text command. Next, the resulting text is semantically validated and mapped through the ontological database in order to, in case of success, return a valid command for the ERP within its present context. So, the proposed framework (Fig. 1) is then divided into three modules: 1) Voice Activity Detection (VAD) Module; 2) Speech Recognition Module 3) and Ontological Module, which is subsquentely described.

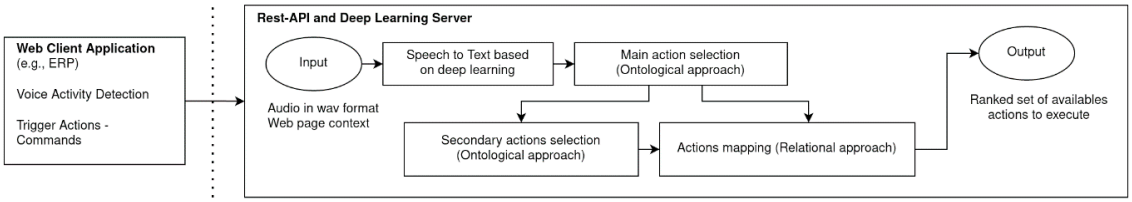

Fig. 1. Voice Interaction Framework architecture proposed

3.1. Audio Capture and VAD proposal module

The implemented module focuses in the development of an application to be used in an office environment where the existent background noise amounts, mostly, to sounds emitted by workers and/or devices located in the room. Consequently, an audio signal captured by the browser (microphone input) will contain, not only the application user utterances, but the existent background noise, as well. In such cases and for speech processing algorithms, it is advisable to include a noise-robust VAD pre-processing module in order to analyse input audio signal for speech and non-speech portions, considering a range noise level. For the proposed framework, an energy thresholding VAD approach is implemented, having a binary decision output where audio frames are classified as speech or non-speech. The implementation is divided in stages, where the first step is the downsampling of input audio to desired sample rate, after which follows the features extraction that enable the characterisation of the input signal, and support the decision of whether an analysed frame contains speech plus noise, or noise only. In this first stage, the signal is analysed considering energy-based features extraction for a defined speech frequency range. Upon launch of the VAD module, an initial pre-defined portion of input signal containing noise only is evaluated for thresholds computation. Subsequently, based on extracted values from a signal frame and estimated thresholds, a decision is made, with speech

classified frames streamed to back end for speech processing. Fig. 2 schematises stages and processes for the implemented VAD module.

VAD module decisions and the segmented speech portions are graphically represented in Fig. 3, with the vertical lines "spch on" and "spch off" signalling the start and end of audio portions classified as speech. The analysed audio consists of voice commands, which have been recorded from vocalisations made within an office working context. A laptop built-in microphone has been used for acquiring diversified audio input that, in turn, allowed to gather a dataset for a preliminary use-case application.

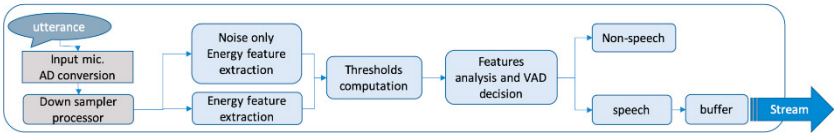

Fig. 2 VAD module processing scheme.

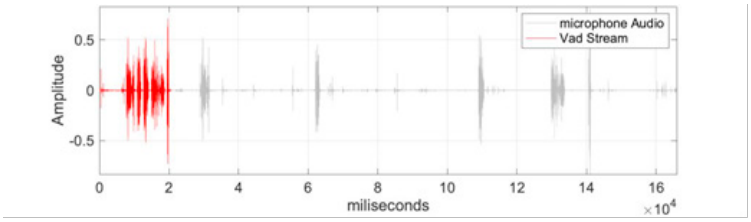

Fig. 3 VAD input audio detection decisions (top) and output segmented speech audio (bottom) graphical representation.

3.2. Speech Recognition Module

Three open-source speech-to-text inference engines are included in this module: a) Mozilla DeepSpeech, b) CMU Sphinx, and c) Vosk. Mozilla DeepSpeech [12] use a Recurrent Neural Network (RNN) architecture implemented with TensorFlow framework. CMU Sphinx [13] is a popular platform among the scientific community, sorted with a wide range of tools and flexible design. It allows an easy and fast alternative to handle ASR-related applications development. Finally, Vosk API [14] is a speech recognition toolkit that integrates offline models for 17 languages. The Speech Recognition module proposal aims to guarantee a trade-off between accuracy and inference time in three languages: English, Spanish and Portuguese. As such, we consider that the three aforementioned engines might be of useful integration, since, besides the referred capabilities, they offer models already trained for the English language. Vosk also covers inference in Spanish and Portuguese, oppositely to DeepSpeech and CMU Sphinx. Taking int account these aspects, we have implemented some training approaches to include all of those capabilities in the workflow of the framework under proposal.

