Multilingual Voice Control for Endoscopic Procedures

Simão Afonso, Isabel Laranjo, Joel Braga, Victor Alves(✉), and José Neves

CCTC - Computer Science and Technology Center, University of Minho, Braga, Portugal
{simaopoafonso, joeltelesbraga}@gmail.com,
{isabel, valves, jneves}@di.uminho.pt

Abstract. In this paper it is present a solution to improve the current endoscopic exams' workflow. These exams require complex procedures, such as using both hands to manipulate buttons and pressing a foot pedal at the same time, to perform simple tasks like capturing frames for posterior analysis. In addition to this downside, the act of capturing frames freezes the video. The developed software module was integrated with the *MIVbox* device, a device for the acquisition, processing and storage of the endoscopic results It uses libraries developed by the PocketSphinx project to recognize a small amount of commands. The module was fine-tuned for the Portuguese language which presents some specific difficulties with speech recognition. It was obtained a Word Error Rate (WER) of 23.3 % for the English model and 29.1 % for the Portuguese one.

Keywords: Automatic speech recognition · Hidden Markov Models · Pocketsphinx · Sphinxtrain · Endoscopic procedures

1 Introduction

Nowadays it is accepted by most healthcare professionals that information technologies and informatics are crucial tools to enable a better healthcare practice. The Pew Health Professions Commission (PHPC) recommended that all healthcare professionals should be able to use information technologies in their workout [1]. Indeed the technological evolution has led to an enormous increase in the production of diagnostic tests [2].

EsophagoGastroDuodenoscopy (EGD) and Colonoscopy occupy relevant positions amongst diagnostic tests, since they combine low cost and good medical results. The current endoscopic systems do not fully utilize the current advances in technology, and require a multi-step process to perform simple tasks such as video acquisition and frame capturing. A gastroenterologists needs to press a programmable button on the endoscope to freeze the image and then press the pedal to capture and save the displayed image [3]. These procedures are not optimal and raise several issues, such as limiting the range of possible movements of everyone involved and distracting the gastroenterologist from the objective of the exam: diagnosing anomalies. A possible solution to this problem could consist in adding a voice recognition module to the

R. Giaffreda et al. (Eds.): IoT360 2014, Part I, LNICST 150, pp. 229–235, 2015.
DOI: 10.1007/978-3-319-19656-5_33

video acquisition system, providing hands-free control. This module, named *MIV-control*, will be integrated into the device named *MIVbox*.

The module should be speaker-independent and have a very low error rate, even in noisy environments, and it should be able to capture audio from a microphone continuously, so that it can run in the background unattended, without human intervention. This requires automatic word segmentation to make recognition possible. *Barnett* et al. [4] confirmed that not every language can be recognized with the same accuracy. Although no definitive theory is provided, they present data that confirms the claims. Some possible reasons for some languages being harder to recognize than other include increased frequency of smaller words, language-specific phonemes, and lack of training data [4].

2 Speech Recognition

Automatic Speech Recognition (ASR) is a process by which a computer processes human speech, creating a textual representation of the spoken words [5]. *Aymen* et al. presented the theoretical foundation of Hidden Markov Models (HMM) for automatic speech recognition that underpin most recent implementations [6]. There are several HMM accomplishments, but the most mature ones are the Hidden Markov Model Toolkit (HTK) [7] and the CMU Sphinx system [8]. HTK is a set of libraries used for research in automatic speech recognition, implemented using HMM. Its last release was launched in 2009 and since then it has been largely abandoned [7]. The CMU Sphinx project started on 1990 and already produced 4 (four) versions of its recognizer [8–11]. *Vertanen* [12] tested both the HTK and the Sphinx systems with the Wall Street Journal (WSJ) *corpus* and found no significant differences in error rate and speed which is corroborated by independent researchers [13]. *Huggins-Daines* et al. [14] optimized SPHINX-II for embedded systems, primarily for those with ARM architecture. This work has led to the creation of the PocketSphinx project, a Large Vocabulary Continuous Speech Recognition system developed at the CMU University as an open source initiative [14]. The PocketSphinx project has been used for many different idioms, from Native American and Roma [15], Mexican Spanish [16], Mandarin [17], Arabic [18], and Swedish [19]. These examples show that the PocketSphinx system is flexible enough so that it is relatively easy for people with phonetics training to extend it to other languages, with acceptable results. *Harvey* et al. [5] focused on the creation of models and general optimization tasks, and managed to create a multilingual system that has a 2-s processing time on embedded systems, with error rates below 30 % [5]. *Kirchhoff* et al. suggested other methods to improve the ASR systems' performance, such as including non-acoustic data [20].