Thereby, the training process runs on a desktop computer with Linux Ubuntu 18.04.5 LTS, Python v3.6.9, Sphinx4 and DeepSpeech v0.9.3, supported over an Intel(R) Core (TM) i9-9900K CPU @ 3.60GHz with 64 GB of RAM and an 8 GB NVIDIA GeForce RTX 2070. We use the free Common Voice datasets for training. The training datasets include a) Mozilla Common Voice Spanish corpus, and b) Mozilla Common Voice Portuguese corpus. The former (Spanish) combines about 324 of validated hours of more than 19000 different voices, from which 25% are from Spain, 36% concerns to participants with ages ranging between 19-39 years old, and 50% are male subjects. The latter one (Portuguese) incorporates around 50 validated hours of audio composed of more than 1000 different voices, from which 67% belong to subjects with ages between 19-39 years old and 81% of them are males.

Comparison functions such as Levenshtein Distance (LD) and Word Error Rate (WER) are valid methods to evaluate the accuracy of the ASR engines. Although, the WER has two major drawbacks: (i) lower word error rate is erroneously associated to higher accuracy, and (ii) measurements consider the difference between the sequences at the word level instead of considering the phoneme level, as LD does. In addition, LD relies on the minimum number of single-character combinations required to shape an individual word or sentence into variants. Therefore, it is suitable to use in this work.

3.3. Ontological Component Module

As described above, the Ontological Component (or ontological layer) is a developed module that comprises the entire process related to the semantic mapping of the ERP and its implementation in companies is always related to the context in which the company operates. Thus, that characterises the company, in first place it is necessary to identify and analyse the organisational context, as well as the sector of activity that characterises the company. Afterwards, we must focus on the analysis of the company's internal processes, documented activities and tasks performed in each department, as well as the actors that carry them out. In this stage, the Unified Modelling Language (UML) based use-cases were focused, as they allow to observe the organisation as a whole and, together with the company, decide which area or areas make sense to be targeted through ontological implementation. All this contextualization is very important to understand the range of products that exist and all their specificities, something that would not be possible just observing the functioning of the ERP in particular.

Next, the analysis must be oriented to the existing data model or reference architecture within the domain of interest. If, at least, one of these artefacts can be found, it is analysed and detailed, while scenarios are built as a way to validate whether the defined ontological layer of the model covers the project's intervention areas. Otherwise, or if some specific requirement is defined by the stakeholder, a proper data model is designed to fit the needs, which is then used as starting point to establish an ontological database schema that is populated according to the client's specifications and terminology. In the end, we proceed with the integration of this ontological database in a visualisation tool, wherein the stakeholder can apply filters, edit the database or add new terminology and relationships, in a more user-friendly way. Fig. 4, depicts the aforementioned process, which has been supporting several other projects.

The ontological processing and mapping were performed by one of the most used graphical database engines, Neo4J [24], which facilitates the development of ontological components. The language used by this tool is Cypher (Query Language) and its syntax allows to combine node patterns and relationships in graphs, visually and intuitively.

Fig. 4 Process Used in Ontology Design.

4. Results and discussion

All the technology has to meet several requirements to achieve the desired impact on users. In this case, processing time, from the capture of an audio signal to the execution of an associated operation, is vital to achieve a smooth and seamless work flow. Relatively long processing times can negatively influence the acceptance and adoption of the system Besides that, there are two major concern issues for both end users and company managers, which are privacy and IT security that, according to [25], are increasingly popular topics among the research community. Both topics are nevertheless out of the scope of this work.

4.1. VAD module results

For applications of speech transmission, it is desirable to save bandwidth by disabling streaming when no speech content is detected. For speech recognition tasks, CPU demands can be decreased if avoiding unnecessary, no-speech content being processed. On the other hand, if speech content is erroneously classified as noise, the module won't be transmitting whole information required by the recognition processing. For the implemented VAD performance evaluation, a criteria focusing on the correct segmentation of speech content has been assumed. Primarily, it should be guaranteed that no speech data is blocked from being transmitted to the recognition processing module. An

objective evaluation parameter relying on speech hit rate (SHR) is employed, indicating the percentage of correct detection of speech content. Additionally, a noise suppression ratio (NSR) is used to measure the amount of noise that has been blocked from further processing, relative to total noise present in the signal segment. For the performance evaluation, the NOIZEUS noisy speech corpus database has been used [26]. From the database, five SNR levels (Signal-to-Noise Ratio) have been tested (clean, SNR=0, 5, 10 and 15) on three speech in babble noise audio samples (Sp01, Sp02 and Sp03). The test audio segments for each SNR level are composed of the three samples separated by noise regions. Fig.5 illustrates for SNR = 10, VAD decisions in relation to test audio segment. In Table 1) test results for each SNR level is summarized.

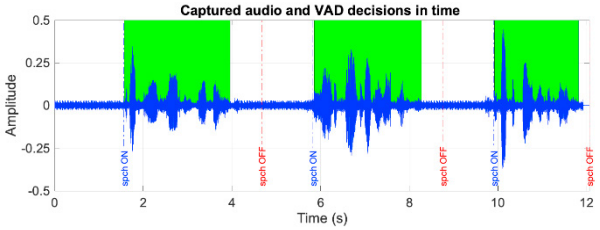

Fig. 5 Graphical representation of SNR = 10 test audio VAD decisions (dashed spch ON and spch OFF vertical lines) and, real portions of speech data (hand marked, yellow background).

Table 1 Summary of Noizeus database selected test samples with respective speech hit rate (SHR) and noise suppression ratio (NSR) evaluation results.

SNR	SHR sp01	SHR sp02	SHR sp03	NSR
clean	0.98	0.98	0.98	0.81
15	0.98	0.98	0.98	0.70
10	0.98	0.98	0.98	0.71
5	0.97	0.99	0.99	0.80
0	1.0	1.0	1.0	0.09

The achieved results indicate that the presented VAD algorithm does not guarantee an effective discrimination of speech from noise in conditions of SNR lower than 5. The case of SNR = 0 illustrates the case, not being capable of discrimination, it classifies the whole signal as speech data.

4.2. Speech-to-text module results

The DeepSpeech training includes a data augmentation process for 15% of the initial dataset. The main hyperparameters set for the training process were: (i) drop_source_layers=5 to adjust all model weights, (ii) batch=16 according to hardware capabilities, (iii) epochs=200 and (iv) n_hidden=2048 as DeepSpeech suggests. Finally, the learning_rate=0.0001 and dropout_rate=0.05 return acceptable results after a trial-and-error process. The CMU Sphinx training process uses sphinxtrain4, compiled for Linux. The train steps include setting up: a) a dictionary with the desired vocabulary; b) the phoneme file, and c) a language model using the online CMU Sphinx tool. Besides, we performed a five-fold cross-validation process using 80% for training and 20% for validation.

To compare the Vosk, DeepSpeech and CMU Sphinx results, we used the testing data set belonging to the fold that obtained the best results in the cross-validation process for the CMU Sphinx engine. The LD and the inference time were computed for each audio to plot its histogram for each language (English, Spanish and Portuguese), as depicted in Fig. 6 and Fig. 7. The results (Fig. 6) show that Vosk obtains similar results when compared with Google Speech API (a commercial tool). Also, it outperforms the accuracy for the three languages with results over 85%. According to our experiments, DeepSpeech obtains the second best results using the LD. The inference times for Portuguese, Spanish and English languages are shown in Fig. 7. In all cases, CMU Sphinx engine shows to be the fastest, with response times equal to or less than 1 second, followed by the Vosk engine with responses between 1.0 and 1.5

seconds, similar to the Google Speech API times. Finally, it is shown that DeepSpeech performs better with the one with slowest response times, with averages above 2.5 seconds for the English dataset, and above 3 seconds for the Spanish dataset.