3 Voice-Controlled Endoscopic Exams Acquisition

MyEndoscopy is a web-based system developed to link different entities and standardize the patient's clinical process management, in order to promote sharing of information between different entities [21]. It acts like a private cloud, with several

devices providing and using services via common protocols. As more health institutions need this kind of services, it can be useful to pool users in common clouds, with costs shared between all the institutions, increasing the scale at which services are provided, to lower individual costs [21].

The *MIVbox* device is part of that system, and was developed to tackle the problems that healthcare professionals face when performing endoscopic procedures, including replacing the current analogue video acquisition with a more up-to-date integrated digital system [21]. Currently, gastroenterologists use a pedal to capture relevant frames. The main goal of the *MIVcontrol* module is to replace the pedal with voice commands collected from a microphone that interact with the *MIVacquisition* module. As presented on Fig. 1, the *MIVacquisition* module receives the video that is feed directly from the endoscopic tower and provides it to all the other *MIVbox* devices [22]. By integrating the *MIVcontrol* module on the gastroenterologist's workflow, the system can perform frame capturing and video control on the fly, without the need for any extra buttons. In addition, it is much easier to extend it to accept new functionalities, which stand for new commands.

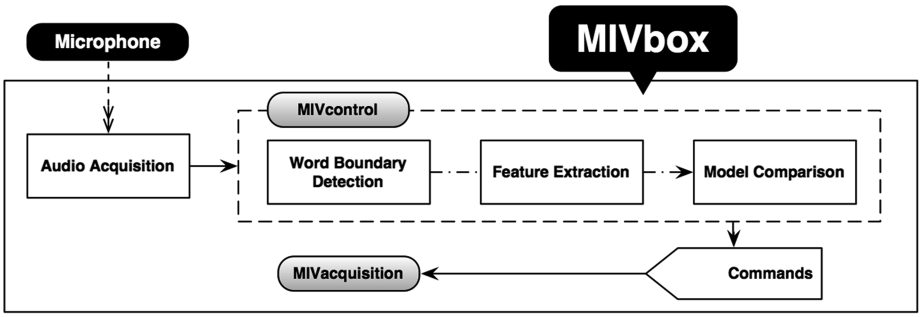

Fig. 1. *MIVcontrol* global architecture

4 Implementation

The creation of the speech model used in the *MIVcontrol* module requires annotated audio and consists of two phases: creating the text model and creating the acoustic model. The language model is a high-level description of all valid phrases (*i.e.* combination of words) in a certain language. It may be classified either as statistical models [23] or as Context-Free Grammars [24]. SphinxBase requires the grammar to be defined in Java Speech Grammar Format (JSGF) [25]. The statistical language model is automatically created based on the command list. The dictionary is a map between each command and the phonemes it contains. A phoneme is defined as the basic unit of phonology, which can be combined to form words. Since the list of required commands is small, all the dictionaries were created manually. It is presented on Fig. 2 as **LmCreate**. The acoustic model is trained using SphinxTrain and maps audio features to the phonemes they represent, for those included in the dictionary.

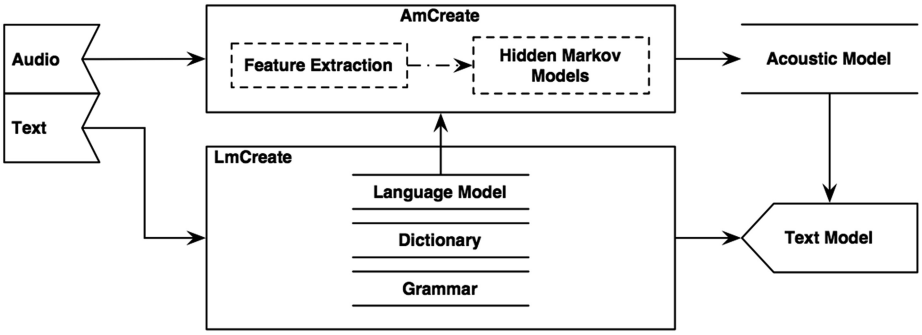

Fig. 2. *MIVcontrol* model training procedure

The training performed by SphinxTrain requires previous knowledge of the dictionary and a transcription for each utterance, in order to map each utterance to its corresponding phonetic information. It is presented on Fig. 2 as **AmCreate**. The audio is split into utterances by tracking silent periods between them, and processed to create a set of features that feed into the HMM. The final result is the most likely command contained in its dictionary.