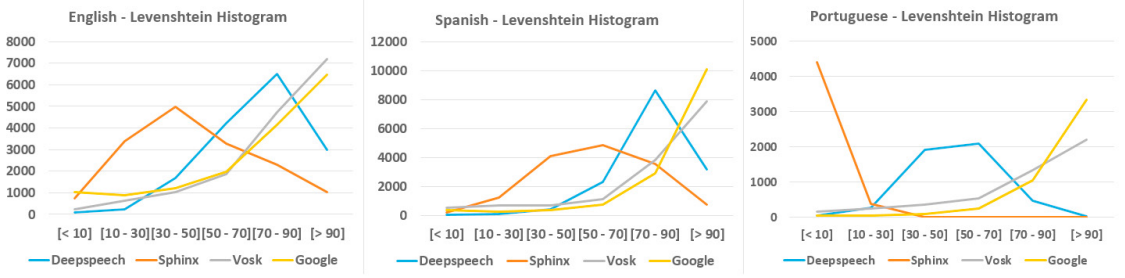

Fig. 6 Histograms of the Levenshtein distance for a) English, b) Spanish and c) Portuguese.

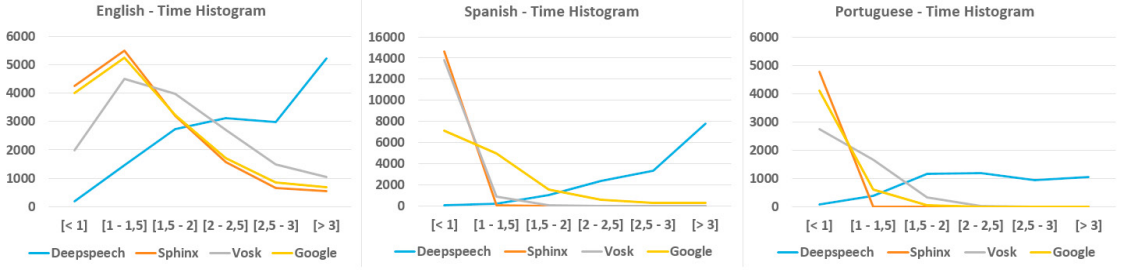

Fig. 7 Histogram of the inference time for a) English, b) Spanish and c) Portuguese.

4.3. Ontological module results

The component that works offline concerns the ontological mapping of the entire context present in the ERP. The process to be carried out to the ontological mapping is dependent on the Speech Recognition module. Taking into account the generation of expressions and words that “Speech to Text" makes available, a mapping of these same expressions is made, associating them to a range of synonyms. The mapping is done in a hierarchical way, that is, in order to ensure that all expressions that are present in the same menu can have access to common commands.

The ontological database contains the mapping of the entire ERP interface, namely its menus, the existing functionalities in each one of the menus and even the expressions that each page of functionalities contains. In order to overcome the use of different words for a same meaning and compose all this mapping, it became necessary to add a set of synonyms for each of the expressions existing in the ERP. This allows for the optimisation of the search engine and guarantees that whatever search is carried out, by voice command, the result will be returned correctly. On the one hand, this method avoids the existence of expressions with exactly the same name and, on the other hand, it allows for a better and faster execution of each command. Thus, the objective presented here is the development of a virtual assistant integrated in an ERP, enabling the end user to search and fill out forms through voice commands.

The ontological database proposed in this work has very specific characteristics where it is intended to organise the database through a combination of language, geography and ERP version (PT-Portugal-V1, ES-Spain-v1 or EN-English-V1). In addition, each of these databases must have a hierarchical structure with the following nodes: Module (page); Functionalities; Expressions and Synonyms. In addition to these structural features of the ontological database, nodes are composed of properties or attributes such as Name, ID, CommandID (this one with the intention of redirecting to a script or url in ERP). The "Expressions" nodes also have the "ERP Product Edition" and "Geography" (country where ERP is used) properties available.

Thus, tests were carried out on the prototype developed in order to ensure full alignment with the customer’s needs. These tests were carried out in a near real context, where company employees were selected to test the queries in the

ontological database. These tests allowed to assess the quality of response times and the quality of information that is returned to the end user. Finally, tests were also carried out to validate the integration between the three modules (VAD, Speech Recognition and Ontological). That said, in the following Fig. 8 it is possible to visualise the final result of this mapping.

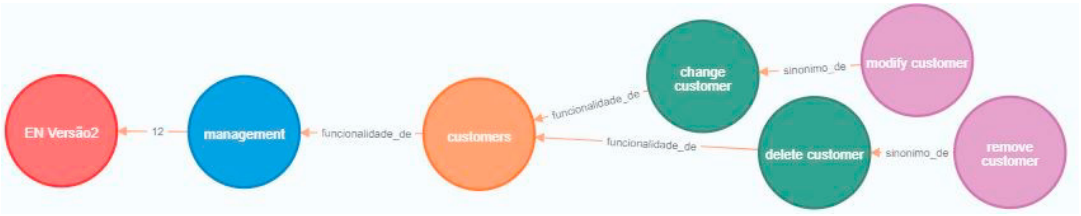

Fig. 8 Excerpt from the ontological mapping performed in Neo4J.

4.4. A typical use cases

The main objective of these developments, explained throughout the paper, is to allow the end user to interact with the ERP through voice commands, as with different virtual assistants on the market (e.g., Siri at Apple; Google Home, Amazon Alexa, between others). That said, as mentioned throughout the paper, we approach the development of three main modules, namely Voice Activity Detection; Automatic Speech Recognition and Ontological Component Module. The operation of the system can be demonstrated based on analysing the flow of information through the modules.

Upon successfull loading of the ERP components on a page (e.g., customer list page), the system is ready to recognize and interpret the audio sequences to suggest actions to the users. As an example, as soon as the user pronounces the following sentence "append customer" - in English -, the VAD module detects and sends the resulting audio to the Speech Recognition module. For use cases such as described above, the Speech Recognition module yields an average LD of 0.1 and conversion time is 930 milliseconds. Next, the Ontological component will return the complying lexical-aware command (e.g., “insert customer”) and a set of suggestions (e.g., “edit customer”) in 10 milliseconds.

5. Conclusions and future works

The proposed voice interaction framework allows IS’s users to introduce information and to navigate in complex functionalities faster, increasing its productivity. Besides, other users with physics deficiencies (hands-free and eyes-free) will be enabled to use this system, until now not allowed.

VAD algorithm testing revealed works adequate results for signals with SNRS > 5. For all scenarios, the inclusion of speech portions is prioritised, for example, for SNRS < 5 the algorithm struggles on the discrimination, thus classifying noise as speech. Further tuning of the algorithm may address minor issues verified, as a slow detection for speech onset events.

After analysing the results returned derived from the three Speech-to-text engines evaluation, we can conclude that, the DeepSpeech engine showed better results than CMU Sphinx in terms of accuracy, but is considerably slower in terms of inference time. The Vosk engine shows the best accuracy results, above 85% - similar to Google Speech API - and low inference times. We recommend the use of Vosk as it seems to show a proper trade-off regarding conversion quality and low inference times.

Regarding the Ontological component we mapped the entire ERP interface (menus, features and expressions), associating a catalogue of terminologies and synonyms that are also mapped in the ontological database. This global mapping allowed to optimise the queries carried out by the engine. When the voice is converted into text, that term follows as a query to the ontological database, from which a corresponding ERP functionality is retrieved.

As future works, we propose deeper discussions on these different technologies' focusing on whether they can be fitted, accepted and adopted for such tasks by the end users. It is not impossible that even if the system meets the

expected functional requirements, it may not be suitable for the tasks it has to perform, therefore, it may not be accepted and adopted by users. Also, regarding more complex scenarios yet to tackle, a greater integration and more detailed frontoffice and backoffice functionalities are among the plans for future improvements.

References

[1] P. Barry, P. Crowley, Modern embedded computing: designing connected, pervasive, media-rich systems, Elsevier, 2012.

[2] Z.-H. Tan, B. Lindberg, Low-complexity variable frame rate analysis for speech recognition and voice activity detection, IEEE Journal of Selected Topics in Signal Processing 4 (5) (2010) 798{807.

[3] J. Kola, C. Espy-Wilson, T. Pruthi, Voice activity detection, Merit Bien (2011) 1{6.

[4] S. Furui, Speech recognition - past, present, and future, NTT Review, Vol. 7, No. 2, pp. 13-18 (1995).

[5] R. S. Rocha, P. Ferreira, I. Dutra, R. Correia, R. Salvini, E. Burnside, A speech-to-text interface for mammoclass, in: 2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS), 2016, pp. 1{6. doi:10.1109/CBMS.2016.25.

[6] L. Besacier, E. Barnard, A. Karpov, T. Schultz, Automatic speech recog-nition for under-resourced languages: A survey, Speech Communication 56 (2014) 85{100. doi:https://doi.org/10.1016/j.specom.2013.07.008.

[7] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M.Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., The kaldi speech recognition toolkit, in: IEEE 2011 workshop on automatic speech recognition and understanding, no. CONF, IEEE Signal Processing So-ciety, 2011.

[8] A. H. Vineel Pratap, wav2letter++: The fastest open-source speech recognition system, CoRR, vol. abs/1812.07625 (2018).