5 Discussion

The parameters that have a larger impact on the model's accuracy, and so will be tested, are the number of tied states used in the HMM and the number of Gaussian mixture distributions. To test the accuracy of the model, 10-fold cross-validation is performed on the data. The audio corpus in which the system was tested contained two languages, Portuguese and English, with a total of 1405 recordings, totalling 25 min of speech, recorded by 5 female and 7 male speakers, recorded in both noisy and quiet conditions. The vocabulary used was chosen so that it would be useful for direct application is the context of endoscopic procedures. The results are presented as precision-recall matrices. The OTHER label consists of unrecognizable commands or out-of-vocabulary predictions.

To the English model, the best results were obtained for 100 Gaussian mixture distribution and 8 tied states (Table 1). This model has a Total Error Rate of 23.27 %, corresponding to 128 errors in 550 commands.

Table 1. Confusion Matrix for the English model, for 100 Gaussian mixtures and 8 tied states

	"continue"	"end"	"hold"	"start"	"take picture"	OTHER	RECALL
"continue"	**69**	8	3	2	21	7	**62.73 %**
"end"	3	**74**	3	17	9	4	**67.27 %**
"hold"	0	4	**89**	8	0	9	**80.91 %**
"start"	0	7	4	**95**	0	4	**86.36 %**
"take picture"	1	4	2	2	**95**	6	**86.36 %**
OTHER	0	0	0	0	0	**0**	***0***
PRECISION	**94.52 %**	**76.29 %**	**88.12 %**	**76.61 %**	**76.00 %**	***30***	***550***

For the Portuguese (pt-PT) model, the best results were obtained for 150 Gaussian mixture distributions and 8 tied states (Table 2). This model has a Total Error Rate is 29.1 %, corresponding to 249 errors in 855 commands. The difference can be explained by the fact that the similarity among the Portuguese commands is superior, and some sounds might not be detected by the recognizer.

Table 2. Confusion Matrix for the pt-PT model, for 150 Gaussian mixtures and 8 tied states

	"acaba"	"começa"	"continua"	"pausa"	"tira imagem"	OTHER	RECALL
"acaba"	**160**	1	0	0	0	10	**93.57 %**
"começa"	22	**87**	2	2	41	17	**50.88 %**
"continua"	19	2	**110**	0	13	27	**64.33 %**
"pausa"	44	2	0	**102**	2	21	**59.65 %**
"tira imagem"	14	2	3	0	**147**	5	**85.96 %**
OTHER	0	0	0	0	0	**0**	***0***
PRECISION	**61.78 %**	**92.55 %**	**95.65 %**	**98.08 %**	**72.41 %**	***80***	***855***

6 Conclusions

This paper presents a voice recognizer for a very small vocabulary to be used as a command and control system, integrated on the *MyEndoscopy* system, leveraging the capabilities of the CMU Sphinx project, particularly the PocketSphinx libraries. It was created to respond to issues with the current solutions reported by gastroenterologists, and can be presented as an alternative to cloud-based solutions, such as Google Speech API. In a medical environment, cloud-based solutions pose certain challenges that might degrade their desirability, such as security and privacy issues. Legal reasons on systems that deal with sensitive data also have to be accounted with. Having a system that can be installed inside the healthcare institutions' network without external dependencies is a plus for the reasons presented above. The results obtained required an extensive tuning to the PocketSphinx parameters, particularly for the Portuguese model. This tuning is necessary because the system was not designed to recognize Romance languages like Portuguese one. Future work may involve the creation of models adapted to each specific user, instead of the one-size-fits-all approach followed in this work.

Acknowledgments. This work is funded by ERDF - European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT - *Fundação para a Ciência e a Tecnologia* (Portuguese Foundation for Science and Technology) within project PEst-OE/EEI/UI0752/2014.