[9] B. Iancu, Evaluating google speech-to-text api's performance for roma-nian e-learning resources., Informatica Economica 23 (1) (2019).

[10] V. Kepuska, G. Bohouta, Comparing speech recognition systems (mi-crosoft api, google api and cmu sphinx), Int. J. Eng. Res. Appl 7 (03) (2017) 20{24.

[11] N. Anggraini, A. Kurniawan, L. K. Wardhani, N. Hakiem, Speech recognition application for the speech impaired using the android-based google cloud speech api, TELKOMNIKA (Telecommunication Comput-ing Electronics and Control) 16 (6) (2018) 2733{2739.

[12] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen,R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al., Deep speech: Scaling up end-to-end speech recognition, arXiv preprint arXiv:1412.5567 (2014).

[13] P. Lamere, P. Kwok, E. Gouvea, B. Raj, R. Singh, W. Walker, M. War-muth, P. Wolf, The cmu sphinx-4 speech recognition system, in: IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong, Vol. 1, 2003, pp. 2{5.

[14] A. Cephei, Vosk alpha cephei API, https://alphacephei.com/vosk/ (2020).

[15] Y. He, Z. Xiang, J. Zheng, Y. Lin, J. Overton, E. Ong, The extensible ontology development (xod) principles and tool implementation to support ontology interoperability, Journal of Biomedical Semantics 9 (01 2018). doi:10.1186/s13326-017-0169-2.

[16] G. da Silva Serapi~ao Leal, W. Guedria, H. Panetto, An ontol-ogy for interoperability assessment: A systemic approach, Jour-nal of Industrial Information Integration 16 (March) (2019). doi:10.1016/j.jii.2019.07.001.

[17] G. Leal, W. Guedria, H. Panetto, An ontology for interoperability as-sessment: A systemic approach, Journal of Industrial Information Integration 16 (07 2019). doi:10.1016/j.jii.2019.07.001.

[18] A. L. Fraga, M. Vegetti, H. P. Leone, Ontology-based solutions for interoperability among product lifecycle management systems: A systematic literature review, Journal of Industrial Information Integration 20 (2020) 100176. doi:https://doi.org/10.1016/j.jii.2020.100176.

[19] S. Nadkarni, R. Prugl, Digital transformation: a review, synthesis and opportunities for future research, Management Review Quarterly 71 (2021) 233{341. doi:10.1007/s11301-020-00185-7.

[20] B. Gajsek, M. Sternad, Information Flow in the Context of the Green Concept, Industry 4.0, and Supply Chain Integration, in: Integration of Information Flow for Greening Supply Chain Management, 2020, pp. 297{323. doi:10.1007/978-3-030-24355-516:

[21] N. M. Nawel Amokrane, Jannik Laval, Philippe Lanco, Mustapha Derras, Analysis of Data Exchanges, Towards a Tooled Approach for Data Interoper-ability Assessment, in: Intelligent Systems: Theory, Research and Innovation in Applications, Vol. 864 of Studies in Computational Intelligence, Springer International Publishing, 2020, pp. 345{363. doi:10.1007/978-3-030-38704-4.

[22] T. Hagedorn, B. Smith, S. Krishnamurty, I. Grosse, Interoperability of dis-parate engineering domain ontologies using basic formal ontology, Journal of Engineering Design 30 (2019) 625 { 654.

[23] Andrei Tara, Alex Butean, Constantin Zam rescu, Robert Learney., An On-tology Model for Interoperability and Multi-organization Data Exchange, in: Arti cial Intelligence and Bioinspired Computational Methods, Vol. 1225 of Advances in Intelligent Systems and Computing, Springer International Publishing, 2020, pp. 284{296. doi:10.1007/978-3-030-51971-1.

[24]Z. Zhu, X. Zhou, K. Shao, A novel approach based on neo4j for multi-constrained exible job shop scheduling problem, Computers Industrial En-gineering 130 (2019) 671{686. doi:https://doi.org/10.1016/j.cie.2019.03.022.

[25] S. Goyal, M. Ahuja, J. Guan, Information systems research themes: A seventeen year data driven temporal analysis, Communications of the Association for Information Systems 43 (2018) 404{431. doi:https://doi.org/10.17705/1CAIS.04323.

[26] Y. Hu, Subjective evaluation and comparison of speech enhancement algo-rithms, Speech Communication 49 (2007) 588{601.