References

1. O'Neil, E.H.: Recreating Health Professional Practice For A New Century, p. 106. Pew Health, San Francisco (1998)
2. Summerton, N.: Positive and negative factors in defensive medicine: a questionnaire study of general practitioners. BMJ **310**, 27–29 (1995)

3. Canard, J.M., Létard, J.-C., Palazzo, L., et al.: Gastrointestinal Endoscopy in Practice. 1st ed., p. 492. Churchill Livingstone, Paris (2011)
4. Barnett, J., Corrada, A., Gao G., et al.: Multilingual speech recognition at dragon systems. In: Proceeding Fourth International Conference on Spoken Language Process, ICSLP 1996, pp. 2191–2194. IEEE (1996)
5. Harvey, A.P., McCrindle, R.J., Lundqvist, K., Parslow, P.: Automatic speech recognition for assistive technology devices. In: Proceedings Of The 8th International Conference On Disability Virtual Reality And Associated Technologies. Valparaíso, pp 273–282 (2010)
6. Aymen, M., Abdelaziz, A., Halim, S., Maaref, H.: Hidden Markov Models for automatic speech recognition. In: 2011 International Conference on Communications, Computing and Control Applications, pp. 1–6. IEEE (2011)
7. Young, S., Evermann, G., Kershaw, D., et al.: HTK speech recognition toolkit. http://htk.eng.cam.ac.uk/. Accessed 3 February 2014
8. Lee, K.-F., Hon, H.-W., Reddy, R.: An overview of the SPHINX speech recognition system. IEEE Trans. Acoust. **38**, 35–45 (1990)
9. Huang, X., Alleva, F., Hon, H.-W., et al.: The SPHINX-II speech recognition system: an overview. Comput. Speech Lang. **7**, 137–148 (1993)
10. Seltzer, M.: SPHINX III signal processing front end specification, vol. 31, pp. 1–4 (1999)
11. Lamere, P., Kwok, P., Gouvea, E., et al.: The CMU SPHINX-4 speech recognition system. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003). Hong Kong, pp. 2–5 (2003)
12. Vertanen, K.: Baseline WSJ Acoustic Models for HTK and Sphinx: training recipes and recognition experiments. Cavendish Laboratory University, Cambridge (2006)
13. Ma, G., Zhou, W., Zheng, J., et al.: A comparison between HTK and SPHINX on chinese mandarin. In: IJCAI International Joint Conference on Artificial Intelligence, pp. 394–397 (2009)
14. Huggins-Daines, D., Kumar, M., Chan, A., et al.: Pocketsphinx: a free, real-time continuous speech recognition system for hand-held devices. In: 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. I-185–I-188 (2006)
15. John, V.: Phonetic decomposition for speech recognition of lesser-studied languages. In: Proceedings of 2009 International Conference on Intercultural Collaboration, p. 253. ACM Press, New York (2009)
16. Varela, A., Cuayáhuitl, H., Nolazco-Flores, J.A.: Creating a Mexican Spanish version of the cmu sphinx-iii speech recognition system. In: Sanfeliu, A., Ruiz-Shulcloper, J. (eds.) CIARP 2003. LNCS, vol. 2905, pp. 251–258. Springer, Heidelberg (2003)
17. Wang, Y., Zhang, X.: Realization of Mandarin continuous digits speech recognition system using sphinx. In: 2010 International Symposium on Computer, Communication, Control and Automation, pp. 378–380 (2010)
18. Hyassat, H., Abu Zitar, R.: Arabic speech recognition using SPHINX engine. Int. J. Speech Technol. **9**, 133–150 (2008)
19. Salvi, G.: Developing Acoustic Models For Automatic Speech Recognition (1998)
20. Kirchhoff, K., Fink, G.A., Sagerer, G.: Combining acoustic and articulatory feature information for robust speech recognition. Speech Commun. **37**, 303–319 (2002)
21. Laranjo, I., Braga, J., Assunção, D., Silva, A., Rolanda, C., Lopes, L., Correia-Pinto, J., Alves, V.: Web-based solution for acquisition, processing, archiving and diffusion of endoscopy studies. In: Omatu, S., Neves, J., Corchado Rodriguez, J.M., Paz Santana, J.F., Gonzalez, S.R. (eds.) Distributed Computing and Artificial Intelligence. AISC 217, pp. 317–324. Springer, Heidelberg (2013)
22. Braga, J., Laranjo, I., Assunção, D., et al.: Endoscopic imaging results: web based solution with video diffusion. Procedia Technol. **9**, 1123–1131 (2013)

23. Clarkson, P., Rosenfeld, R.: Statistical language modeling using the CMU-cambridge toolkit. In: 5th European Conference on Speech Communication and Technology, ISCA Archive, Rhodes, Greece, pp. 2707–2710 (1997)
24. Bundy, A., Wallen, L.: Context-free grammar. In: Bundy, A., Wallen, L. (eds.) Catalogue of Artificial Intelligence Tools, pp. 22–23. Springer, Heidelberg (1984)
25. Hunt, A.: JSpeech Grammar Format (2000)